Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Etudes épidémiologiques randomisées :

Erreurs conceptuelles, erreurs méthodologiques

par Bernard Beauzamy

septembre 2020

Résumé opérationnel

Nous mettons en évidence des erreurs conceptuelles majeures, intervenant dans la définition même des études épidémiologiques randomisées. Elles proviennent du fait que les praticiens connaissent mal les lois fondamentales des probabilités, et confondent une somme et une moyenne.

Si deux sous-groupes sont constitués de manière aléatoire, à partir d'un groupe de N personnes, les effectifs des sous-groupes ne tendront pas à être égaux, mais différeront en général de $\sqrt{N}$, quantité faible devant N, mais qui tend vers l'infini avec N. De même, les profils des sous-groupes ne seront pas identiques, ni pour l'âge, ni pour quelque critère que ce soit.

De manière générale, il est malsain de vouloir laisser le hasard régler un problème que l'on ne sait pas résoudre. La pratique de "randomisation" ne permet pas d'obtenir l'égalité des profils, comme les épidémiologistes l'espèrent, et elle est à bannir de manière générale. On préférera les choix faits à partir de la constitution d'histogrammes, comme expliqué ici.

1. Les lois du hasard

Prenons un jeu de pile ou face : A et B jouent, et celui qui gagne reçoit un Euro de celui qui perd. Le jeu est honnête et équilibré : le résultat de la $n^{ème}$ partie est décrit par une variable aléatoire X_n qui vaut ± 1, avec probabilité $1/2$ dans chaque cas. La somme $S_N = \sum_{n=1}^{N} X_n$ indique le résultat cumulé sur N parties, c'est-à-dire le gain cumulé du joueur A (c'est l'opposé pour B). Ce gain cumulé peut prendre des valeurs entre $+N$ (cas où A gagne tout le temps) et $-N$ (cas où A perd tout le temps).

Ce que dit la loi forte des grands nombres, c'est que la moyenne $\frac{1}{N}\sum_{n=1}^{N} X_n$ tend vers 0 lorsque le nombre de parties augmente. Mais, contrairement à ce que l'on croit, le gain cumulé de chaque joueur oscille : viendra un moment où A aura gagné un million d'Euros, viendra un moment où A aura perdu un milliard d'Euros. Les règles quantitatives sont connues ; elles sont pour l'essentiel dues à Khintchine (voir aussi le livre de l'auteur [SRW]).

Les erreurs conceptuelles que l'on rencontre souvent dans les études épidémiologiques randomisées proviennent du fait que les auteurs confondent la somme et la moyenne.

2. Découper un groupe en deux

Supposons que nous disposions d'un groupe de N personnes, que nous voulons diviser en deux sous-groupes égaux : au premier, nous donnerons un médicament, au second un placebo. On voudrait que les deux sous-groupes aient le même nombre de personnes, mais en outre qu'ils présentent le même profil : même nombre de jeunes, même nombre de personnes en surpoids, etc.

Il est très difficile de le faire de manière complètement déterministe. Par exemple, si nous séparons la France en deux régions, Nord et Sud, en affectant au premier groupe ceux qui viennent du Nord et au second groupe ceux qui viennent du Sud, il est vraisemblable que les effectifs ne seront pas égaux, et il peut y avoir en outre des différences dans les modes de vie.

L'approche déterministe la plus acceptable serait de les ranger par ordre alphabétique et d'affecter les $N/2$ premiers au groupe A et les $N/2$ derniers au groupe B (peu importe ici que N soit pair ou non). On se dit en effet que l'ordre alphabétique n'influe ni sur l'âge, ni sur le poids, etc.

On peut se demander : peut-on faire cette répartition de manière complètement aléatoire, sans aucune intervention humaine, si bien qu'on ne pourra pas nous accuser de "tricher" ? La réponse est évidemment oui, mais ce n'est pas aussi simple qu'on pourrait le croire.

L'idée la plus immédiate est : faisons un tirage de pile ou face pour chaque personne ; si on obtient pile, on envoie la personne dans le groupe A, si on trouve face, on l'envoie dans le groupe B. Le choix est complètement aléatoire et aucune tricherie n'apparaît. Le problème est que, constitués de cette manière, les deux groupes ne seront pas égaux.

En effet, le cardinal de A, c'est-à-dire le nombre de personnes constituant ce groupe, est $S_N = \sum_1^N Y_n$, où $Y_n = 1$ si le $n^{ème}$ tirage donne pile et 0 s'il donne face ; ce n'est pas exactement la même chose que ± 1. La somme S_N est bien le nombre de fois où l'on a obtenu pile. En général, S_N prend des valeurs qui dépassent $\sqrt{N}$. Autrement dit, si j'ai à répartir un groupe de 10 000 personnes et que je procède par tirages répétés de pile ou face, en général les deux sous-groupes ainsi constitués auront des effectifs différents d'environ 100 personnes. Cette différence d'effectif peut altérer les résultats de l'étude.

Comment tirer un sous-groupe au hasard ? L'idée est simple : On écrit les noms sur des bouts de papier, on met ces papiers dans une urne et quelqu'un, à l'aveugle, extrait $N/2$ papiers. On peut bien sûr automatiser ceci : on affecte à chaque personne un numéro d'ordre, de 1 à N, et on demande à Excel de tirer un numéro au hasard entre ces bornes. On élimine celui-là, on renumérote de 1 à $N-1$, et on recommence.

La fonction Excel à utiliser s'appelle RANDBETWEEN(bottom, top), avec ici bottom = 1, top = N, puis top=N-1, etc. La question n'est pas de savoir si le générateur Excel est parfait : il ne l'est pas, mais il est bien suffisant pour ce type de tâche.

3. Profils des sous-groupes

Supposons qu'un groupe de $N = 10\,000$ personnes ait été diagnostiqué positif au "Covid-19". La question est de comparer l'efficacité de divers médicaments, entre eux ou par rapport à un placebo (qui est supposé n'avoir aucune action).

Comme expliqué au paragraphe précédent, on constitue aléatoirement deux sous-groupes, notés A et B, chacun de $M = N/2 = 5\,000$ personnes. On donnera le médicament aux personnes du sous-groupe A et le placebo aux personnes du sous-groupe B et, à la fin, on verra dans quel sous-groupe les résultats sont les meilleurs (ou le nombre de décès est le plus faible).

La difficulté tient au fait que les sous-groupes A, B peuvent différer de multiples manières : par exemple, il peut y avoir plus de personnes âgées dans le sous-groupe A, ce qui risque de fausser les résultats. Les épidémiologistes aimeraient que les deux sous-groupes aient le même "profil", sauf pour le fait que l'un reçoit le médicament et pas l'autre.

Les épidémiologistes croient que, du fait que la séparation en deux sous-groupes a été faite de manière aléatoire, les profils des deux sous-groupes seront identiques et identiques au profil du groupe tout entier, et ce d'autant plus que l'effectif total N est plus élevé. Mais c'est là que se situe l'erreur fondamentale. Bien au contraire, plus N est grand et plus les profils des sous-groupes seront différents.

Pour bien faire comprendre ceci, restreignons-nous à un seul critère, par exemple celui de l'âge, et restreignons-nous encore à une distinction binaire : être jeune (moins de 40 ans) ou être vieux (plus de 40 ans) ; admettons ici pour simplifier que la médiane d'âge du groupe tout entier est à 40 ans (autant de personnes au-dessous, autant au-dessus).

Nous avons donc les nombres JA, VA, JB, VB qui dénotent respectivement le nombre de jeunes et de vieux dans les sous-groupes A, B. Bien entendu, par définition $JA+VA=M,\ JB+VB=M.$

L'idée incorrecte consiste à croire que, du fait de la randomisation, les deux sous-groupes vont avoir le même profil en termes d'âge, et en particulier que $JA \approx JB$ et $VA \approx VB,$ la proximité étant d'autant meilleure que N est plus grand. Si, dans le groupe initial, on avait autant de jeunes que de vieux, on croit qu'il doit en être de même dans chacun des sous-groupes. Or ceci est radicalement faux, et la différence d'effectif ne fait que croître avec $N,$ comme nous allons le voir.

La loi des grands nombres assure que la moyenne des âges des sous-groupes est proche de la moyenne des âges du groupe tout entier, et ce d'autant plus que N est grand, et par conséquent les moyennes des âges des deux sous-groupes vont être peu différentes entre elles. Mais ceci n'est d'aucun intérêt pour le traitement : on voudrait être sûr que le nombre de jeunes et celui de vieux sont sensiblement les mêmes d'un sous-groupe à l'autre, et non pas seulement la moyenne. En langage mathématique, il y a une différence essentielle entre la somme $S_N = \sum_{n=1}^{N} X_n$ et la moyenne $E_N = \frac{1}{N}\sum_{n=1}^{N} X_n.$

Continuons avec la présentation mathématique simplifiée de notre exemple. Le groupe comporte $N = 10\,000$ personnes et autant de jeunes que de vieux. On notera X_n la variable aléatoire qui indique si la $n^{ème}$ personne est jeune ou vieille : si elle est jeune $X_n = 0,$ si elle est vieille $X_n = 1$; on est en présence d'une suite de variables indépendantes de même loi : $P(X_n = 0) = P(X_n = 1) = \frac{1}{2}.$

Le sous-groupe A est constitué des M premières personnes et le sous-groupe B des suivantes. La quantité $VA = \sum_{n=1}^{M} X_n$ représente le nombre de vieux dans le premier sous-groupe, puisqu'on compte 1 à chaque fois qu'une personne est vieille. De même, $VB = \sum_{n=M+1}^{N} X_n$ représente le nombre de vieux dans le second sous-groupe.

La loi de probabilité de la variable VA est facile à obtenir : il s'agit d'une loi binomiale de paramètres $\left(M, \frac{1}{2}\right)$ et de même pour $VB.$ Cela veut dire que, pour tout $k = 0,...,M$:

$$P(VA = k) = \binom{M}{k}\frac{1}{2^M}$$

Notons la variable aléatoire $Z = VA - VB$, la différence du nombre de vieux entre les deux sous-groupes : c'est elle qui nous intéresse, car elle caractérise la différence de profil entre eux. La loi de Z est facile à établir à partir des lois de VA, VB :

$$P(Z=d)=\sum_{k_1+k_2=d} P(VA=k_1)P(VB=k_2)$$

ou encore :

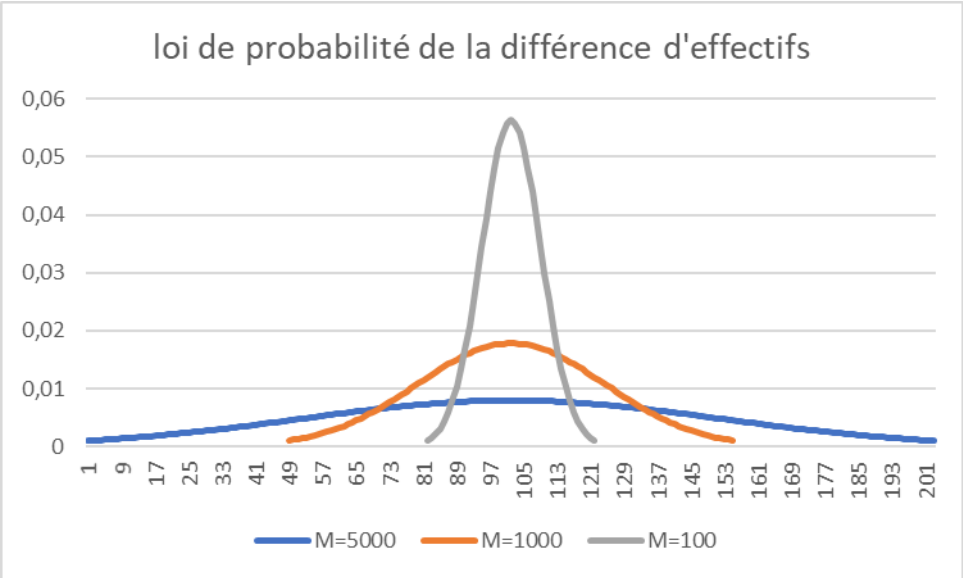
$$P(Z=d)=\frac{1}{2^{2M}}\sum_{k_1+k_2=d}\binom{M}{k_1}\binom{M}{k_2}$$

En utilisant l'identité de Vandermonde [Vandermonde], on obtient, pour $d=-M,...,M$:

$$P(Z=d)=\frac{1}{2^N}\binom{N}{M-d}$$

Cette formule est complètement explicite et permet l'évaluation des diverses probabilités. C'est une loi binomiale, symétrique par rapport à $d=0$: il est aussi probable de trouver $+d$ que $-d$.

La courbe a une forme "en cloche" ; le point intéressant est que, plus M est élevé plus la forme est évasée (contrairement à ce que l'on croit) :



Dans le cas $M=100,$ la courbe est plus resserrée, dans le cas $M=1\,000,$ elle est plus plate; et encore plus plate si $M=5\,000.$ Il en résulte évidemment que, plus M est élevé, plus la probabilité d'avoir des grandes valeurs pour la différence est élevée.

Cette loi binomiale a pour espérance 0 et pour écart-type $\sigma=\sqrt{M}.$ D'après le théorème de Moivre-Laplace, elle peut être approximée par une loi de Gauss de même espérance et de même écart-type ; la densité est :

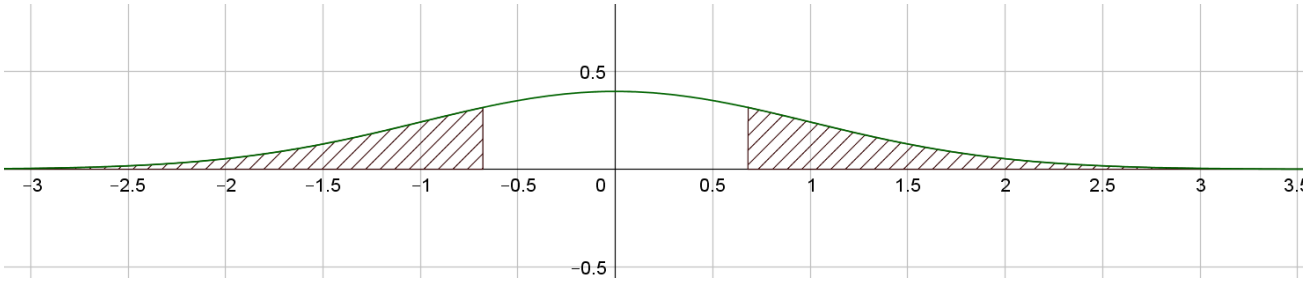
$$f(t)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-t^2}{2\sigma^2}\right)$$

Cette expression facilite beaucoup les calculs numériques.

Fixons un seuil de probabilité égal à $1/2.$ On a :

$$\int_{-\alpha\sqrt{M}}^{\alpha\sqrt{M}} f(t)\,dt=\int_{-\alpha}^{\alpha}\exp\left(-\frac{t^2}{2}\right)\frac{dt}{\sqrt{2\pi}}=\frac{1}{2}$$

pour $\alpha\approx 0.68.$ C'est l'aire non hachurée, sous le graphe, dans la figure ci-dessous.



Il en résulte que la somme des deux aires hachurées fait aussi $1/2,$ et donc, pour tout M :

$$P\left(|Z|\geq\alpha\sqrt{M}\right)\approx\frac{1}{2}.$$

Dans le cas $M=100,$ on trouve $\alpha\sqrt{M}=6.80,$ et donc les situations avec $|d|\geq 7$ ont probabilité $\frac{1}{2}$. Cela signifie que, si l'on constitue, comme expliqué précédemment, deux sous-groupes de taille 100, on a une chance sur deux de trouver que le nombre de vieux, entre les deux, diffère d'au moins 7 personnes.

Dans le cas $M=1000,$ on trouve $\alpha\sqrt{M}=21.50,$ et les situations avec $|d|\geq 22$ ont probabilité $\frac{1}{2}.$ Cela signifie que, si l'on constitue deux sous-groupes de taille 1 000, on a une chance sur deux de trouver que le nombre de vieux, entre les deux, diffère d'au moins 22 personnes.

Dans le cas $M=5\ 000,$ on trouve $\alpha\sqrt{M}=48.08,$ et on a une chance sur deux de trouver que le nombre de vieux, entre les deux, diffère d'au moins 48 personnes, et ainsi de suite.

Le point important est que la différence d'effectifs, entre les deux sous-groupes, en ce qui concerne les vieux, augmente comme $\sqrt{M}$. Cette différence ne tend pas vers 0, contrairement à ce que l'on croit souvent. Elle devient négligeable devant la taille du groupe tout entier lorsque

celle-ci est grande ; elle ne l'est pas si M est petit. Par exemple, pour $M = 100,\ \sqrt{M} = 10,$ exemple que nous allons maintenant étudier.

Pour chaque sous-groupe, on a tiré au hasard 100 fois, avec résultat 0 ou 1 avec probabilité $\frac{1}{2}$: "0" signifie "Jeune" et "1" signifie "Vieux". Voici les résultats obtenus :

	Jeunes	Vieux
G1	53	47
G2	43	57

Bien entendu, les résultats seront différents d'un tirage aléatoire à l'autre.

Cette différence de composition, d'un sous-groupe à l'autre, peut avoir des conséquences si on cherche à tester l'efficacité d'un médicament, surtout lorsque les populations sur lesquelles il peut être bénéfique sont en proportion très faible.

4. Efficacité d'un médicament

Considérons une maladie qui n'a aucun effet sur les Jeunes et qui, au pire, tue 5% des Vieux. On administre un médicament au groupe G2 et rien au groupe G1. On espère que, grâce au médicament, la proportion de décès chez les Vieux ne sera plus que de 3%, à $\pm 1\%$ près.

Dans le groupe G1, le nombre de décès sera $\frac{5 \times 47}{100} \approx 2.$

Dans le groupe G2, le nombre de décès sera $\frac{3 \times 57}{100} \approx 1.7 \simeq 2$; mais il peut être compris entre $\frac{2 \times 57}{100} \approx 1.1 \simeq 1$ et $\frac{4 \times 57}{100} \approx 2.28 \simeq 2$; autrement dit, du fait de la faible taille des échantillons et de la différence de composition des sous-groupes G1 et G2, l'efficacité du médicament ne peut être avérée.

5. Ne pas s'en remettre au hasard

Laisser le hasard décider, en espérant ainsi obtenir des "panels" les plus égaux possible, n'est certainement pas une bonne idée : les lois fondamentales des probabilités montrent, à l'opposé, que les panels ainsi réalisés seront fondamentalement différents. La "randomisation" n'est certainement pas une bonne pratique.

De manière générale, il n'est pas sain de s'en remettre au hasard pour régler une question que l'on ne sait pas régler de manière explicite. Dans le cas qui nous intéresse, il existe des méthodes parfaitement explicites qui répondent bien au besoin exprimé, comme nous allons le voir.

6. Utilisation d'histogrammes

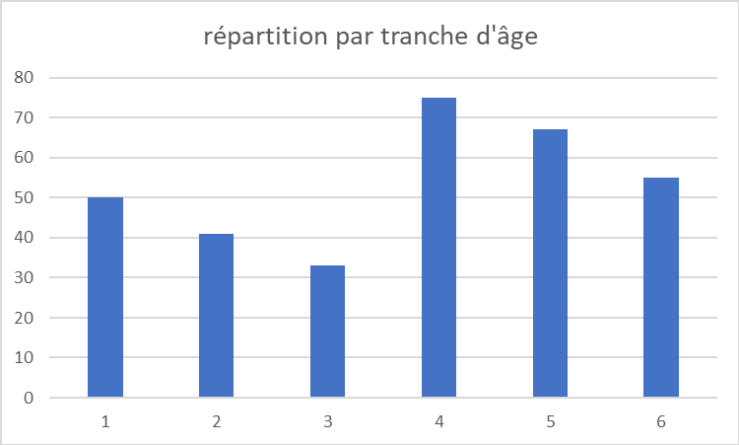
Plutôt que de laisser le hasard décider, on peut procéder de la manière suivante. On construit l'histogramme de la population concernée, par exemple en ce qui concerne l'âge. Cela signifie que l'on range la population par tranches : 0-10 ans, 10 - 20 ans, etc. Toutes les personnes de la même tranche sont considérées comme équivalentes du point de vue de l'âge. On prend une personne sur deux dans chaque tranche et on la range dans le sous-groupe A, une personne sur deux dans le sous-groupe B (selon les tranches, on peut avoir une difficulté du fait de la parité, mais c'est sans importance ici). On obtient deux sous-groupes qui ont le même profil vis-à-vis de l'âge.

Voici un exemple très simple :

On dispose d'un panel de 321 personnes, réparties en tranches d'âge selon le schéma suivant :

âge	10-20	20-30	30-40	40-50	50-60	60-70	total
effectif	50	41	33	75	67	55	321

Voici la représentation sous forme d'histogramme :



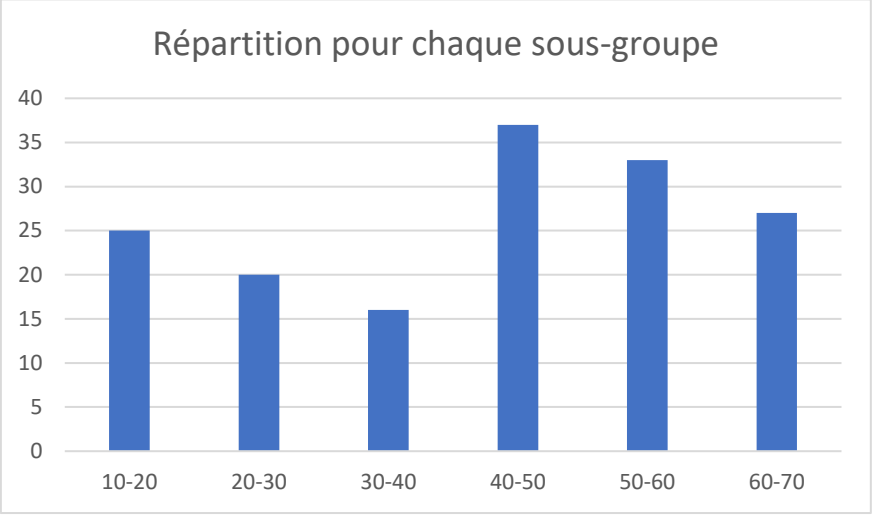
On veut répartir ce panel en deux sous-groupes, aussi identiques que possible du point de vue du profil d'âge.

Pour la première tranche, on fait 25, 25 et il ne reste rien ;
Pour la seconde tranche, on fait 20, 20 et il reste une personne
Pour la troisième tranche, on fait 16, 16 et il reste une personne
Pour la quatrième tranche, on fait 37, 37 et il reste une personne
Pour la cinquième tranche, on fait 33, 33 et il reste une personne
Pour la sixième tranche, on fait 27, 27 et il reste une personne

Voici le résultat de l'opération de répartition :

âge	10-20	20-30	30-40	40-50	50-60	60-70	total
effectif 1	25	20	16	37	33	27	158
effectif 2	25	20	16	37	33	27	158
reste	0	1	1	1	1	1	5

et le résultat sous forme d'histogramme :



La forme est bien la même que dans le panel tout entier, et les deux sous-groupes sont en tout point identiques. Quant aux 5 personnes restantes, on peut au choix :

- Les incorporer alternativement à chacun des panels (la personne dans 20-30 dans le panel 1, celle de 30-40 dans le panel 2, etc.), ce qui modifiera légèrement les profils de chaque panel ;
- Les éliminer de l'étude, ou les mettre de côté en attendant que d'autres personnes viennent compléter les panels.

Si deux critères sont nécessaires, par exemple l'âge et le poids, on fait la même chose pour ce second critère : un découpage en tranches, par exemple 0 - 10 kg, 10 - 20 kg, etc. Ensuite, on décompose en deux pour chaque situation croisée : par exemple, toutes les personnes satisfaisant âge entre 40 et 50 ans et poids entre 60 et 70 kg sont divisées en deux sous-groupes.

Voici un exemple, comme précédemment ; il porte sur un panel de 1205 personnes :

poids\âge	10-20	20-30	30-40	40-50	50-60	60-70	total
40-50	5	10	33	75	67	55	245
50-60	12	23	45	87	45	56	268
60-70	14	36	65	58	46	45	264
70-80	20	22	33	45	89	39	248
80-90	12	5	31	31	68	33	180
							1205

Chaque cellule ayant un nombre pair est exactement divisée en deux ; si elle comporte un nombre impair, il reste 1. Voici le résultat, pour chacun des deux sous-groupes :

poids\âge	10-20	20-30	30-40	40-50	50-60	60-70	total
40-50	2	5	16	37	33	27	120
50-60	6	12	22	43	22	28	133
60-70	7	18	32	29	23	22	131
70-80	10	11	16	22	44	19	122
80-90	6	2	15	15	34	16	88

Chaque sous-groupe comporte 594 personnes.

Voici les restes :

poids\âge	10-20	20-30	30-40	40-50	50-60	60-70	total
40-50	1	0	1	1	1	1	5
50-60	0	0	1	1	1	0	3
60-70	0	0	1	0	0	1	2
70-80	0	0	1	1	1	1	4
80-90	0	1	1	1	0	1	4

Il reste un total de 18 personnes ; que l'on peut traiter comme précédemment : les affecter alternativement à l'un ou l'autre sous-groupe, ou bien les mettre de côté.

Annexe

Calcul de l'espérance de $|Z|$

Dans la pratique, on s'intéresse à la valeur absolue de la différence d'effectifs entre les deux sous-groupes ; cette différence étant symétrique, la valeur absolue a pour densité :

$$g(t) = \frac{2}{\sigma\sqrt{2\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right), \ t \geq 0, \ = 0 \ \text{si} \ t < 0.$$

Donc :

$$E(|Z|) = \int_0^{+\infty} \frac{2t}{\sigma\sqrt{2\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right) dt = \frac{2\sigma}{\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(\frac{-x^2}{2}\right) dx = \frac{\sigma}{\sqrt{2\pi}} \int_0^{+\infty} \exp\left(\frac{-y}{2}\right) dy = \frac{2\sigma}{\sqrt{2\pi}}$$

et donc finalement :

$$E(|Z|) = \sqrt{\frac{2M}{\pi}} = \sqrt{\frac{N}{\pi}}$$

En moyenne, l'écart jeunes-vieux au sein d'un même sous-groupe est proportionnel à la racine de l'effectif complet.

Les mots "en moyenne" signifient que, si on répète un grand nombre de fois l'expérience de répartition d'un groupe de N personnes en deux sous-groupes, en conservant le même N, l'écart moyen entre les effectifs des deux sous-groupes sera celui indiqué par la formule ci-dessus.

Références

Inégalité de Vandermonde
https://fr.wikipedia.org/wiki/Identité_de_Vandermonde

Livre classique sur ces questions :
Johnson, N.L. et Leone, F.C. "Statistics and Experimental Design in Engineering and The Physical Sciences", Wiley, 1977.