Bonjour,

Nous allons aujourd'hui analyser un biais méconnu dans la théorie des sondages.

1 - Définition et importance

Un sondage c'est une technique qui permet de tirer des conclusions sur une population à partir d'un échantillon issu de celle-ci. Les sondages sont utilisés pour contrôler la qualité des pièces sortant d'une chaine de fabrication, pour contrôler la véracité des écritures comptables d'un dossier financier, pour tenter de mesurer l'efficacité d'un médicament ou le risque de tomber malade, pour obtenir des informations sur les goûts des consommateurs ou pour tirer des projections sur les tendances des électeurs vis-à-vis d'un ou plusieurs candidats. Le champ d'application est très large.

2 - Rappels du fondement scientifique des probabilités

La technique est basée entièrement sur la théorie des probabilités. Or cette dernière est elle-même fondée uniquement sur la situation qui suit et que nous allons rappeler à travers un exemple :

Soit un sondage d'opinions électorales avec une population de 40 millions d'électeurs et un échantillon issu "au hasard" de celle-ci d'un effectif de 400 sondés. On s'intéresse au fait que les sondés voteront ou non pour un candidat MARTIN. La théorie mathématique ne s'applique qu'à la première condition initiale : que les électeurs aient tous la même probabilité d'être sélectionnés dans l'échantillon. Précisons ici que les mathématiciens peuvent aussi étudier une situation où les éléments de la population initiale ne sont pas équiprobables en la ramenant à une situation d'équiprobabilité de la manière suivante : si par exemple une personne donnée a deux fois plus de chances d'être tirée que ses voisines, il suffira de considérer que cette personne particulière est présente en deux exemplaires dans la population. Les deux situations étant en effet équivalentes.

La théorie mathématique définit alors la probabilité d'un évènement donné comme le rapport entre le nombre de cas favorables sur celui du nombre de cas possibles.

Dans notre exemple, en supposant que la population contient 45% de MARTIN (personnes votant pour MARTIN), il est possible de calculer la probabilité d'obtenir un échantillon contenant lui aussi 45% de MARTIN, c'est-à-dire ici 180 MARTIN sur les 400 interrogés. Mais il est aussi possible de calculer la probabilité d'obtenir d'autres échantilons : ceux ayant moins de 45% de MARTIN et ceux ayant plus de 45% de MARTIN. On obtient alors l'histogramme de la figure n°1 :

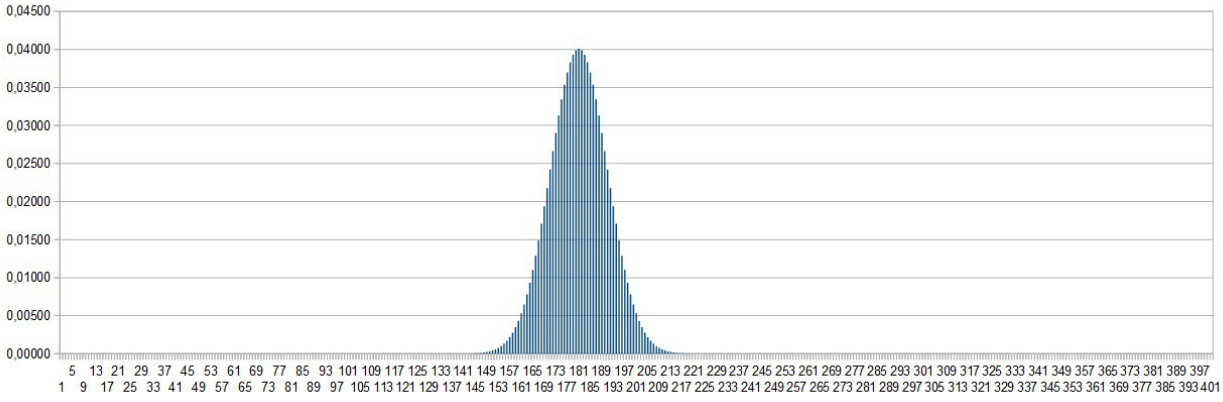


figure n°1 :
Chaque rectangle représente la probabilité d'avoir la proportion de MARTIN signalée en abcisse. En réalité, il s'agit d'une densité de probabilité car la probabilité elle-même n'est pas représentée par la hauteur du rectangle, donc la valeur figurant sur l'axe des ordonnées, mais par la surface du rectangle. NB: bien que représentée sous forme d'histogramme, les calculs ont été réalisés ici en utilisant la loi normale.

3 - Les biais bien connus

Le calcul exact des probabilités utiliserait une loi hypergéométrique. Elle est justifiée parce que les tirages sont exhaustifs : il n'est pas interrogé une même personne deux fois. Mais il est utilisé à la place une loi binomiale adaptée seulement pour les tirages non exhaustifs parce qu'elle est plus facile à calculer. L'erreur est minime pour des populations grandes et des échantillons petits devant celle de la population. Il est en effet peu probable d'interroger une même personne deux fois de suite quand on choisit au hasard 400 personnes parmi 40 millions. C'est une première appoximation.

La loi binomiale tend vers la loi de Gauss qui est une courbe continue. Il s'agit donc d'une situation limite, celle où la population devient infinie. Avec 40 millions, notre population donne des résultats voisins de l'infini. Il est donc valide d'effectuer cette seconde approximation.

Ces deux biais sont parfaitement maitrisés.

D'autres biais sont parfaitement connus mais moins bien maitrisés comme la difficulté de tirer les électeurs au hasard, la qualité de la question qui leur est posée qui peut être ambigüe, la qualité de leur réponse qui peut être influencée par la peur d'avouer un choix politique plutôt qu'un autre, etc. etc. Il est trivial de considérer qu'un sondage d'opinion est moins fiable que celui portant sur la qualité des pièces sortant d'une chaine de fabrication.

Mais maitrisés ou non, ces biais sont parfaitement connus. Y en a-t-il d'autres ?

4 - Une erreur de logique ?

Un sondage consiste donc à extraire un échantillon d'une population pour en tirer des informations sur cette population. L'opération logique en cause est l'implication :

J'ai telle information sur l'échantillon => j'en déduis telle propriété de la population

La meilleure approche de la logique mathématique se fait par la théorie des ensembles. La situation A => B est alors schématisée ainsi :

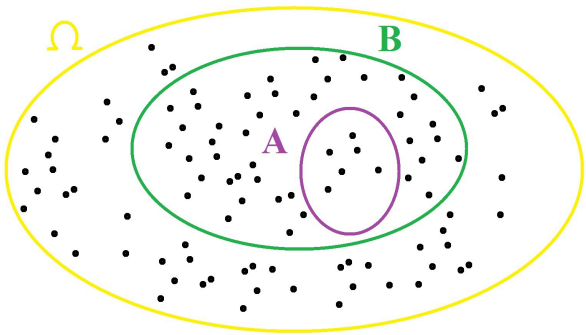


figure n°2
Chaque point noir représente un univers possible. Oméga, en jaune, représente tous les univers possibles. Tous les univers présents dans A sont aussi

présents dans B car A est inclus dans B. Donc "être dans A" => "être dans B"
Ici la taille des ensembles n'a pas d'importance. Ce qui est important c'est que tous les univers de A soient aussi dans B.

Mais pour présenter avec des ensembles la théorie des probabilités, il faudrait rendre importante la taille des ensembles. En effet, la probabilité d'avoir A, en tirant un univers au hasard dans l'ensemble Omega des univers possibles, se calcule en divisant le nombre d'univers de A par le nombre d'univers de Omega. De la même manière, sachant que B est réalisé, la probabilité de A est alors Card(A)/Card(B). Le cardinal d'un ensemble étant le nombre de ses éléments.

Pour tenir compte de ces considérations de taille, il y a deux manières.

On peut représenter précisément les univers individuels (en noir) comme fait sur la figure 3

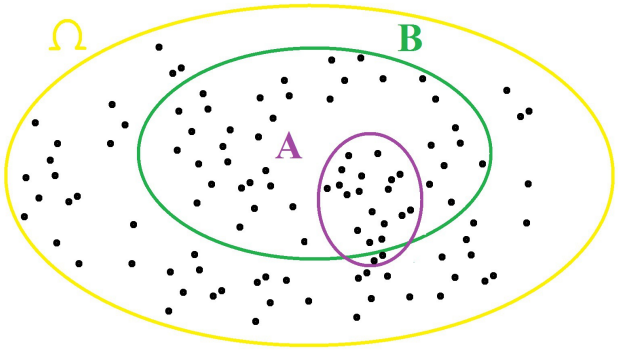


figure n°3
L'ensemble A contient 20 univers dont 2 seulement ne sont pas dans B. Ce schéma représente la situation où 90% des univers de A sont dans B donc celle où l'on peut affirmer avec la probabilité 90% de ne pas se tromper que si un univers est dans A alors il est aussi dans B. Ce que l'on peut noter ainsi :
A =>(90%) B

On peut aussi dessiner les ensembles de telle manière que leur taille soit proportionnelle à leur effectif. Une représentation avec des courbes est alors ambigüe pour l'esprit. Il est immédiat de constater qu'une représentation en surface serait meilleure et en conséquence, qu'une représentation utilisant des rectangles permettrait mieux à l'esprit de comprendre ce qui se passe.

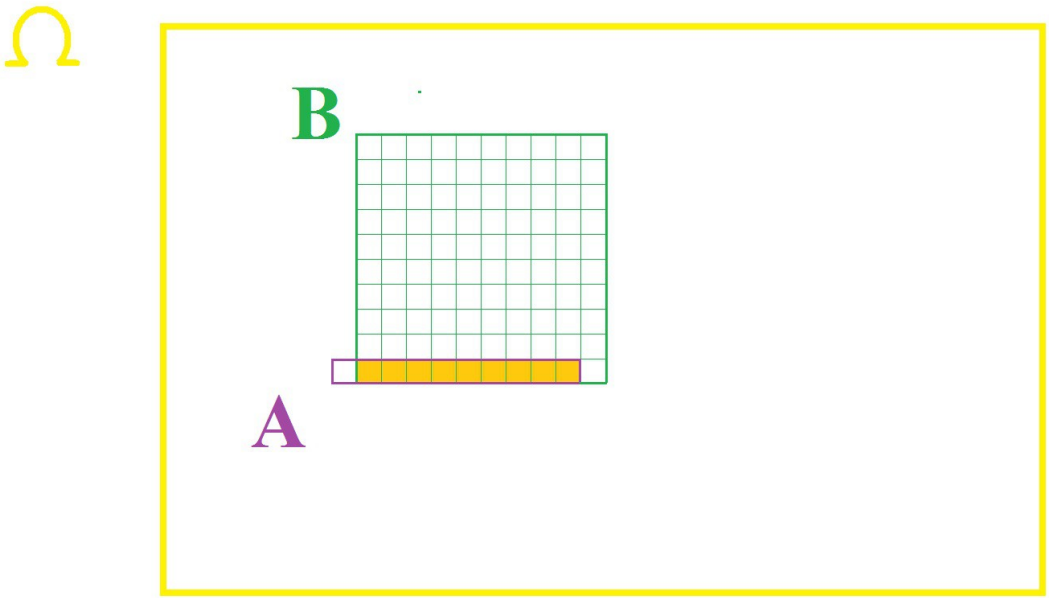


figure n°4
Les petits carrés permettent de visualiser les quantités. Ici l'ensemble A est aux 9/10èmes inclus dans B. On peut donc affirmer avec 90% de chances d'avoir raison que si un univers est dans A, alors il est aussi dans B. NB : Il importe peu que B soit petit ou grand tant qu'il est capable d'ingurgiter les 9/10ème de A.

Choisissons d'écrire la situation avec cette notation:

A =>(90%) B

Dans le cas de notre sondage :
A est l'ensemble des univers qui possèdent une population d'électeurs ayant une proportion de 45% votant pour MARTIN.
B est l'ensemble des univers qui donnent un échantillon de 45%.

Le problème est que la justification avancée généralement pour ne pas dire systématiquement par les techniciens des sondages est l'inverse. Ils choisissent un intervalle de confiance de 95% et démontrent en effet que :

Si une population contient 45% de MARTIN alors il y a 95% de chances que l'échantillon contienne entre 40% et 50% de MARTIN

Et en concluent que, ayant obtenu un échantillon de 45% alors la population de départ dispose d'une proportion de MARTIN comprise entre 40% et 50% avec une probabilité de 95% !

La justification présentée est donc la suivante :

A =>(95%) B => B =>(95%) A

Ce qui n'est absolument pas prouvé. Dans l'exemple représenté sur la figure n°4, B contient 100 cases dont 9 sont communes avec A. Si on sait qu'un univers appartient à B, la probabilité qu'il soit dans A n'est alors que de 9/100.

Les techniciens qui tentent de justifier leur pratique par le seul calcul précédant passent de:

J'ai telle propriété de la population => j'en déduis telle information sur l'échantillon

à

J'ai telle information sur l'échantillon => j'en déduis telle propriété de la population

Ce qui est totalement impossible du point de vue logique. Pourtant les mathématiciens ont abordé la question de l'inférence autrement dit des possibilités de déduire quelque chose sur la population à partir d'un échantillon tirée de celle-ci avec notamment les travaux de Thomas Bayes et les loi Beta mais il semble que tout cela soit ignoré. Malgré cela, la pratique des sondeurs n'est pas forcément fausse, ils peuvent avoir raison sur la conclusion mais c'est la justification qu'ils en donnent qui est fausse ou à minima incomplète.

5 - Complètons-là

Nous allons prendre ici un exemple plus petit. Cela nous permettra d'utiliser une loi hypergéométrique et donc de ne pas être concernés par les approximations que constituent la loi binomiale d'abord, la loi normale ensuite.

Soit une population qui contient 30 éléments. Ces éléments peuvent être classés en deux catégories : rouges ou non rouges.

Nous en tirons 10 au hasard et sans remise et nous nous intéressons au nombre de rouges dans le tirage. On ignore quel est le nombre de rouges dans la population de ces 30 éléments. Ce nombre peut donc varier de 0 à 30 soit 31 possibilités. Après avoir tiré au hasard 10 éléments, nous notons que 4 d'entre eux sont rouges. Si la population a la même proportion de rouge que notre échantillon, alors il y aurait 12 rouges parmi les 30 éléments de celle-ci. Mais ce n'est qu'une hypothèse.

Dans cette hypothèse, celle où la population contient 12 rouges parmi les 30, nous pouvons calculer exactement les probabilités pour chacun des tirages contenant respectivement 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 et 10 rouges. Ces probabilités se calculent de manière approchée à partir des loi binomiale ou de la loi normale mais elles se calculent de manière exacte avec une loi hypergéométrique selon la formule suivante:
N = 30 (nombre d'éléments dans la population)
R = 12 (nombre de rouges dans la population)
n =10 (nombre d'éléments tirés au hasard dans la population formant notre échantillon)
k = {0,..., 10} (diverses valeurs possibles du nombre de rouges dans l'échantillon)

$$\mathbb{P}(X = k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}$$

avec :

$$\binom{a}{b} = C_a^b = \frac{a\,!}{b!\,(a-b)\,!}$$

figure n°5

Ainsi, la probabilité pour que, la population contenant 12 rouges sur les 30, on trouve un échantillon de 4 rouges sur les 10 est de 0,31. Les chiffres présentés dans le tableau suivant (figure n°6) ont été arrondis à deux décimales ce qui explique les 0 des extrémités du tableau. On vérifie que la somme de ces probabilités donne 1.

	0	1	2	3	4	5	6	7	8	9	10
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0

figure n°6

Si on somme les probabilités pour k = 2, 3, 4, 5 et 6 on trouve 0,95% (figure n°7):

	0	1	2	3	4	5	6	7	8	9	10
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
			0,95								

figure n°7

On peut donc affirmer que, si on a une population de 30 éléments parmi lesquels 12 sont rouges,

alors il ya une probabilité de 95% pour qu'un échantillon issu de cette population tiré au hasard donne un résultat compris entre 2 et 6 (2 et 6 compris) ce qui correspond à des proportions de 20% et 60%.

Mais, nous l'avons vu, il est impossible à ce stade d'inverser la proposition logique et d'affirmer que, puisque nous avons un échantillon de 4/10 (soit 40% de rouges) alors la proportion de la population est compise entre 20% (soit 6/30) et 60% (soit 18/30).

Il est même impossible de conclure quoi que ce soit sur la population à l'exception du fait qu'elle contenait avant le tirage au moins 4 rouges, puisqu'ils sont dans l'échantillon, et au plus 24 car il y a 6 éléments dans l'échantillon qui ne sont pas rouges.

Jusque-là, nous avons pratiqué comme les techniciens de sondage, ou au moins comme ils nous justifient leur pratique. Mais allons plus loin.

Nous l'avons dit, les proportions de rouges dans la population peuvent aller de 0/30 à 30/30. Il y a donc 31 possibilités. Pour chacune d'elles, il est possible de calculer les différentes probabilités pour k = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 et 10 comme nous l'avons fait pour la proportion 12/30 (soit 40%). Cela donne le tableau suivant (figure n°8):

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,06	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,03	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,02	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,01	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82

figure n°8

En haut, horizontalement sur fond orange, les valeurs de k de 0 à 10
A gauche, verticalement sur fond magenta, les valeurs de R (soit le nombre de rouges dans la population)
En vert horizontalement, les probabilités que nous avons calculées plus haut concernant le cas de la population ayant 12/30 (40%) de rouges.
Les autres probabilités sont sur fond gris.
Chaque ligne correspond donc à une hypothèse donnée sur la proportion de rouges dans la population. On vérifie que la somme des probabilités d'une même ligne donne bien 1 comme nous l'avons fait précédemment pour la ligne verte.
De ce fait, la somme de toutes les lignes doit donner 31 x 1 soit 31. En calculant les sommes de probabilités verticalement (qui donnent toutes 2,82) on peut vérifier la cohérence d'ensemble. En effet la somme des 11 cases contenant 2,82 donne bien 31.

Maintenant, nous avons bien tous les cas possibles. La réalité que nous cherchons est obligatoirement représentée sur notre tableau puisque celui-ci répertorie les 31 cas de populations possibles. Dans chaque cas, il y a diverses possibilités d'extraire un échantillon de ces populations et toutes les possibilités sont bien présentes.

Mais nous avons effectué un tirage et nous avons un échantillon qui donne 4/10 (40%) de rouges. Nous savons donc que nous sommes dans la colonne correspondante, en bleu sur le tableau de la figure n°9 :

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,06	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,03	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,02	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,01	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82

figure n°9

L'univers des possibles est maintenant réduit à la zone bleue. On voit que le total de toutes les probabilités fait 2,82 or il doit faire 1. Ce n'est qu'une question d'unité de mesure. Rappelons-nous que ces probabilités sont en fait des surfaces qui représentes soit les cas possibles soit les cas favorables.
Donc ici, les cas possibles sont représentés par la surface totale des cases bleues. On parle de la totalité chiffrée et pas la taille des cases qui sont toutes identiques. On a donc un total de 2,82.
Regardons la surface que représente les proportions dans la population comprises entre 20% (soit 6/30) et 60% (soit 18/30) :

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,06	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,03	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,02	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,01	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82	2,82

figre n°10
Les cases représentant les populations comprises entre 20% (soit 6/30) et 60% (soit 18/30) sont sur fond bleu foncé. Le total est de 2,66.

Donc pour calculer la probabilité d'avoir une population avec une proportion de rouges comprise entre 20% (soit 6/30) et 60% (soit 18/30), il faut faire le rapport des cas favorables, ici 2,66 sur les cas possibles soit 2,82. Le résultat est de 0,94%.

Il est plutôt proche des résultats des techniciens sondeurs.

Voyons l'explication graphiquement.

Les propabilités pour la population 12/30 données dans le tableau de la figure n°6 peuvent se représenter graphiquement ainsi:

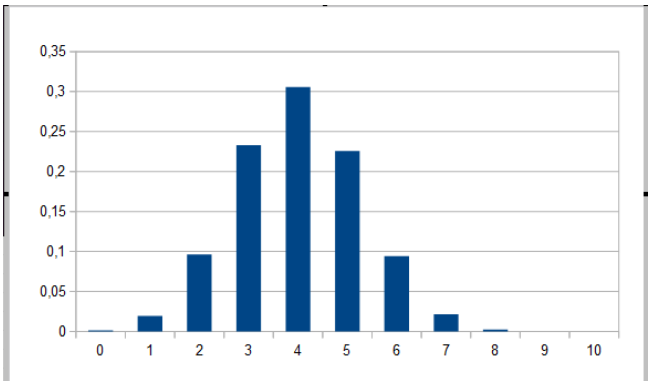


figure n°11
L'échantillon le plus probable est celui contenant la proportion la plus proche de la population qui est de 12/30 (40%), c'est donc l'échantillon avec 4/10 (40%).

Si on examine maintenant le graphique de la proportion la plus proche de 4/10 dans les échantillons, c'est 3/10 (30%). Cela correspond à une population qui serait elle aussi de 30% soit 9/30. La voici représentée sous forme d'histogramme:

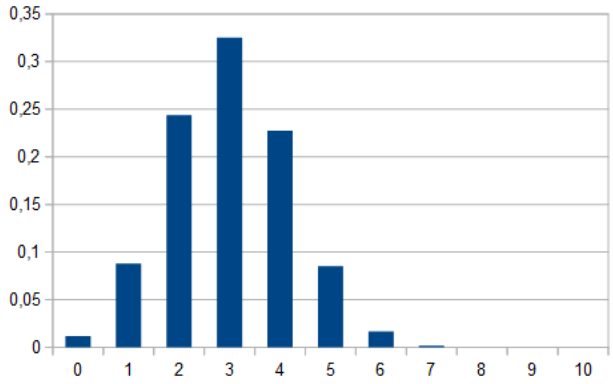


figure n°12

On voit que la forme est très similaire à la précédente. Elle semble simplement décalée vers la gauche. Superposons-les comme réalisé sur la figure n°13:

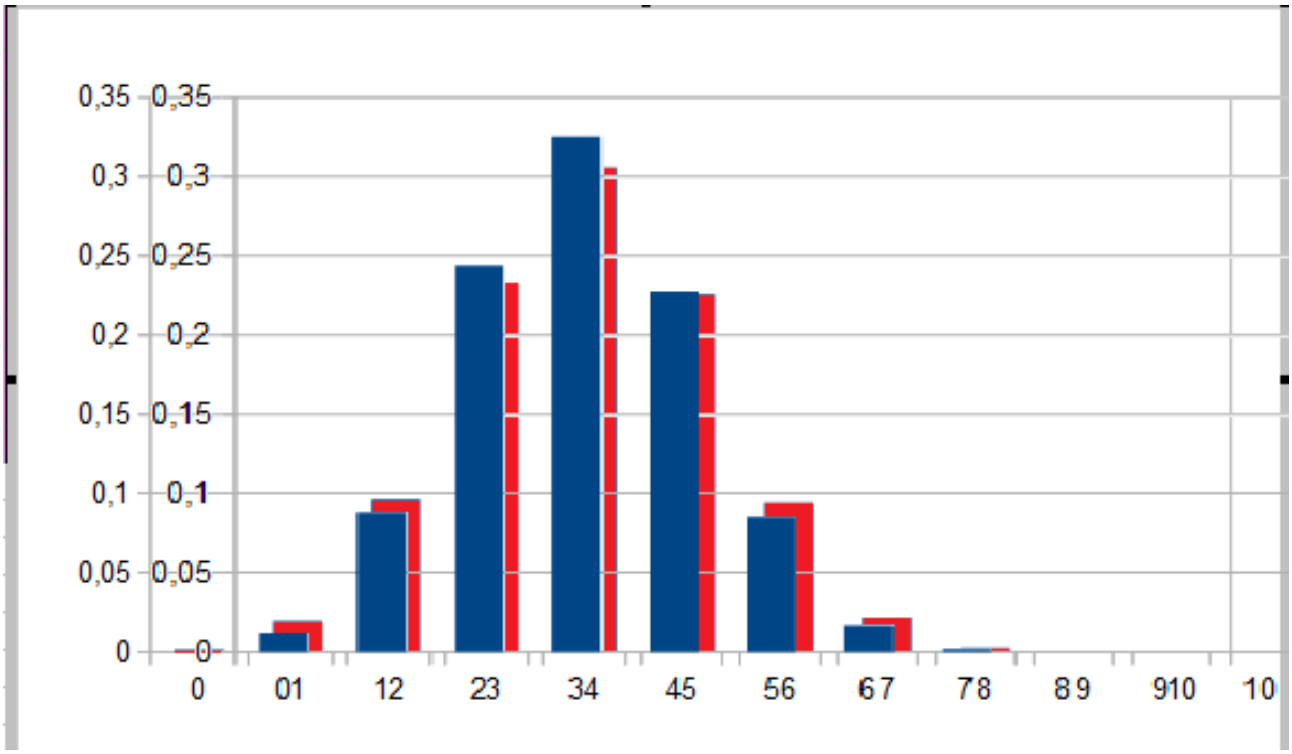


figure n°13
Superposition des deux représentations bleue pour la population à 30% et rouge pour celle à 40%. On voit que les graphiques se superposent assez précisément. Il a fallu bien entendu décaler le graphique bleu ce qui fait que le sommet de cette population (9/30), obtenue pour une abcisse de 3 est maintenant superposée au 4 de la population en rouge (12/30).

Remettons-les à peu près à leur place, "à peu près" car on souhaite voir ce qui se passe derrière.

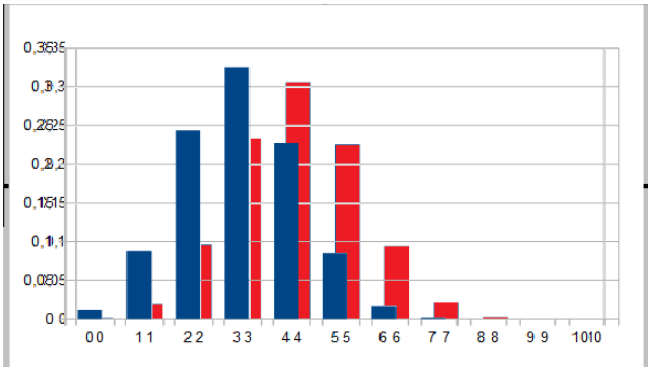


figure n°14

On voit maintenant que si l'on désire calculer la probabilité ou le nombre de tirages qui donne 4/10 dans cette population de 9/30 représentée en bleu, il est possible de se référer au rectangle qui représentait 5/10 pour la population contenant 12/30 rouges:

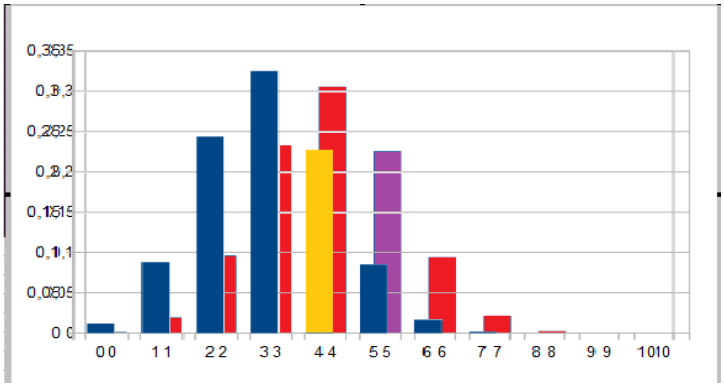
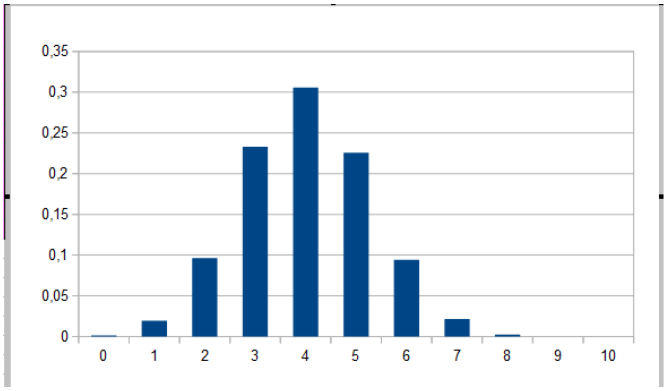


figure n°15

Les deux valeurs représentées en jaune (qui appartenait au graphique bleu) et celle représentée en magenta (qui appartenait au graphique rouge) sont à peu près égales. Et c'est aussi vrai pour les valeurs suivantes.

Certes, sur ces graphiques, les valeurs sont proches mais légèrement différentes et c'est visible. En revanche, si on avait utilisé un exemple plus important en taille comme par exemple une population de 1 000 000 éléments avec un tirage de 1 000, nous n'aurions pas eu d'écart visible ni sur le graphique ni probablement dans les nombres sauf à utiliser un nombre invraisemblable de décimales.



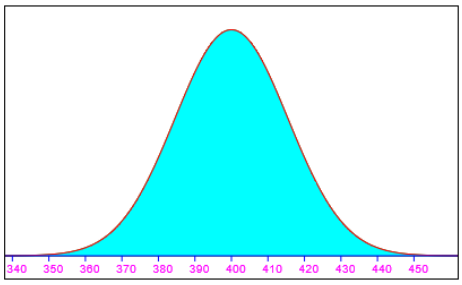


figure n°16

A gauche notre distribution pour un tirage de 10 éléments parmi n dans une population de 30 éléments contenant 40% (donc 12/30) rouges

A droite la distribution normale correspondant à la même proportion de 40% mais avec des effectif supérieurs: 1 000 éléments tirés d'une population infinie avec 40% de rouges.

La précision est nettement meilleure puisque les probabilités s'étalent entre 350 et 450 soit les nombre 3,5 et 4,5 de notre cas. De plus la courbe est nettement symétrique.

En contractant la courbe pour faire coincider les graduations en X (pas en Y) on voit que l'on est très loin du bord autrement dit des cas où les échantillons auraient 0 ou 100% de rouges.

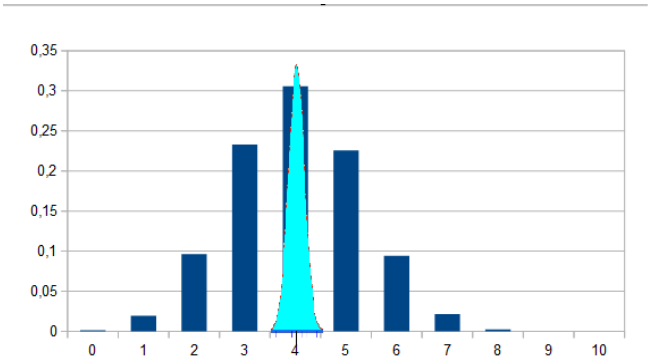
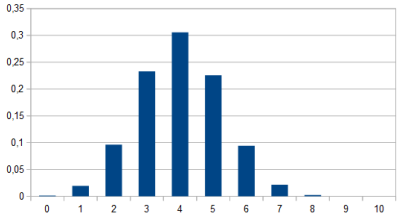
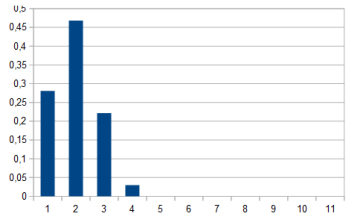


figure n°17

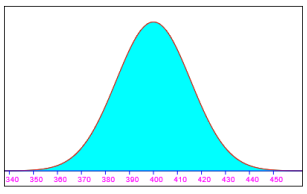
De ce fait, l'effet de bord qui va déformer la courbe jusqu'à la rendre compètement asymétrique n'est pas visible avec de gros effectifs.



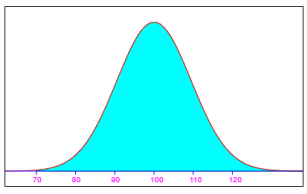
N = 30 et R = 12 p = 40%



N = 30 et R = 3 p = 10%



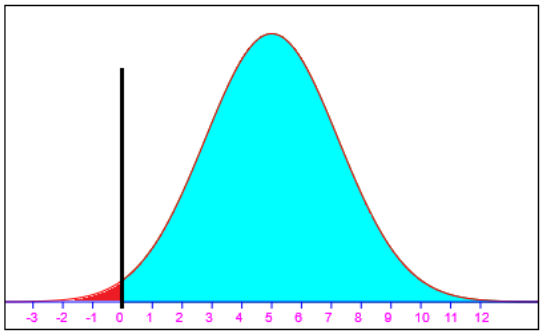
N = infini p = 40%



N = infini et p = 10%

figure n°18

L'effet de bord disparait pour la plupart des distributions quand on augmente la précision donc les nombres de tirages.
En haut avec de petits effectifs:
A gauche la distribution que nous avons observée
A droite, celle qui correspond à une distribution plus proche du bors à savoir 3/30 (10%)
En bas avec de gros effectifs mais les mêmes proportions de rouges que ci-dessus:
A gauche pour 40%
A droite pour 10%
Avec des effectifs élevés, 10% n'est pas suffisamment petit pour voir l'effet de bord apparaitre. Il faudrait des proportions bien plus proches de 0%.



N = infini p = 1%

figure n°19
Avec une proportion de 1%, la loi normale ne peut plus être utilisée car elle déborde sur les négatifs.

Tout ceci, c'est-à-dire la validité de considérer les distributions des populations voisines comme identiques à celle qui correspond à notre échantillon (notre superposition) et l'absence d'effet de bord pour des valeurs éloignées des extrêmités (notre symétrie), semble donner enfin la justification des techniciens sondeurs, celle qui n'apparait jamais ou en tous les cas suffisamment rarement pour que nous ne l'ayons pas trouvée.

6 - Y a-t-il confirmation ou non de la justification de la pratique des sondeurs ?

Oui... mais non.

En fait, cette explication n'est valable que dans un cas très précis. En effet, pour élaborer cette démonstration, il nous a fallu faire une hypothèse suppémentaire, celle que les 31 populations ou plus précisément les 31 proportions possibles étaient équiprobables. Et cette hypothèse, les techniciens n'en parlent pas. Ils font comme s'ils l'ignoraient et de ce fait, il ne vérifient pas que cette hypothèse est plausible, compatible avec la réalité qu'ils comptent apréhender avec l'outil mathématique.

Voyons donc ce que cela change.

La fourchette des techniciens ayant pour eux 95% de chance de contenir la proportion réelle de rouges dans la population s'étale de 20% à 60%. Multiplions par un coefficient les probabilités des populations qui sortent de cet intervalle de confiance par exemple celles qui sont supérieures à 60% donc les populations qui ont 19 et plus rouges sur 30 éléments. Avec un coefficient de 2, le total de 2,82 se transforme en 2,94 mais comme nous n'avons pas appliqué le coefficient à la zone en bleu foncé, le nombre 2,66 ne change pas.

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,12	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,07	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,03	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,01	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	2,94	2,82	2,82	2,82	2,82	2,82	2,82

figure n°20

Dans ce cas, la probabilité des populations entre 20% et 60% n'est plus la même puisqu'elle est maintenant de 2,66/2,94 soit 0,91.

NB: les nombres 2,66 et 2,94 sont arrondis mais le rapport a été calculé avec les nombres précis.

On voit donc que l'intervalle de confiance n'est déjà plus le même puisqu'il passe de 0,94% à 0,91%.

Poursuivons. Augmentons le coefficient. En multipliant ces mêmes probabilités par 3 (figure 21) ou 4 (figure 22) on obtient les chiffres suivants :

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,18	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,1	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,05	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,02	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0,01	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	3,06	2,82	2,82	2,82	2,82	2,82	2,82

figure n°21

	0	1	2	3	4	5	6	7	8	9	10
0	1	0	0	0	0	0	0	0	0	0	0
1	0,67	0,33	0	0	0	0	0	0	0	0	0
2	0,44	0,46	0,1	0	0	0	0	0	0	0	0
3	0,28	0,47	0,22	0,03	0	0	0	0	0	0	0
4	0,18	0,42	0,31	0,09	0,01	0	0	0	0	0	0
5	0,11	0,34	0,36	0,16	0,03	0	0	0	0	0	0
6	0,07	0,26	0,37	0,23	0,07	0,01	0	0	0	0	0
7	0,04	0,19	0,34	0,29	0,12	0,02	0	0	0	0	0
8	0,02	0,13	0,3	0,32	0,17	0,05	0,01	0	0	0	0
9	0,01	0,09	0,24	0,33	0,23	0,09	0,02	0	0	0	0
10	0,01	0,06	0,19	0,31	0,27	0,13	0,03	0	0	0	0
11	0	0,03	0,14	0,28	0,3	0,18	0,06	0,01	0	0	0
12	0	0,02	0,1	0,23	0,31	0,23	0,09	0,02	0	0	0
13	0	0,01	0,06	0,19	0,29	0,27	0,14	0,04	0,01	0	0
14	0	0,01	0,04	0,14	0,27	0,29	0,18	0,06	0,01	0	0
15	0	0	0,02	0,1	0,23	0,3	0,23	0,1	0,02	0	0
16	0	0	0,01	0,06	0,18	0,29	0,27	0,14	0,04	0,01	0
17	0	0	0,01	0,04	0,14	0,27	0,29	0,19	0,06	0,01	0
18	0	0	0	0,02	0,09	0,23	0,31	0,23	0,1	0,02	0
19	0	0	0	0,01	0,24	0,18	0,3	0,28	0,14	0,03	0
20	0	0	0	0	0,14	0,13	0,27	0,31	0,19	0,06	0,01
21	0	0	0	0	0,07	0,09	0,23	0,33	0,24	0,09	0,01
22	0	0	0	0	0,03	0,05	0,17	0,32	0,3	0,13	0,02
23	0	0	0	0	0,01	0,02	0,12	0,29	0,34	0,19	0,04
24	0	0	0	0	0	0,01	0,07	0,23	0,37	0,26	0,07
25	0	0	0	0	0	0	0,03	0,16	0,36	0,34	0,11
26	0	0	0	0	0	0	0,01	0,09	0,31	0,42	0,18
27	0	0	0	0	0	0	0	0,03	0,22	0,47	0,28
28	0	0	0	0	0	0	0	0	0,1	0,46	0,44
29	0	0	0	0	0	0	0	0	0	0,33	0,67
30	0	0	0	0	0	0	0	0	0	0	1
	2,82	2,82	2,82	2,82	3,18	2,82	2,82	2,82	2,82	2,82	2,82

figure n°22

Si on progresse on obtient un intervalle dans la confiance diminue. Voici un tableau recuillant quelques valeurs (figure 23).

multiplicateur	intervalle de confiance
2	0,91
3	0,87
4	0,84
10	0,68
20	0,52
50	0,31
100	0,18

figure n°23

Donc pour peu que ces populations soit nettement plus probables que celles dans l'intervalle de confiance des sondeurs, la qualité de cet intervalle diminue. En effet, avec le multiplicateur 10, il n'y a plus que 2 chances sur 3 (0,68) de se trouver dans l'intervalle présenté et avec 100, il ne reste que 18 chances sur 100 !

En poussant à l'extrême, imaginons la situation ou les probabilités de l'intervalle 20% - 60% soient très faibles par rapport aux valeurs extérieures voire nulles. On aurait donc une situation où il est impossible que la proportion de la population soit comprise entre 20 et 60% et en même temps probable à 95% qu'elle y soit !!!!

On pourrait penser que *"oui, mais bon. comme on ne connait pas les probablités de chacune des 31 populations, on n'a donc aucune raison de privilégier l'une plutôt que l'autre et que donc il est légitime de les considérer comme équiprobables"*. Autrement dit on affirmerait ici que le manque de connaissance d'une situation est équivalente à l'équiprobabilité.

Ce n'est pas le cas.

Prenons le cas d'une urne contenant des billes vertes et rouges. Dans le cas 1, on connait la proportion de billes rouges et vertes : 50%-50%. Dans le cas 2, on ignore complètement la proportion.

Dans le cas 1, le calcul probabiliste vous dit qu'il est égal pour vous de vous attendre à une verte ou à une rouge et qu'il ne peut vous conseiller un choix plutôt que l'autre.
Dans le cas 2, le calcul probabiliste ne peut être fait, il ne vous donne donc aucun conseil.

Dans le cas 1, si vous gagnez 30 euros si vous tirez une bille de la couleur que vous avez choisie mais que vous perdez 20 euros si vous vous êtes trompés, le calcul vous dira que vous ave intérêt à jouer.
Dans le cas 2, avec les mêmes gains, les mathématiques ne vous diront rien.

Dans le cas 1, s'il existe des billes vert-clair et des billes vert-foncé, cela ne changera rien.
Dans le cas 2, dans la même situation, si vous souhaitez dire que l'équiprobabilité est valide quand on ignore tout de la proportion dans la population, vous serez amenés à évoquer à nouveau l'équiprobabilité et vous serez alors amenés à considérer 1/3 de rouges, 1/3 de vert-clair et 1/3 de vert-foncé. Il y aura donc maintenant pour vous 1/3 de rouges seulement. Cela n'a pas de sens car la situation n'a pas changé, les billes sont les mêmes, simplement vous savez maintenant qu'il y des billes vertes qui sont différentes mais la proportion réelle n'a pas changé.

Autre exemple que nous avons traité dans d'autre travaux. Celui d'un candidat à un courcours pour un poste de professeur de mathématiques dans une école. Chaque année 100 personnes se présentent à un test et il y a chaque année 10 reçus. Sur ces 10, l'école en choisit 8 au hasard pour occuper les 8 postes disponibles. Il s'agit donc bien de sondages. Un candidat peut donc estimer ses chances à 8/100 avec un intervalle de confiance qu'il calculera comme précédemment. Oui, mais si, alors même que la réalité de ce concours ne change pas, il apprend que le test est un examen de mathématiques alors qu'il sait que, contrairement aux autres candidats, il est lui-même diplômé dans cette matière, il est donc presque sûr de réussir le test. Il a donc maintenant une probabilité d'avoir un poste proche de 8/10. Et rappelons-le, la réalité n'a pas changé, il a juste pris connaissance d'une information qu'il n'avait pas : celle que les candidats ne sont pas équiprobables. Il n'était donc nullement fondé de calculer sa première probabilité d'une chance sur 100. Et il ne l'est d'ailleurs pas plus sur sa seconde estimation sauf s'il peut établir qu'il est bien maintenant dans une situation d'équiprobabilité au niveau du tirage au sort choisissant les 8 gagnants parmi les 10 candidats réussissant le test.

Donc, si vous ne pouvez choisir une situation d'équiprobabilité pour la seule raison que vous n'avez aucun motif de privilégier une situation plutôt qu'une autre, vous ne pouvez considérer que vous avez équiprobabilité **QUE** si vous avez une bonne raison de le faire. Et rappelons que, dans un modèle déterministe, cette situation n'est acquise que dans le cadre d'un très grand nombre de facteurs qui peuvent se compenser comme se conforter sans privilégier un sens ou un autre. Il est donc impératif, avant d'appliquer le modème statistique, de vérifier cette situation.

7 - Synthèse

Concrètement, si on vous demande de tester une machine de fabrication de certaines pièces, que vous disposez de 1000 pièces choisies au hasard sur les 100 000 que cette machine a produit jusque là, que vous trouvez une proportion de 400 pièces défectueuses, vous ne pourrez extrapoler comme le font les sondages, que si vous arrivez à justifier que toutes les porportions sont équiprobables c'est-à-dire que la machine aurait tout aussi probablement produire 100% de pièces défectueuses

que 20% ou 0%.

Pour un sondage d'opinion, en matière électorale par exemple, sans même considérer les biais connus, il vous faudra justifier que toutes les répartitions possibles de la population sont équiprobables. C'est-à-dire qu'il y a par exemple autant de chances qu'il y ait 0% de gens votant pour le candidat MARTIN que 100%. Il est assez évident que cela n'est jamais le cas. Et pourtant, faute de cette situation d'équiprobabilité, votre intervalle de confiance ne vaudra rien. Au moins d'un point de vue scientifique.

Il semble bien que les mathématiciens ayant développé le domaine des statistiques et probabilités aient eu tellement envie de pouvoir estimer une population à partir d'un échantillon qu'ils ont fait fi du résultat pourtant évident à savoir que les résultats établis dans cette discipline scientifique ne fonctionnent, pour le moment, que dans un sens : connaissant une population, je peux calculer les probabilités des tirages issus de celle-ci. C'est ce qui explique que les probabilités soient si efficaces dans le domaine des jeux de hasard. Connaissant la proportion des cartes d'un jeu, je peux calculer la probabilité d'avoir telle main ou telle autre.

Mais pas l'inverse.

8 - En conclusion

Pour effectuer valablement une extrapolation sur la population à partir d'un échantillon, il est nécessaire de connaitre, ou au moins de supposer, la répartition en probabilité des différentes populations ou des différentes proportions dans la population.

S'il y a équiprobabilité, alors la démarche des sondeurs telle qu'ils la présentent est valable même si la manière dont ils la justifie est critiquable.

Si l'équiprobabilité n'est que locale, c'est-à-dire au voisinage de moyenne de l'échantillon, il est encore possible d'accepter la démarche des sondeurs mais en s'assurant qu'en dehors de la zone locale les probabilités ne sont pas démesurées comme vu dans notre simulation avec coefficients.

S'il y a une autre répartition que l'équiprobabilité alors il est nécessaire d'en tenir compte pour calculer l'intervalle de confiance.

Dans les autres cas, donc quand on n'a aucune information sur la répartition en probabilités des différentes proportions possibles, aucun calcul valable n'est possible. Cela reviendrait à multiplier le nombre de pommes contenues dans un panier par le nombre de cerises contenues dans ce même panier pour obtenir le nombre de fruits total du panier. Il n'échappera à personne que parfois, ce calcul tombe juste.

Sources et outils :

https://irem.univ-reunion.fr/spip.php?article657