- Licence Sciences et technologie
- Mention science de la vie – L2

SCIENCES SORBONNE UNIVERSITÉ

INTRODUCTION A LA MODELISATION EN BIOLOGIE (BM1) - LU2SV382

Modelisation statistique (Cours 1)

Martin LARSEN

www.immulab.fr

- Licence Sciences et technologie
- Mention science de la vie – L2

SCIENCES SORBONNE UNIVERSITÉ

Prise en main d'Excel

Martin LARSEN

www.immulab.fr

Plan d'action

1. Statistiques descriptives
2. Visualisation des données
3. Distribution des données,
4. Intervalle de pari/confiance et Inférence statistique

Excel - prise en main

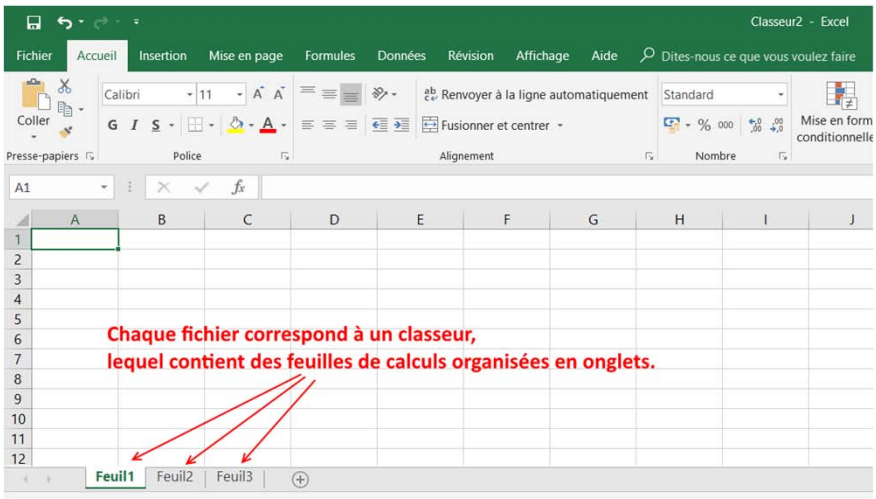
Excel c'est quoi ?

Le logiciel Excel intègre des fonctions :

1. de calcul numérique (à l'aide de formules de calcul)
2. de représentation graphique
3. d'analyse de données (notamment de tableau croisé dynamique)
4. et de programmation, laquelle utilise les macros écrites dans un langage (Visual Basic) – **HORS CURSUS**

Excel - prise en main

Un fichier Excel c'est quoi ?



Excel - prise en main

Un tableau Excel c'est quoi ?

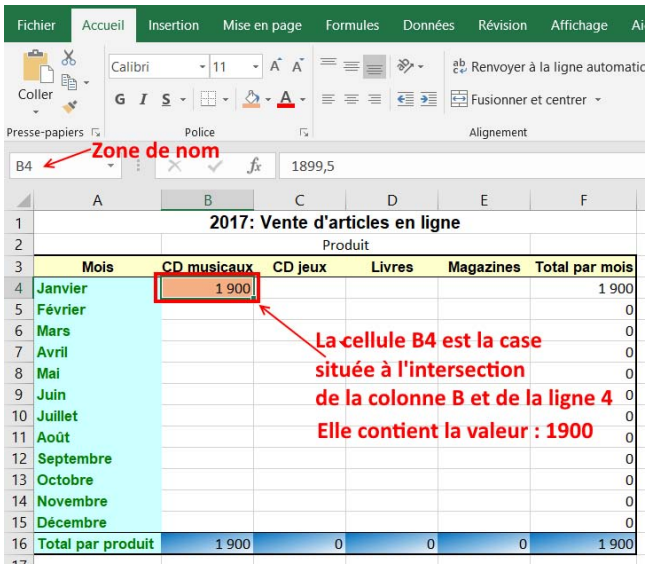
Colonnes : A, B, C ...G ...

	A	B	C	D	E	F
1	2017: Vente d'articles en ligne					
2		Produit				
3	Mois	CD musicaux	CD jeux	Livres	Magazines	Total par mois
4	Janvier	1 900	6 520	2 751	514	11 685
5	Février	1 888	4 522	2 987	687	10 085
6	Mars	1 457	4 523	1 237	125	7 342
7	Avril	1 255	9 563	1 588	385	12 790
8	Mai	957	7 521	2 430	873	11 781
9	Juin	1 250	6 385	2 100	268	10 004
10	Juillet	1 278	2 341	1 251	487	5 356
11	Août	958	4 588	1 955	453	7 953
12	Septembre	1 752	2 384	1 235	365	5 737
13	Octobre	1 521	5 367	2 840	569	10 298
14	Novembre	245	7 851	2 462	235	10 793
15	Décembre	1 238	4 853	3 100	751	9 943
16	Total par produit	15 697	66 420	25 936	5 714	113 767
17						
18						

Lignes : 1, 2, 3, 4, ...16, 17 ....

Excel - prise en main

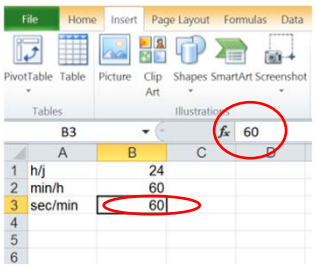
Un tableau Excel c'est quoi ?



	A	B	C	D	E	F
1		2017: Vente d'articles en ligne				
2		Produit				
3	Mois	CD musicaux	CD jeux	Livres	Magazines	Total par mois
4	Janvier	1 900				1 900
5	Février					0
6	Mars					0
7	Avril					0
8	Mai					0
9	Juin					0
10	Juillet					0
11	Août					0
12	Septembre					0
13	Octobre					0
14	Novembre					0
15	Décembre					0
16	Total par produit	1 900	0	0	0	1 900

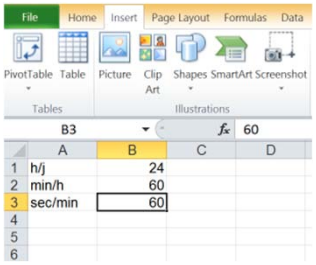
Excel - prise en main

Calcul numérique en Excel

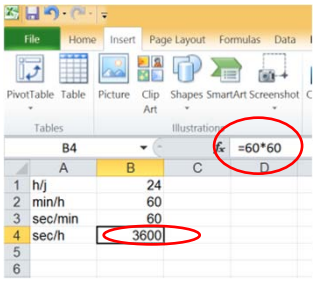


Excel - prise en main

Calcul numérique en Excel

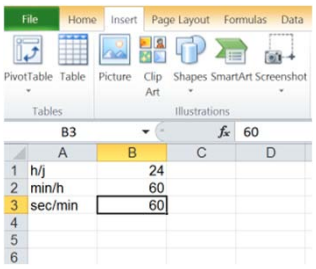


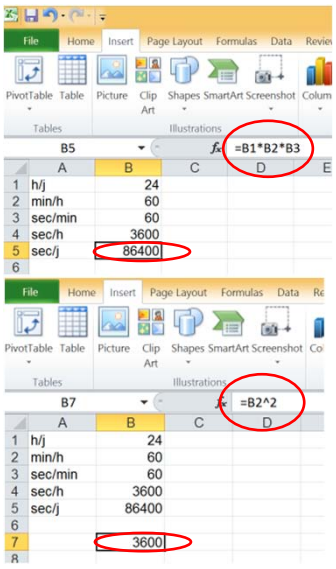
Formule:
Initié par =
Les opérateurs arithmétiques:
+ - * / ^
Les opérateurs relationnels:
=, <>,<, >, <=, >=
Les opérateurs logiques:
ET, OU, NON (des fonctions en Excel)



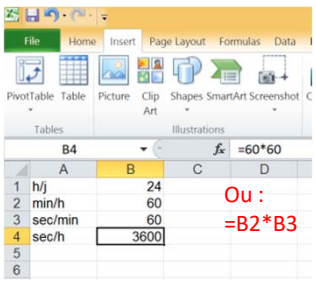
Excel - prise en main

Calcul numérique en Excel





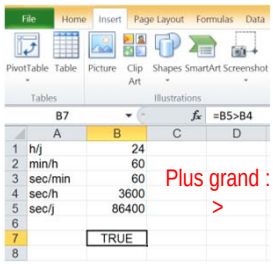
Formule:
Initié par =
Les opérateurs arithmétiques:
+ - * / ^

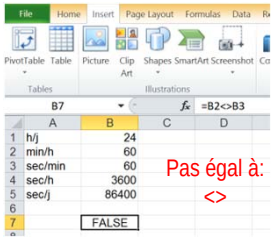


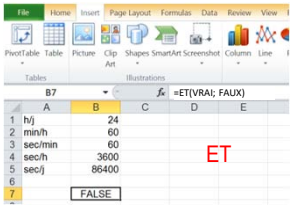
Excel - prise en main

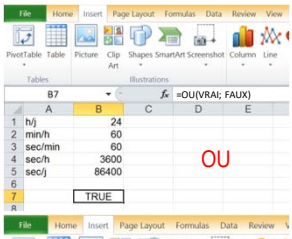
Calcul numérique en Excel

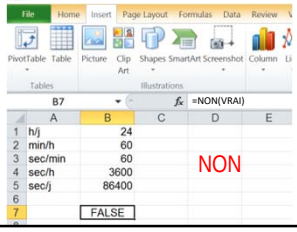
Formule:
Initié par =
Les opérateurs arithmétiques:
+ - * / ^
Les opérateurs relationnels:
=, <>,<, >, <=, >=
Les opérateurs logiques:
ET, OU, NON (des fonctions en Excel)











Excel - prise en main

Des fonctions en Excel

Plage de cellules : Reference 1 **:** Reference 2 (e.g. A1**:**A3)
Decaler(Reference 1, ligne, colonne) (e.g. A3 = Decaler(A1;2;0))

Statistique descriptive:
Moynne(Plage)
Mediane(Plage)
Mode(Plage)
Min(Plage)
Centile(Plage;0,25) (Q1)
Centile(Plage;0,75) (Q3)
Max(Plage)
Ecartype(Plage)=s
Var(Plage)=s^2 (echantillon)

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

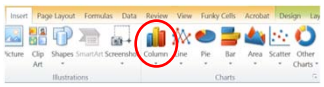
Attention!
Var.P(Plage)= σ^2(population)

$$\sigma^2 = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{n} = \frac{1}{n}\left(\sum_{i=1}^{n}{x_i}^2 - n\mu^2\right)$$

Pour info!
Racine(4) = $\sqrt{4} = 2$

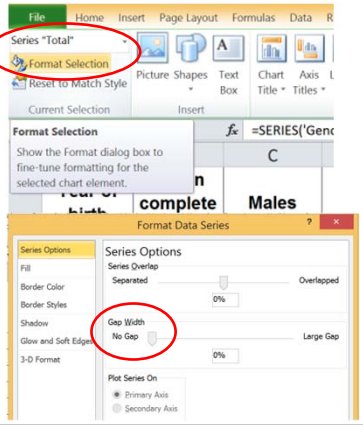
Excel - Diagramme en bâtons (superposés)

Démographie de la population française (Tableau 1):



	A	B	C	D	E
1	Year of birth	Age in completed years	Males	Females	Total
2	2018	0	364,155	347,749	711,904
3	2017	1	370,453	355,472	725,925
4	2016	2	378,518	363,162	741,680
5	2015	3	387,906	372,402	760,308
6	2014	4	399,232	387,042	786,274
7	2013	5	407,611	389,920	797,531
8	2012	6	417,471	396,835	814,306
9	2011	7	418,623	403,349	821,972
10	2010	8	429,919	412,555	842,474
11	2009	9	427,917	408,232	836,149
92	1928	90	55,382	124,322	179,704
93	1927	91	44,797	105,456	150,253
94	1926	92	34,519	91,072	125,591
95	1925	93	27,317	76,447	103,764
96	1924	94	20,525	61,235	81,760
97	1923	95	14,477	48,398	62,875
98	1922	96	10,101	37,882	47,983
99	1921	97	7,239	27,754	34,993
100	1920	98	4,977	19,813	24,790
101	1919	99	2,058	8,273	10,331
102	1918 or before	100 or over	2,976	12,670	15,646
103		Total	32,394,531	34,598,168	66,992,699

1. Affichez le nombre d'individus pour chaque groupe d'âge sous forme de diagramme en bâtons.
2. Formater les données - supprimer les écart entre des bâtons.



Source: https://www.insee.fr/en/statistiques/2382597?sommaire=2382613

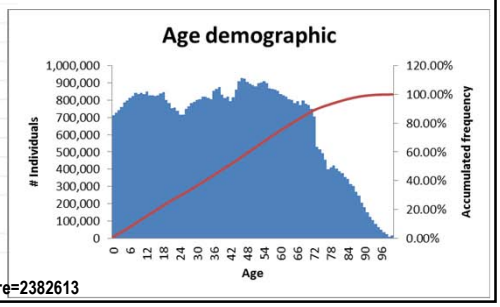
Excel - Diagramme en bâtons superposés

Démographie de la population française (Tableau 1):

	A	B	C	D	E
1	Year of birth	Age in completed years	Males	Females	Total
2	2018	0	364,155	347,749	711,904
3	2017	1	370,453	355,472	725,925
4	2016	2	378,518	363,162	741,680
5	2015	3	387,906	372,402	760,308
6	2014	4	399,232	387,042	786,274
7	2013	5	407,611	389,920	797,531
8	2012	6	417,471	396,835	814,306
9	2011	7	418,623	403,349	821,972
10	2010	8	429,919	412,555	842,474
11	2009	9	427,917	408,232	836,149
92	1928	90	55,382	124,322	179,704
93	1927	91	44,797	105,456	150,253
94	1926	92	34,519	91,072	125,591
95	1925	93	27,317	76,447	103,764
96	1924	94	20,525	61,235	81,760
97	1923	95	14,477	48,398	62,875
98	1922	96	10,101	37,882	47,983
99	1921	97	7,239	27,754	34,993
100	1920	98	4,977	19,813	24,790
101	1919	99	2,058	8,273	10,331
102	1918 or before	100 or over	2,976	12,670	15,646
103		Total	32,394,531	34,598,168	66,992,699

1. Affichez le nombre d'individus pour chaque groupe d'âge sous forme de diagramme en bâtons.
2. Formater les données - supprimer les écart entre des bâtons.
3. Calcule la fréquence accumulée et trace sous forme de ligne. (en cellule F2: =Somme($E$2:E2) et puis copie).

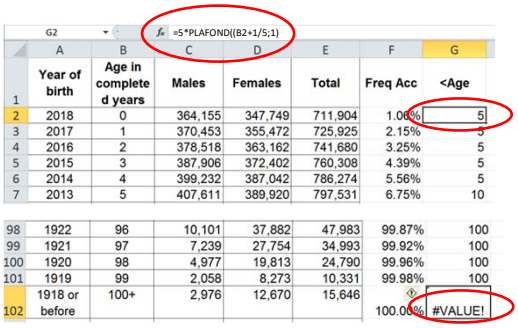
!! Touche de fonction F4 transforme E2 en **$E$2**.
$ Fixe le référence – colonne, ligne ou les deux.



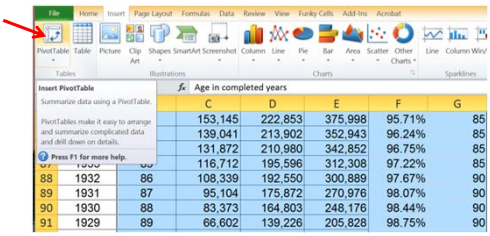
Source: https://www.insee.fr/en/statistiques/2382597?sommaire=2382613

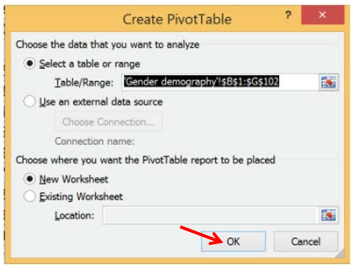
Excel - Tableau Croisé Dynamique (Pivot tables)

Démographie de la population française (Tableau 1):



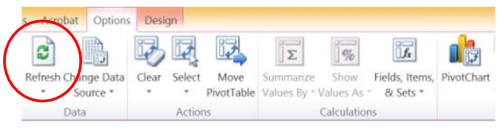
1. Faire des groupes d'âge avec un intervalle de n années:
   = n * Plafond ((B2+1) / n; 1)
   Par exemple. n = 5
2. Corrigez les lignes 102 à 100.
3. Sélectionner des données et créer un tableau croisé dynamique

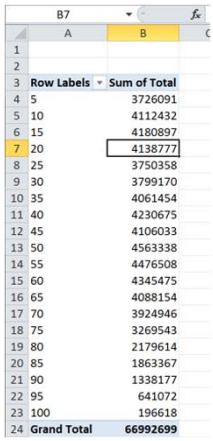


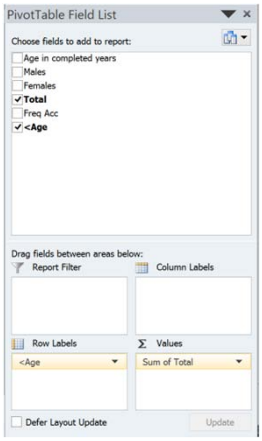


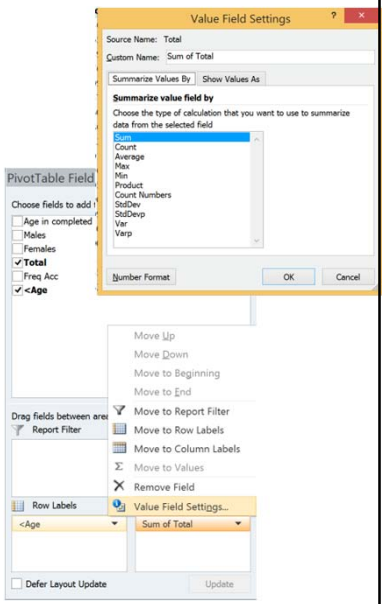
Excel - Tableau Croisé Dynamique (Pivot tables)

Démographie de la population française (Tableau 1):



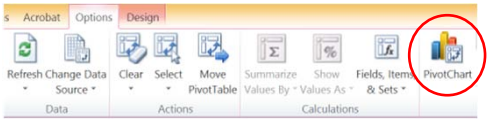


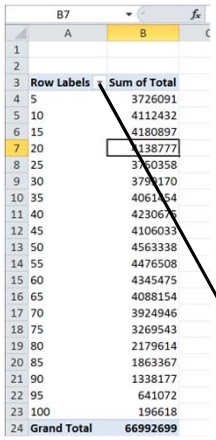




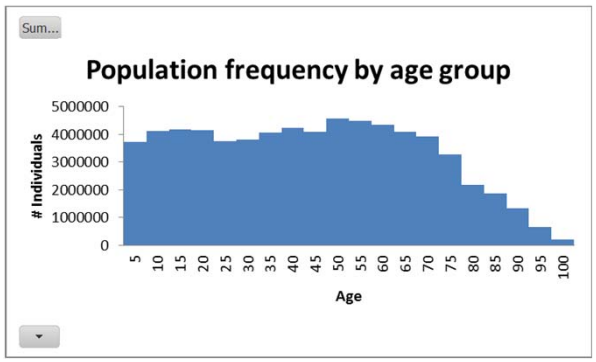
Excel - Tableau Croisé Dynamique (Pivot tables)

Démographie de la population française (Tableau 1):





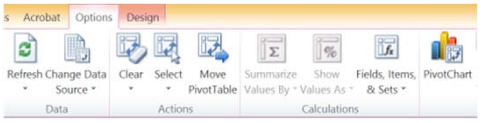
Row Labels	Sum of Total
5	3726091
10	4112432
15	4180897
20	4138777
25	3750358
30	3799170
35	4061454
40	4230675
45	4106033
50	4563338
55	4476508
60	4345475
65	4088154
70	3924946
75	3269543
80	2179614
85	1863367
90	1338177
95	641072
100	196618
Grand Total	66992699

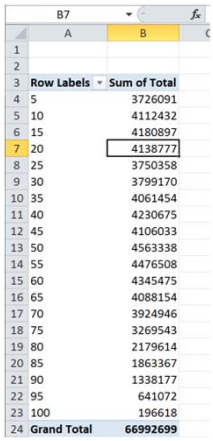


Trier la colonne

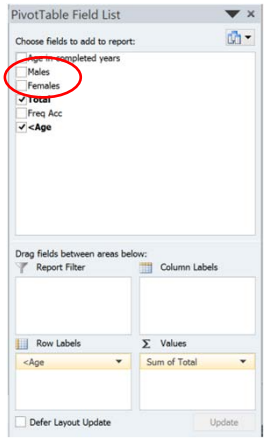
Excel - Tableau Croisé Dynamique (Pivot tables)

Démographie de la population française (Tableau 1):





Row Labels	Sum of Total
5	3726091
10	4112432
15	4180897
20	4138777
25	3750358
30	3799170
35	4061454
40	4230675
45	4106033
50	4563338
55	4476508
60	4345475
65	4088154
70	3924946
75	3269543
80	2179614
85	1863367
90	1338177
95	641072
100	196618
Grand Total	66992699



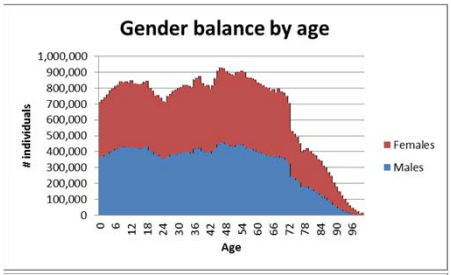
Analyser la parité hommes-femmes stratifiée par «âge» et par «groupes d'âge».

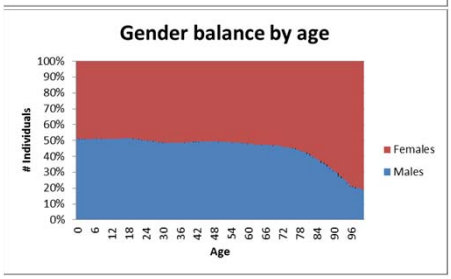
1. Utiliser le jeu de données complet (« age »)
2. Utilisez un tableau croisé dynamique pour effectuer l'analyse stratifiée par groupes d'âge.

Excel - Tableau Croisé Dynamique (Pivot tables)

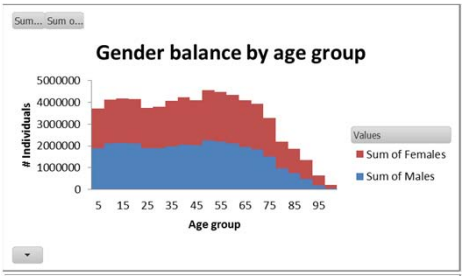
Démographie de la population française (Tableau 1):

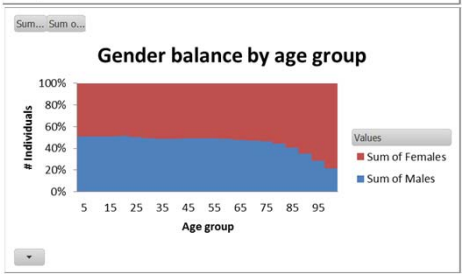
Stratified by "Age"





Stratified by "Age group"





SCIENCES SORBONNE UNIVERSITÉ

Prédiction, inférence et distribution normale

Population / échantillon

Population

Échantillon

Loi connue dans la population
- Loi Binomiale
- Loi Normale
- …

Prédiction de ce qui sera observé dans l'échantillon
⇨ Intervalle de pari

Population / échantillon

Population

Échantillon

Observation

Estimation

Généralisation / Inférence

Moyenne et pourcentage

Intervalle de confiance

Intervalle de Pari

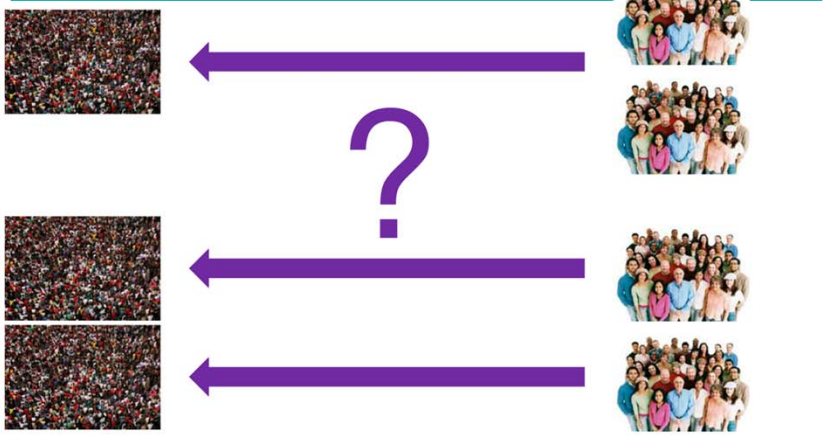
Intervalle de Pari

- $IP_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}}; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$
- $IP_{1-\alpha} = \left[\pi - u_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}; \pi + u_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}\right]$
- A partir de la connaissance de la population on imagine ce que l'on obtiendra dans un échantillon

Intervalle de confiance

- $IC_{1-\alpha} = \left[m - u_\alpha \frac{s}{\sqrt{n}}; m + u_\alpha \frac{s}{\sqrt{n}}\right]$
- $IC_{1-\alpha} = \left[p - u_\alpha \sqrt{\frac{p(1-p)}{n}}; p + u_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$
- A partir de l'observation d'un échantillon on en déduit une information sur la population

Deux échantillons viennent-ils d'une ou de deux populations ?



Démarche

- Construire un intervalle de confiance à partir des données observées et voir si cela correspond à ce que l'on attend en population

Exemple : Le traitement A est il plus efficace que le traitement B ?

- Estimer efficacité avec les deux traitements
- Calculer la différence et l'intervalle de confiance de cette différence
- Hypothèse : Si A = B alors la différence en population $\Delta = 0$
- Voir si l'intervalle de confiance avec un certain risque est compatible avec cette hypothèse

Démarche

- Construire un intervalle de pari à partir des hypothèses et voir si les données de l'échantillon correspondent à ce qui est attendu

Exemple : Le traitement A est-il plus efficace que le traitement B ?

- Hypothèses :
  - Hypothèse nulle H_0 : Si A = B alors la différence en population $\Delta = 0$
  - Hypothèse alternative H_1 : Si A ≠ B alors la différence en population $\Delta \neq 0$
- Trouver une statistique de test correspondant à cette hypothèse
- Etablir un intervalle de pari de cette statistique
- Estimer l'efficacité avec les deux traitements
- Voir si l'intervalle de pari est compatible avec l'estimation

Distribution de X et $\bar{X}$

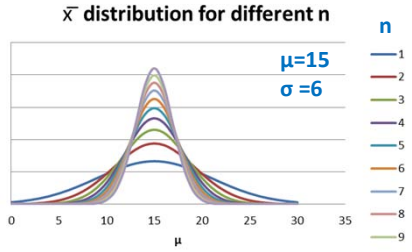
X est une variable aléatoire continuée (VAC)

- Distribution de Echantionnage de Moyenne:

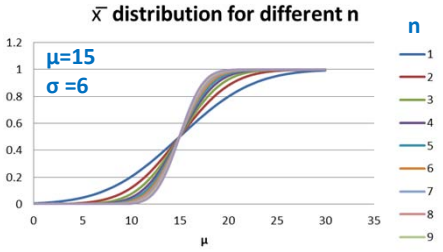
$$X \sim \mathcal{N}(\mu_x, \sigma_x) \Rightarrow \bar{X} \sim \mathcal{N}(\mu_x, \frac{\sigma_x}{\sqrt{n}})$$

Pour n=1, nous avons $X \sim \mathcal{N}(\mu_x, \sigma_x) = \bar{X} \sim \mathcal{N}(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

Fonction de densité de probabilité



Probabilité accumuler



Distribution de X et $\bar{X}$

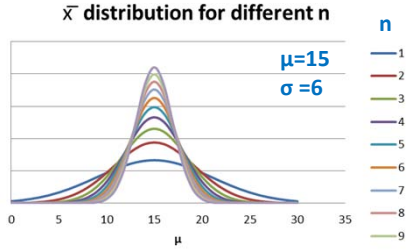
LOI.NORMALE.N(x;μ;σ/RACINE(N);cumulative)

B5 fx LOI.NORMALE.N($A5;$B$1;$B$2/RACINE(B$3);FAUX)

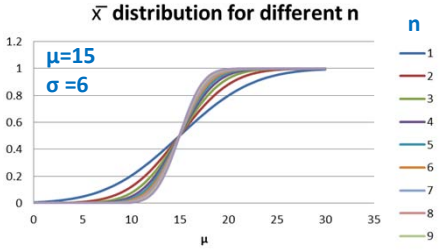
	A	B	C	D	E	F	G	H
1	mu	15						
2	sigma	6						
3	n	1	2	3	4	5	6	7
4	X	1	2	3	4	5	6	7
5	0	0.002921383	0.000182	9.77E-06	4.96E-07	2.43E-08	1.17E-09	5.56E-11
6	1	0.004370315	0.000406	3.27E-05	2.48E-06	1.82E-07	1.31E-08	9.32E-10
7	2	0.006358771	0.00086	0.000101	1.11E-05	1.19E-06	1.25E-07	1.29E-08
8	3	0.008998494	0.001722	0.000285	4.46E-05	6.75E-06	1E-06	1.46E-07
9	4	0.012385194	0.003263	0.000744	0.00016	3.33E-05	6.8E-06	1.37E-06
10	5	0.016579523	0.005847	0.001785	0.000514	0.000143	3.91E-05	1.05E-05
11	6	0.021586266	0.009911	0.003941	0.001477	0.000536	0.000191	6.69E-05
12	7	0.027335012	0.015893	0.008002	0.003799	0.001746	0.000786	0.000349

Pour n=1, nous avons $X \sim \mathcal{N}(\mu_x, \sigma_x) = \bar{X} \sim \mathcal{N}(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

Fonction de densité de probabilité
Cumulative = FAUX



Probabilité accumuler
Cumulative = VRAI



Distribution de X et $\bar{X}$

X est une variable aléatoire continuée (VAC)

- Distribution de Echantionnage de Moyenne:

$$X \sim \mathcal{N}(\mu_x, \sigma_x) \Rightarrow \bar{X} \sim \mathcal{N}(\mu_x, \frac{\sigma_x}{\sqrt{n}})$$

- Théorème centrale limite (TCL):

  X suit n'importe quelle distribution. Pour n> 30 => TCL => $\bar{X} \sim \mathcal{N}(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

- Autre

  En l'absence de distribution connue pour X et n <30, nous ne pouvons pas identifier une distribution pour $\bar{X} \sim$??

Théorème centrale limite

E.g. distribution binomial :

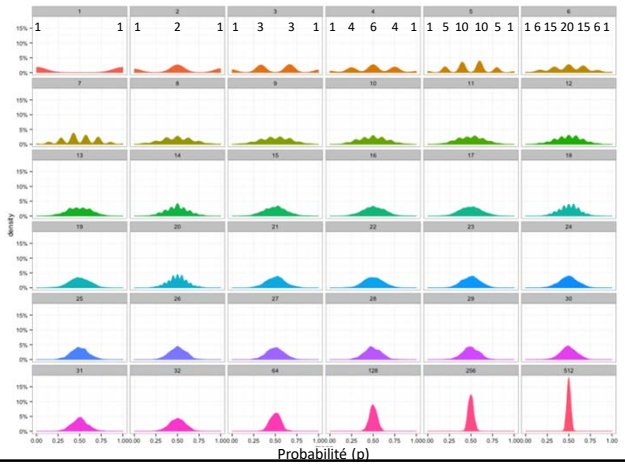
$$X \sim B(n, \pi) \Rightarrow \begin{bmatrix} n\pi > 5 \\ n(1-\pi) > 5 \end{bmatrix} \Rightarrow N\left(\mu = n\pi, \sigma^2 = n\pi(1-\pi)\right)$$

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

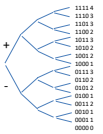
π =0.5 and n∈[1-512]

Fréquence de succès au cours de n épreuves de Bernoulli

Le moyenne de 1000 échantillons était analyse



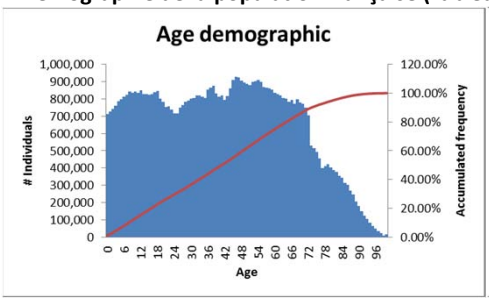
k	$\frac{n!}{k!\,(n-k)!}$
3	1
2	3
1	3
0	1

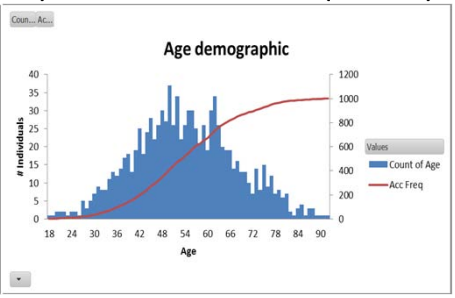


k	$\frac{n!}{k!\,(n-k)!}$
4	1
3	4
2	6
1	4
0	1

Distribution d'Age (statistique descriptive)

Démographie de la population française (Tableau 1) et une échantillon inconnu (Tableau 2):





Descriptive statistics of population:		
n	66,992,699	
μ	41.2 years	
σ^2	586.4 years^2	
σ	24.2 years	

Descriptive statistics of sample		
n	1000	=NB(A2:A1001)
Mean	54.2 years	=moyenne(A2:A1001)
s^2	184.3 years^2	=Var.S(A2:A1001)
s	13.6 years	=ECARTYPE(A2:A1001)

Moyenne et pourcentage

Intervalle de confiance

Intervalle de Pari

Intervalle de Pari

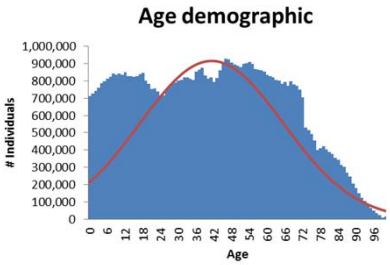
- $IP_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}}; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$
- $IP_{1-\alpha} = \left[\pi - u_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}; \pi + u_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}\right]$
- A partir de la connaissance de la population on imagine ce que l'on obtiendra dans un échantillon

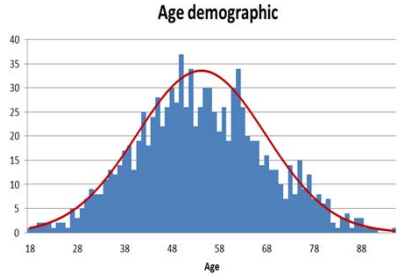
Intervalle de confiance

- $IC_{1-\alpha} = \left[m - u_\alpha \frac{s}{\sqrt{n}}; m + u_\alpha \frac{s}{\sqrt{n}}\right]$
- $IC_{1-\alpha} = \left[p - u_\alpha \sqrt{\frac{p(1-p)}{n}}; p + u_\alpha \sqrt{\frac{p(1-p)}{n}}\right]$
- A partir de l'observation d'un échantillon on en déduit une information sur la population

Intervalle de pari vs. Intervalle de confiance

Démographie de la population française (Tableau 1) et une échantillon inconnu (Tableau 2):





Statistique descriptive (population):

n	66,992,699
μ	41.2 years
σ^2	586.4 years^2
σ	24.2 years

X ne suis pas $\mathcal{N}(\mu,\sigma)$ (Visuellement)

Mais Théorème Central Limite (TCL, n>30):

$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$

$IP_{1-\alpha}$: $\bar{x} \in [\mu - u_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$

= [39.7; 42.7] avec risque α=5% (n=1000)

$N \to \infty$ ↙ Intervalle rétréci

Statistique descriptive (échantillon)

n	1000
Mean	54.2 years
s^2	184.3 years^2
s	13.6 years

$IC_{1-\alpha}$: $\mu \in [\bar{x} - u_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{x} + u_{\alpha/2}\frac{S}{\sqrt{n}}]$

= [53.3;55.0] avec risque α=5%

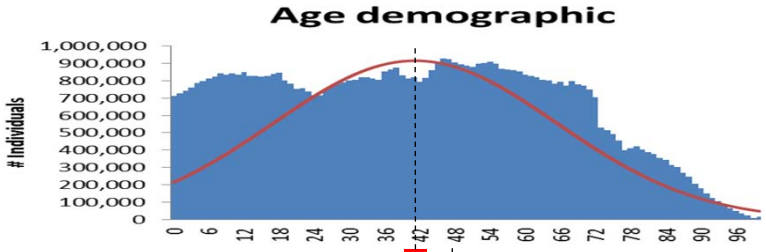
↗ Valider précision de s:

précise (n>30) (loi normale)

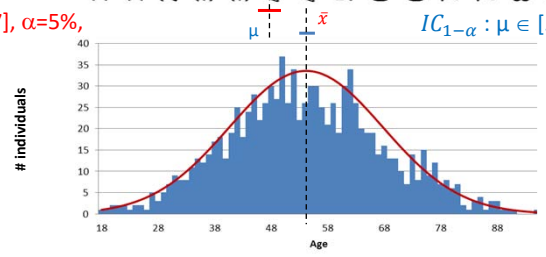
Peu- précise (n<30) (loi de Student)

Intervalle de pari vs. Intervalle de confiance

Démographie de la population française (Tableau 1) et une échantillon inconnu (Tableau 2):



$IP_{1-\alpha}$: $\bar{x} \in$ [39.7; 42.7], α=5%, n=1000



$IC_{1-\alpha}$: μ ∈ [53.3;55.0], α=5%, n=1000

Conclusion:

- μ ∉ $IC_{1-\alpha}$; Inférence: l'échantillon a été sélectionné au hasard dans une autre population avec μ ∈ [53.3;55.0] au risque 5%
- $\bar{x}$ ∉ $IP_{1-\alpha}$: Prédiction: l'échantillon n'a pas été sélectionné au hasard dans la population avec μ = 41.2 au risque 5%

Intervalle de pari vs. Intervalle de confiance

Démographie de la population française (Tableau 1) et une échantillon inconnu (Tableau 2):