Intervalle de confiance à 95%

Jaksic C., Combescure C.
Service d'épidémiologie clinique
Centre de Recherche Clinique

Plan de la présentation

- Notion d'inférence
- Expérience et simulation sur le lancer d'une pièce
- Définition de l'intervalle de confiance
- Principe de l'intervalle de confiance d'une moyenne estimée
- Interprétation des intervalles de confiance: RCT sur le tai-chi
- Recommandations
- Conclusions

Inférence

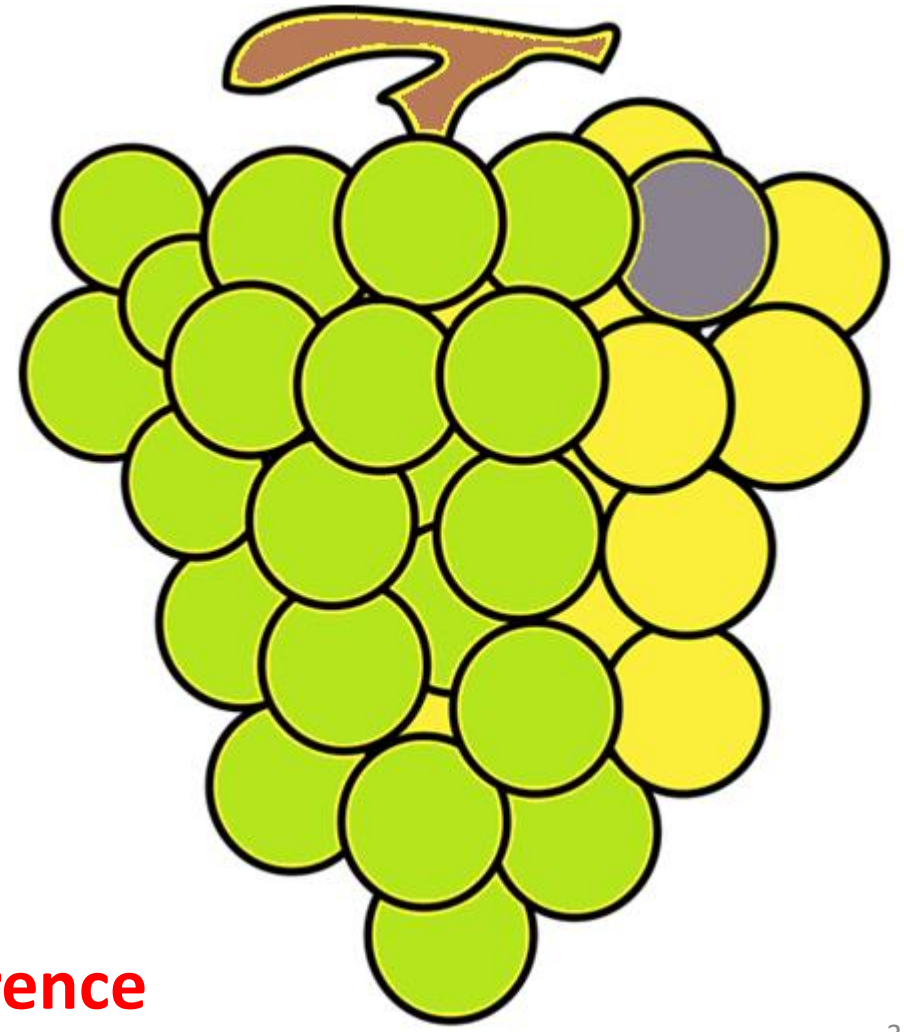
- Comment avoir une idée de la qualité de la grappe de raisin ?

♦ Goûter un des raisins :

si le raisin est bon, vous concluez que la grappe est bonne

si le raisin est mauvais, vous concluez que la grappe est mauvaise

♦ Vous **généralisez** à l'ensemble de la grappe la qualité du raisin testé



Inférence

Inférence

- Comment sélectionner le raisin à goûter ?

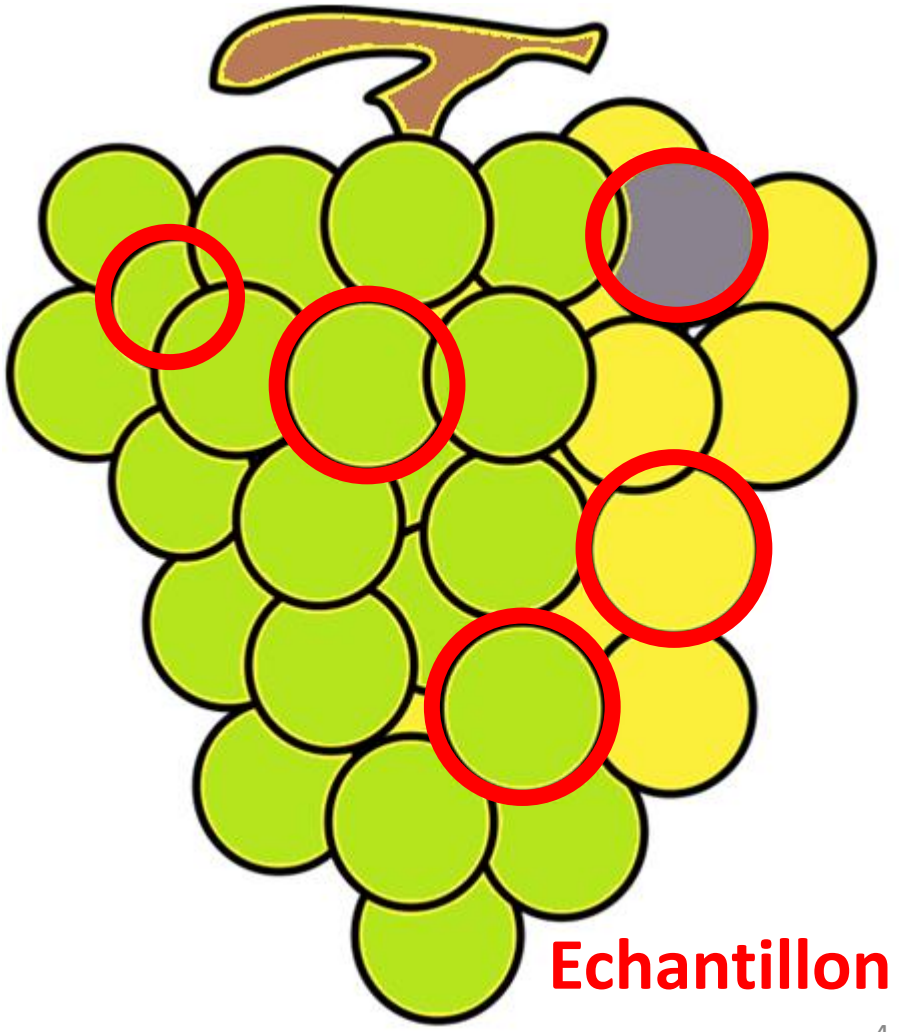
♦ Sélection subjective ?

risque de goûter surtout les jolis raisins => inférence incorrecte

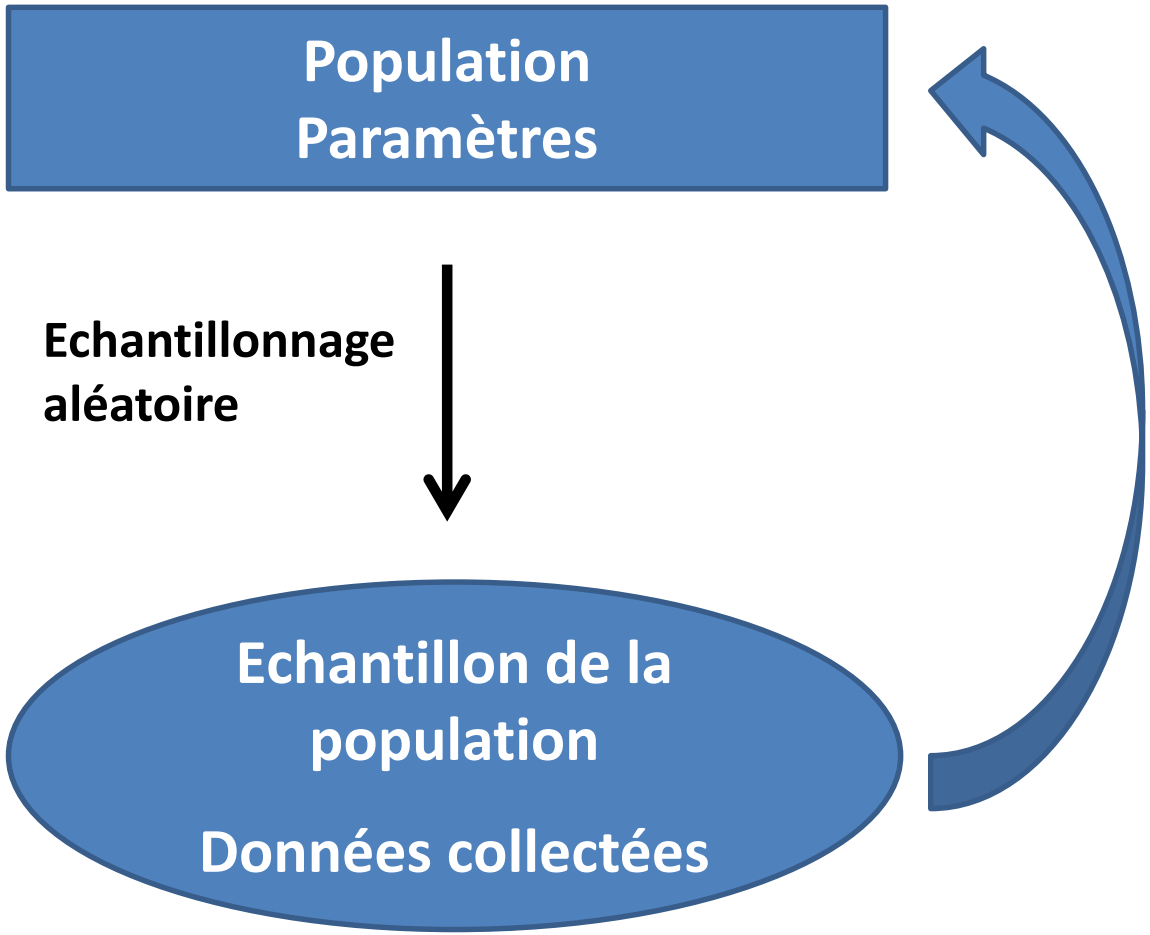
♦ Sélection **aléatoire**

protège d'une sélection subjective

♦ Plusieurs raisins: **échantillon**



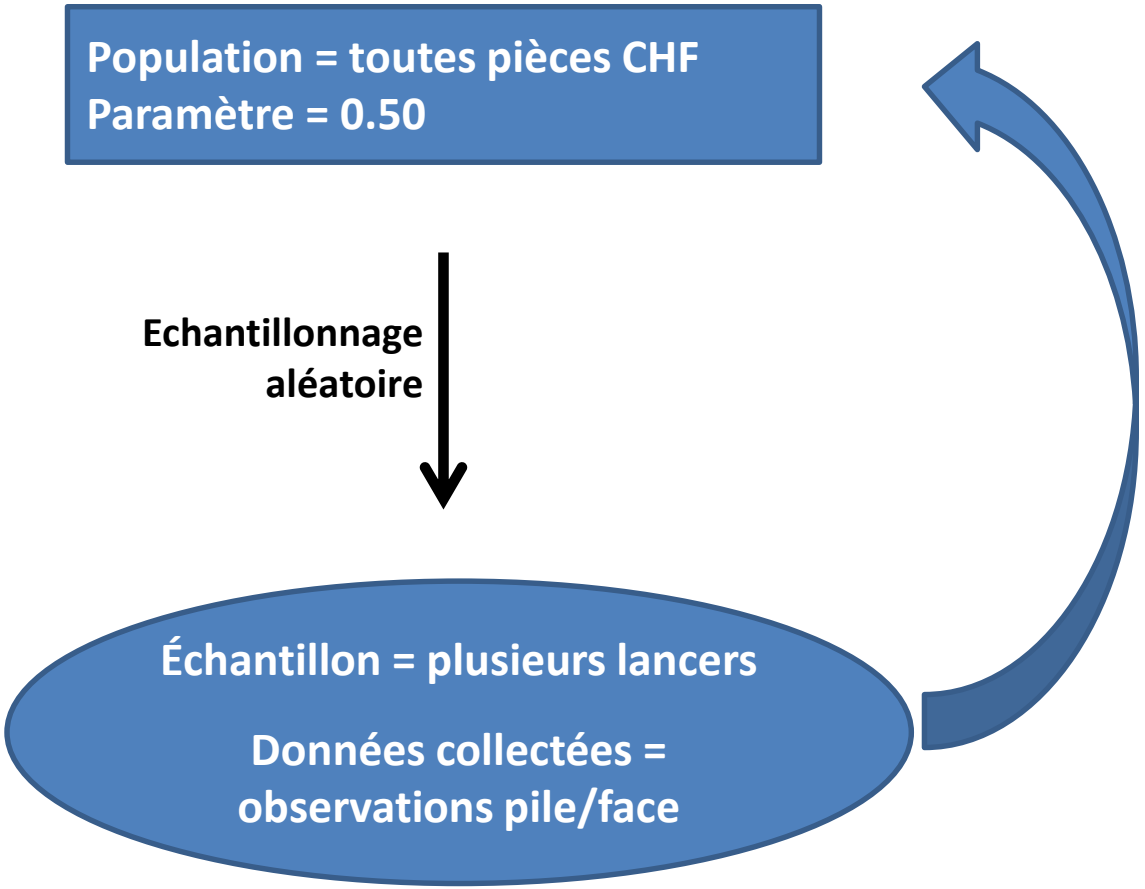
Inférence statistique



Qu'est-ce que les données observées dans l'échantillon me permettent de conclure sur la population (les paramètres)?

Lancer de pièces

Quelle est la probabilité de tomber sur face lorsqu'on lance une pièce de monnaie suisse?



Quelles valeurs du paramètre peuvent expliquer nos observations?

Lancer de pièces: expériences

1ère expérience: veuillez lancer une pièce

Nombre de «pile»: 4
Nombre de «face»: 3
Estimation de la probabilité qu'une pièce tombe sur «face»: 3/7=0,43

2ème expérience: veuillez lancer à nouveau une pièce

Nombre de «pile»:3
Nombre de «face»: 4
Estimation de la probabilité qu'une pièce tombe sur «face»: 4/7=0,57

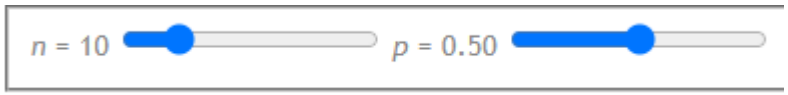
Lancer de pièces: simulation

- Génération de données (pile/face) d'un grand nombre d'échantillons en fixant:
  - la valeur du paramètre: probabilité de «face» = 0,50
  - la taille d'échantillon : n=10 puis n=50 lancers de pièce

- Objectif:
  - Étudier le «comportement» des observations

http://www.distributome.org/V3/exp/BinomialExperiment.html

Lancer de pièces: simulation

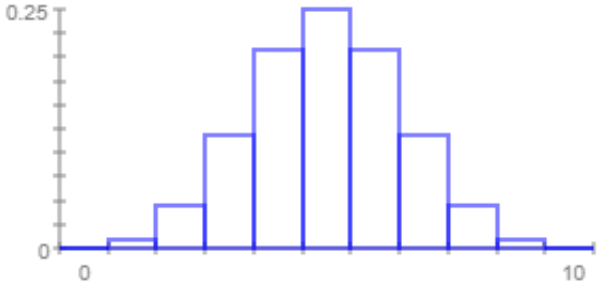


Taille d'échantillon (nombre de lancers de pièce)

Paramètre (probabilité qu'une pièce tombe sur «face»)

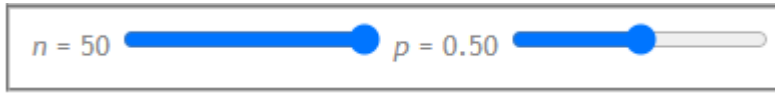
Fréquence des échantillons avec 0 «face», 1 «face»,…

X	Dist
0	0.00
1	0.01
2	0.04
3	0.12
4	0.21
5	0.25
6	0.21
7	0.12
8	0.04
9	0.01
10	0.00



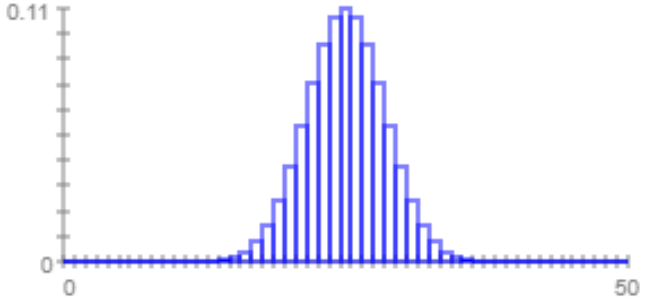
- Avec des échantillons de taille 10 et un paramètre de 0,50:
  - on s'attend à observer 5 «faces» (donc une estimation de la probabilité de «face» égale à 0,50) dans 25% des échantillons (c'est la fréquence la plus élevée)
  - dans 75% des échantillons, l'estimation est différente de 0,50
  - les échantillons avec ≤1 ou ≥9 «faces» sont possibles mais peu fréquents (moins de ~2 ou 3%): lorsque le paramètre est égal à 0,50, on ne s'attend pas à observer ≤1 ou ≥9 «faces» (donc une estimation ≤0,10 ou ≥0,90) dans l'échantillon
  - La fréquence des échantillons avec 0 «face», 1»face»,… correspond à une distribution théorique qui ne dépend que de la valeur du paramètre et de la taille d'échantillon

Lancer de pièces: simulation



Fréquence des échantillons avec 0 «face», 1 «face»,…

Taille d'échantillon (nombre de lancers de pièce)

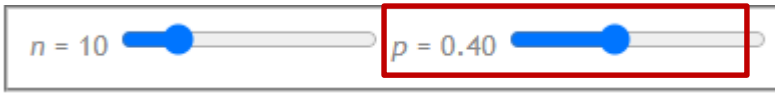


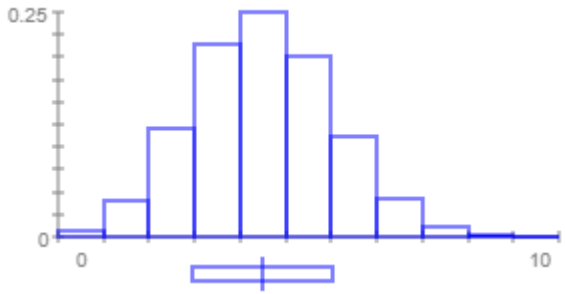
X	Dist
0	0.00
1	0.00
2	0.00
3	0.00
4	0.00
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00
11	0.00
12	0.00
13	0.00
14	0.00
15	0.00
16	0.00
17	0.01
18	0.02
19	0.03
20	0.04
21	0.06
22	0.08
23	0.10
24	0.11
25	0.11

X	Dist
26	0.11
27	0.10
28	0.08
29	0.06
30	0.04
31	0.03
32	0.02
33	0.01
34	0.00
35	0.00
36	0.00
37	0.00
38	0.00
39	0.00
40	0.00
41	0.00
42	0.00
43	0.00
44	0.00
45	0.00
46	0.00
47	0.00
48	0.00
49	0.00
50	0.00

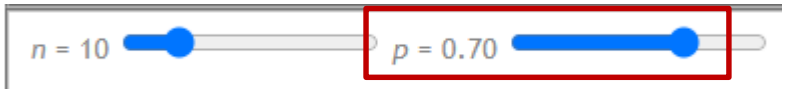
- Avec des échantillons de taille 50 et un paramètre de 0,50:
  - les échantillons avec ≤17 ou ≥33 «faces» sont possibles mais peu fréquents (moins de ~2 ou 3%): lorsque le paramètre est égal à 0,50, on ne s'attend pas à observer ≤17 ou ≥33 «faces» (donc une estimation ≤0,34 ou ≥0,66) dans l'échantillon
  - Les échantillons donnant une estimation éloignée de la valeur du paramètre sont moins fréquents qu'avec des échantillons de taille 10

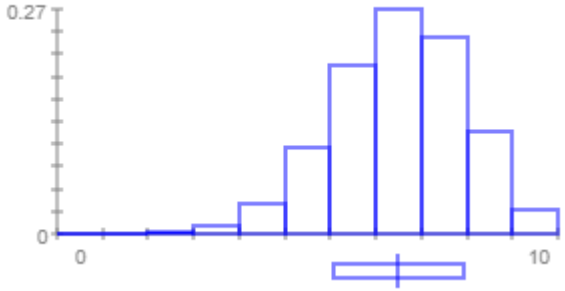
Lancer de pièces: simulation





X	Dist
0	0.01
1	0.04
2	0.12
3	0.21
4	0.25
5	0.20
6	0.11
7	0.04
8	0.01
9	0.00
10	0.00





X	Dist
0	0.00
1	0.00
2	0.00
3	0.01
4	0.04
5	0.10
6	0.20
7	0.27
8	0.23
9	0.12
10	0.03

La chance d'observer 5 «faces» dans un échantillon aléatoire est:

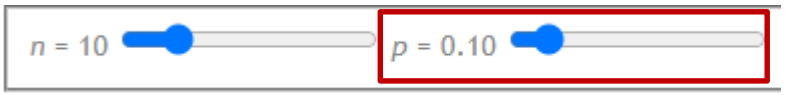
- 20% si le paramètre est égal à 0,40
- 25% si le paramètre est égal à 0,50
- 10% si le paramètre est égal à 0,70

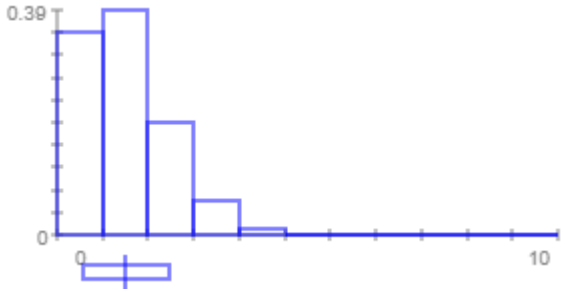
Lancer de pièces: simulation

Si j'observe 5 «faces» dans l'échantillon de mon étude, qu'est-ce que je peux dire de la valeur du paramètre?

Est-elle égale à 0,40? 0,50? 0,70?

Toutes ces valeurs de paramètres sont compatibles avec les données de l'échantillon !

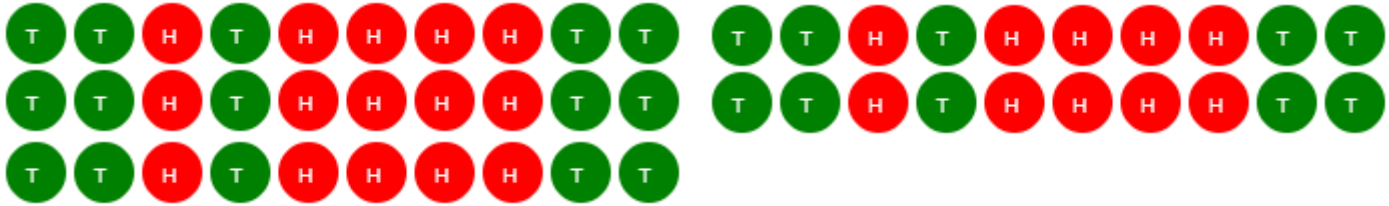




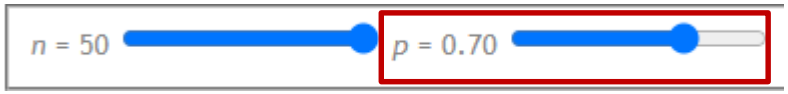
X	Dist
0	0.35
1	0.39
2	0.19
3	0.06
4	0.01
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00

Par contre, si le paramètre valait 0,10, je ne m'attendrais pas à observer 5 «faces» dans l'échantillon (moins de 1% de chance)
0,10 n'est pas une valeur du paramètre compatible avec les données de l'échantillon.

Lancer de pièces: simulation



Si j'observe 25 «faces» dans l'échantillon (n=50) de mon étude, qu'est-ce que je peux dire de la valeur du paramètre?



Par contre, si le paramètre valait 0,70, je ne m'attendrais pas à observer 25 «faces» dans l'échantillon (moins de 1% de chance) 0,70 n'est pas une valeur du paramètre compatible avec les données de l'échantillon.

X	Dist
0	0.00
1	0.00
2	0.00
3	0.00
4	0.00
5	0.00
6	0.00
7	0.00
8	0.00
9	0.00
10	0.00
11	0.00
12	0.00
13	0.00
14	0.00
15	0.00
16	0.00
17	0.00
18	0.00
19	0.00
20	0.00
21	0.00
22	0.00
23	0.00
24	0.00
25	0.00

X	Dist
26	0.00
27	0.01
28	0.01
29	0.02
30	0.04
31	0.06
32	0.08
33	0.10
34	0.11
35	0.12
36	0.12
37	0.11
38	0.08
39	0.06
40	0.04
41	0.02
42	0.01
43	0.00
44	0.00
45	0.00
46	0.00
47	0.00
48	0.00
49	0.00
50	0.00

Lancer de pièces: simulation

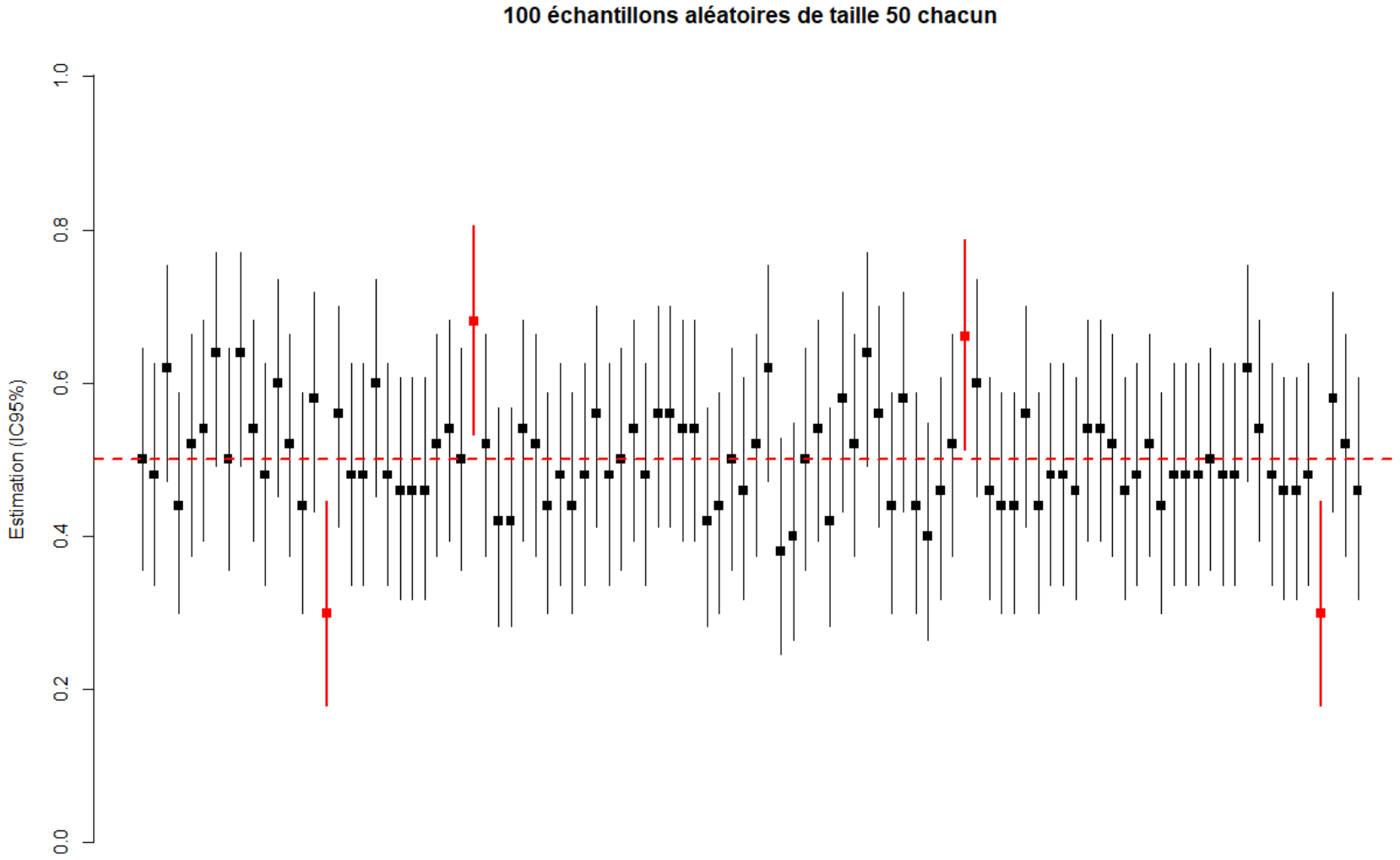
- Enseignements de la simulation:
  - Lorsque l'échantillon aléatoire est de petite taille, l'estimation du paramètre varie bcp d'un échantillon à l'autre
  - A taille identique et à paramètre identique, certains échantillons sont plus probables que d'autres, certains sont peu probables
  - La variabilité (d'un échantillon à l'autre) de l'estimation diminue lorsque la taille d'échantillon augmente
  - Certaines valeurs de paramètre sont compatibles avec les données de l'échantillon et d'autres non
  - Plus la taille d'échantillon est grande, moins il y a de valeurs du paramètre compatibles avec les données observées

Définition de l'intervalle de confiance à 95%

- Intervalle de confiance à 95% (IC95%) = Ensemble des valeurs de paramètres compatibles avec les données de l'échantillon

- Une valeur est «incompatible» si, pour cette valeur, la probabilité d'observer les données obtenues dans l'échantillon de l'étude est faible (=> le 95% de l'intervalle de confiance)

- Attention:
  - il est possible que la valeur du paramètre se trouve en dehors de l'IC95%
  - mais, par construction de l'IC95%, on se donne les moyens que cela arrive rarement (5% des IC95% ne contiennent pas la valeur du paramètre)
  - sur un échantillon spécifique, on ne peut pas savoir si l'IC95% contient le paramètre

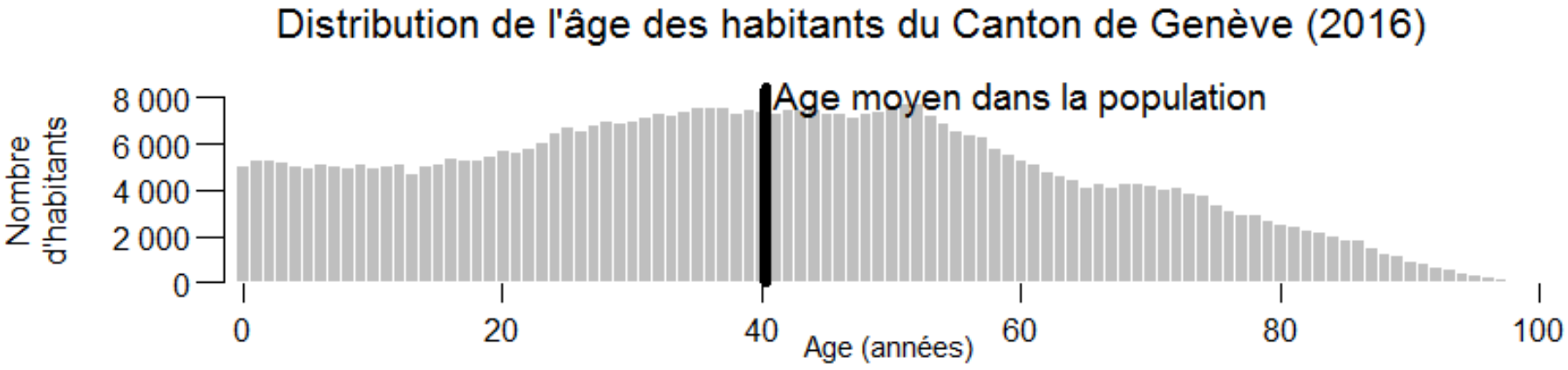


Sur un grand nombre d'échantillons aléatoires de la même population, il est attendu que dans 5% des échantillons, l'IC95% ne contient pas la valeur du paramètre (quelle que soit la taille d'échantillon)

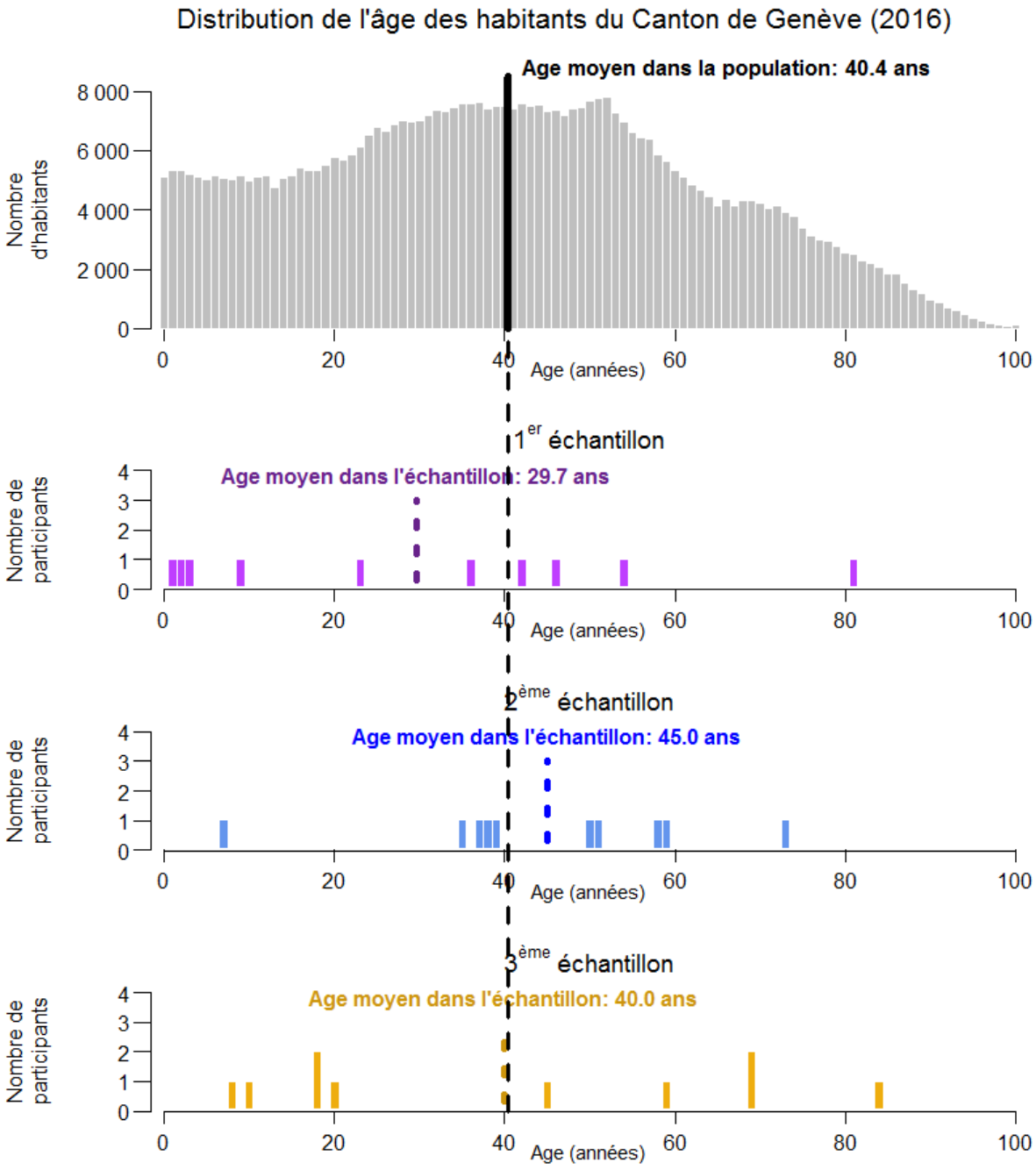
En pratique, le chercheur a un seul échantillon et il ne sait pas si l'intervalle de confiance contient ou non la valeur du paramètre

Autre exemple: estimation de l'âge moyen

- Source des données: OFS, année 2016
- Données de **toute** la population du canton de Genève
- **Paramètre**: âge moyen 40,4 ans (dans la population)



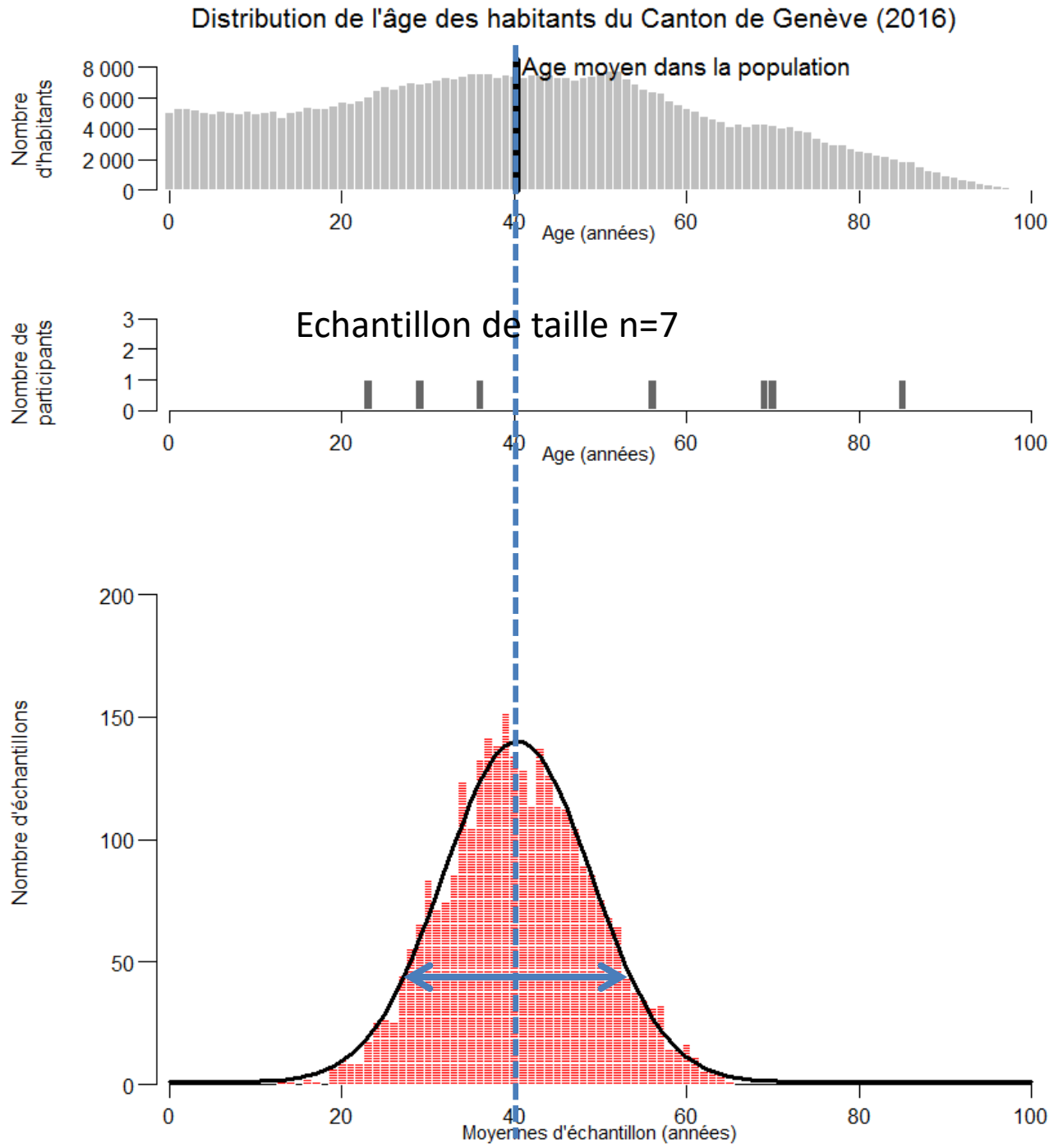
- **Echantillon aléatoire** de cette population
- Age moyen dans l'échantillon => **estimation** de l'âge moyen dans la population
- Comment se comporte l'estimation de l'âge moyen ?



3 échantillons **<u>aléatoires</u>** de la population (n=10)

L'estimation de l'âge moyen dans la population varie d'un échantillon à l'autre

Cette variabilité est causée par l'aléa de l'échantillonnage

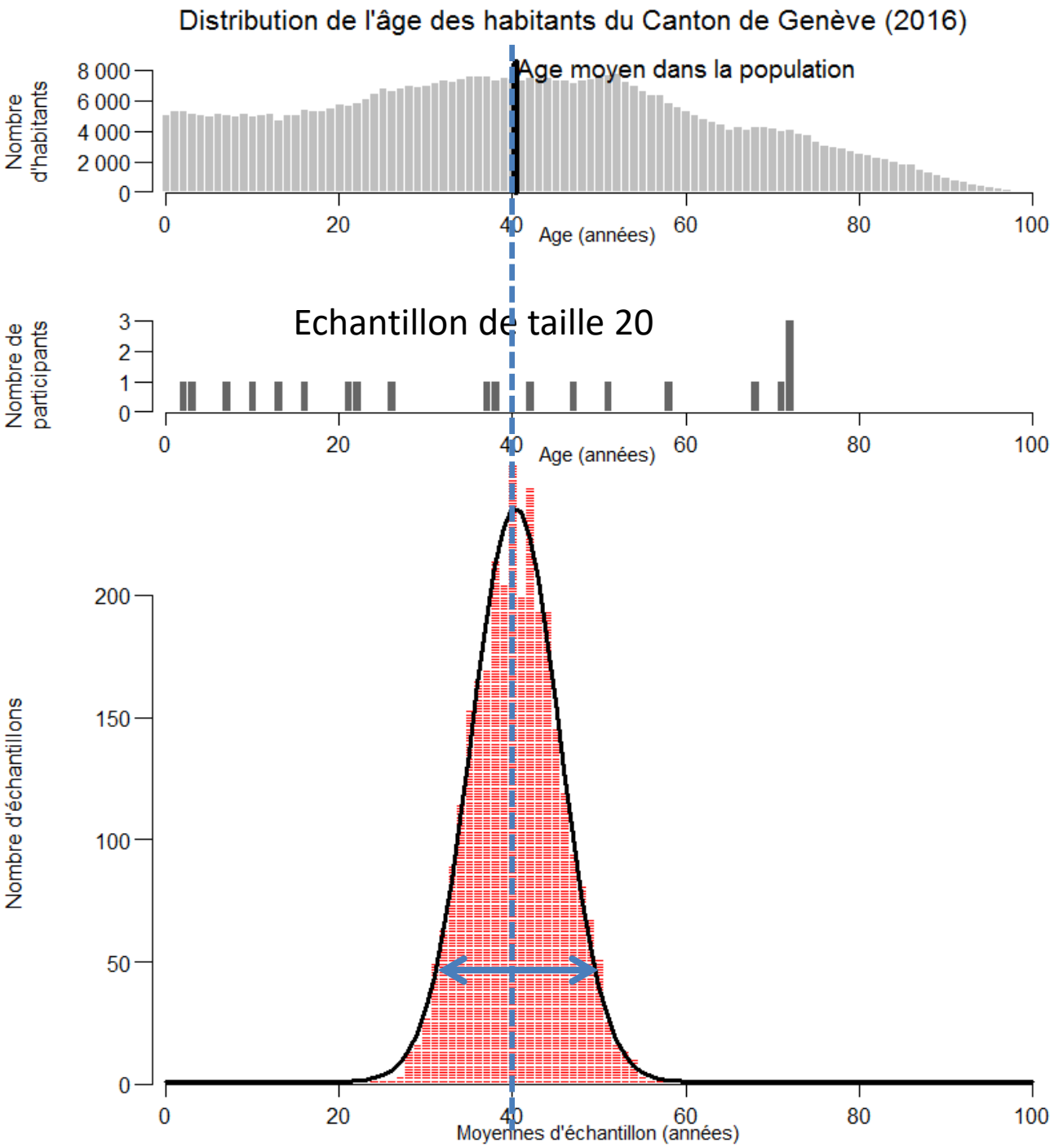


Les moyennes d'échantillon sont centrées sur la valeur du paramètre dans la population

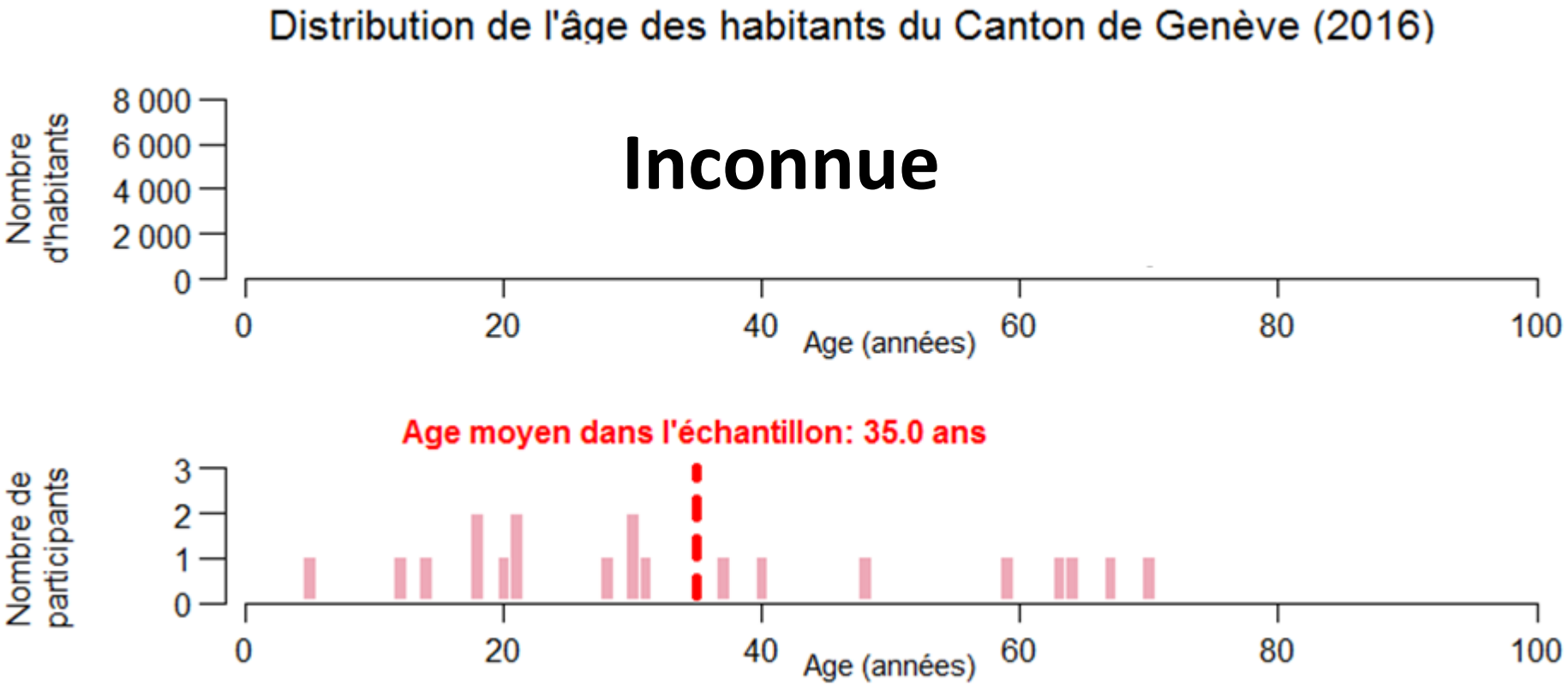
Les moyennes d'échantillon sont variables

Les moyennes d'échantillon ont une distribution en cloche

Vrai même si la variable n'est pas gaussienne dans la population (à condition que n soit suffisamment grand)

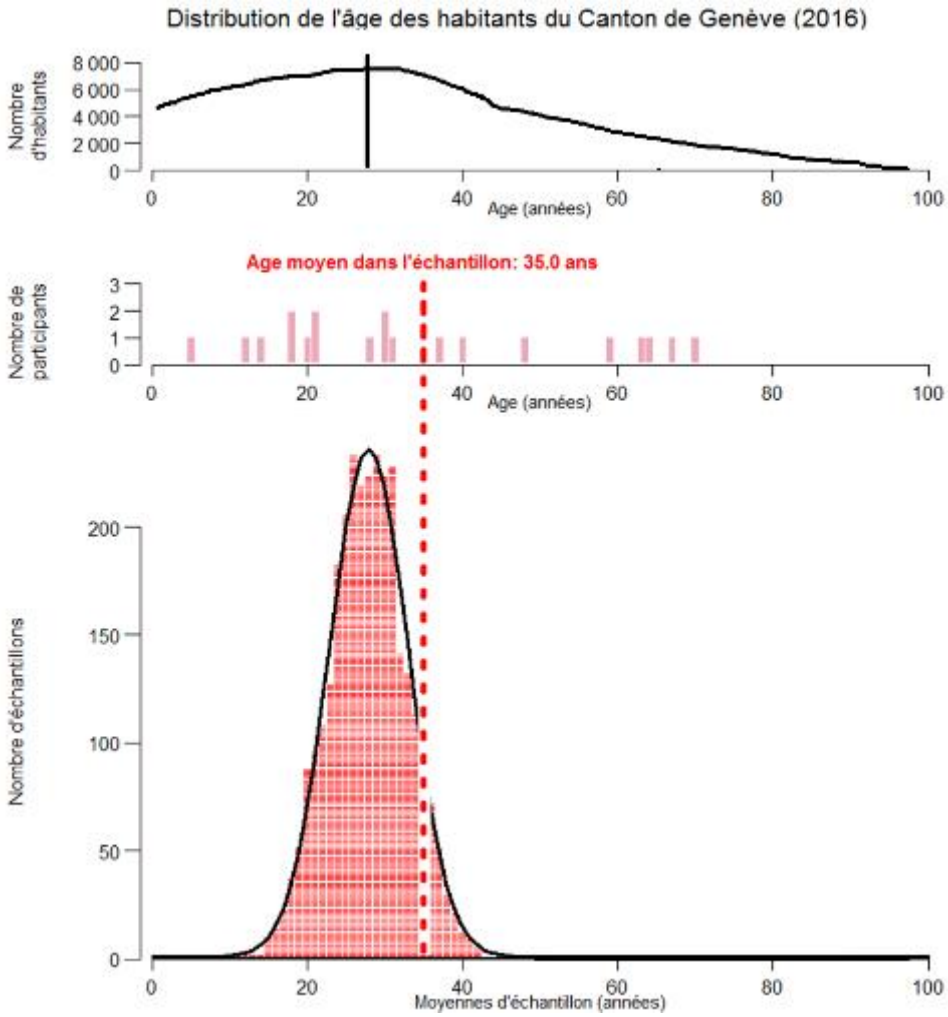


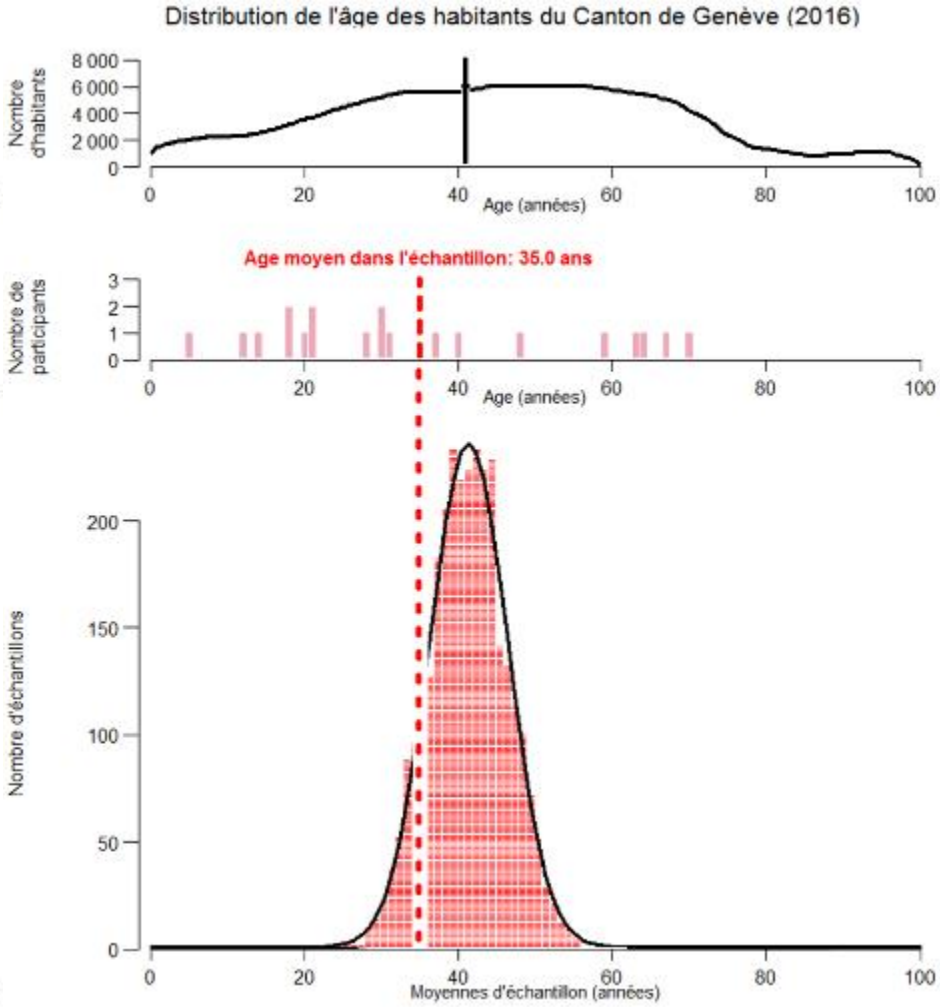
La variabilité des moyennes d'échantillon diminue lorsque la taille d'échantillon augmente

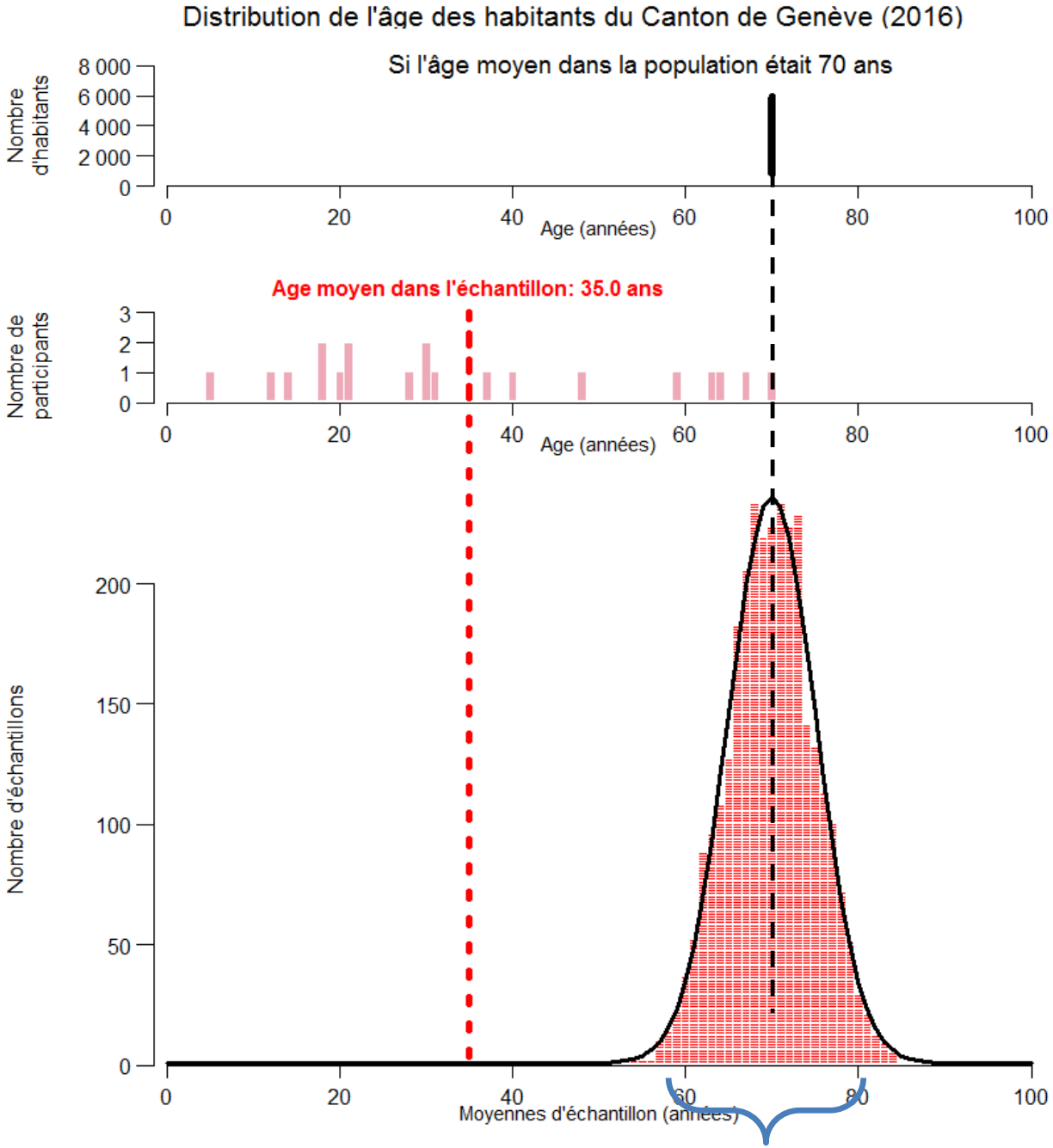


En général, il n'est pas possible de connaître la valeur du paramètre dans la population ou les données de toute la population

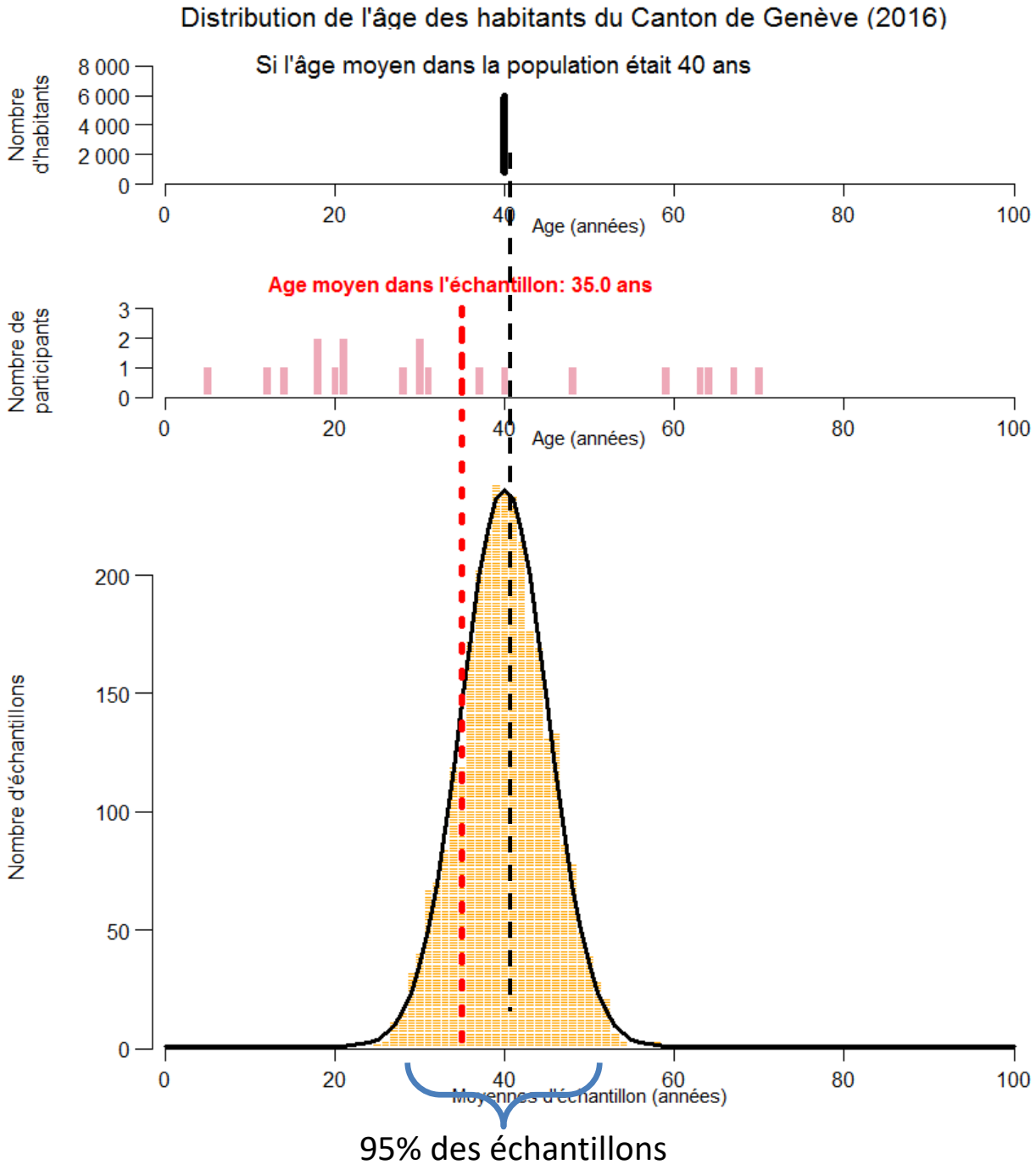
Comment savoir dans quelle situation je suis ?



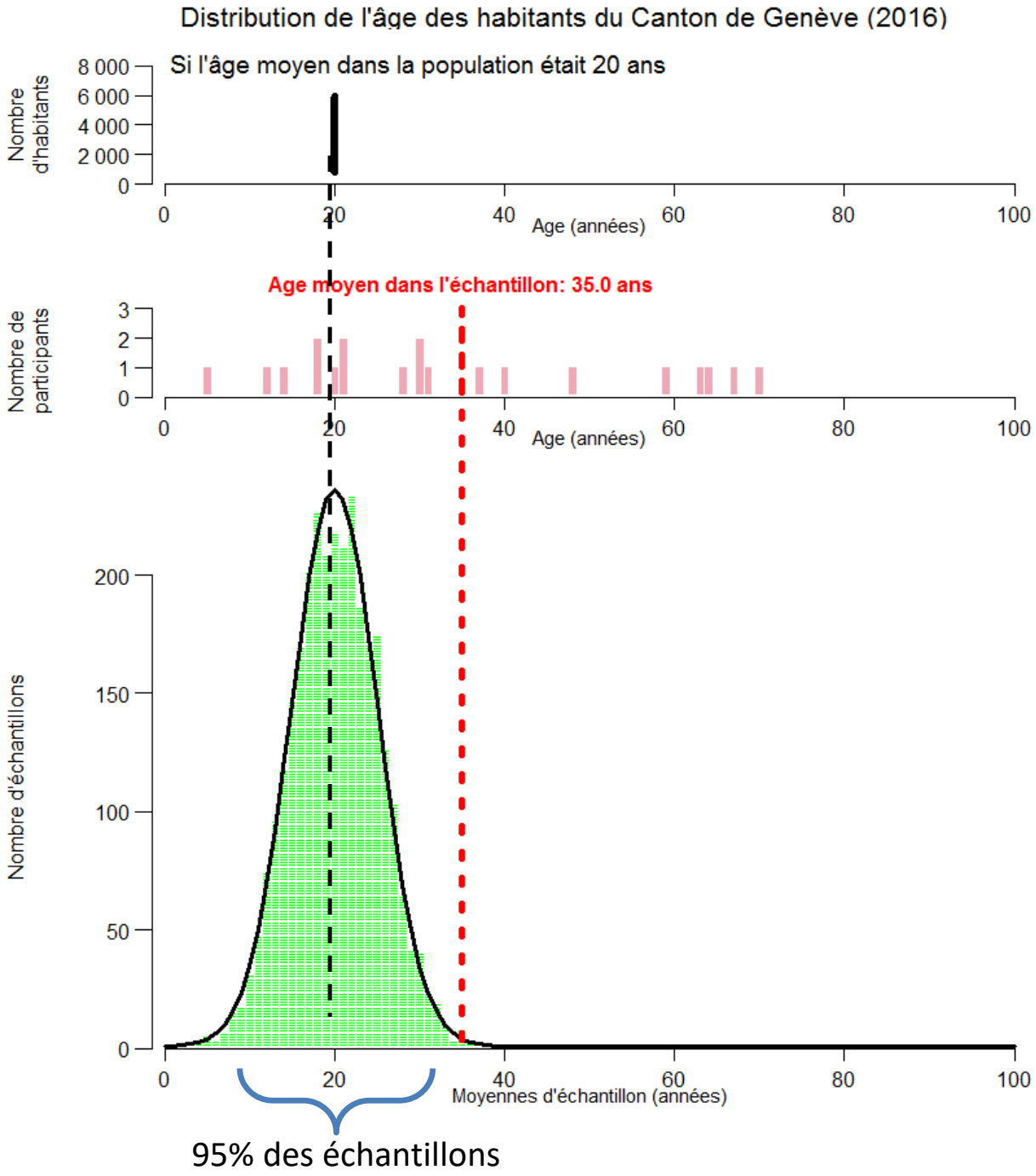




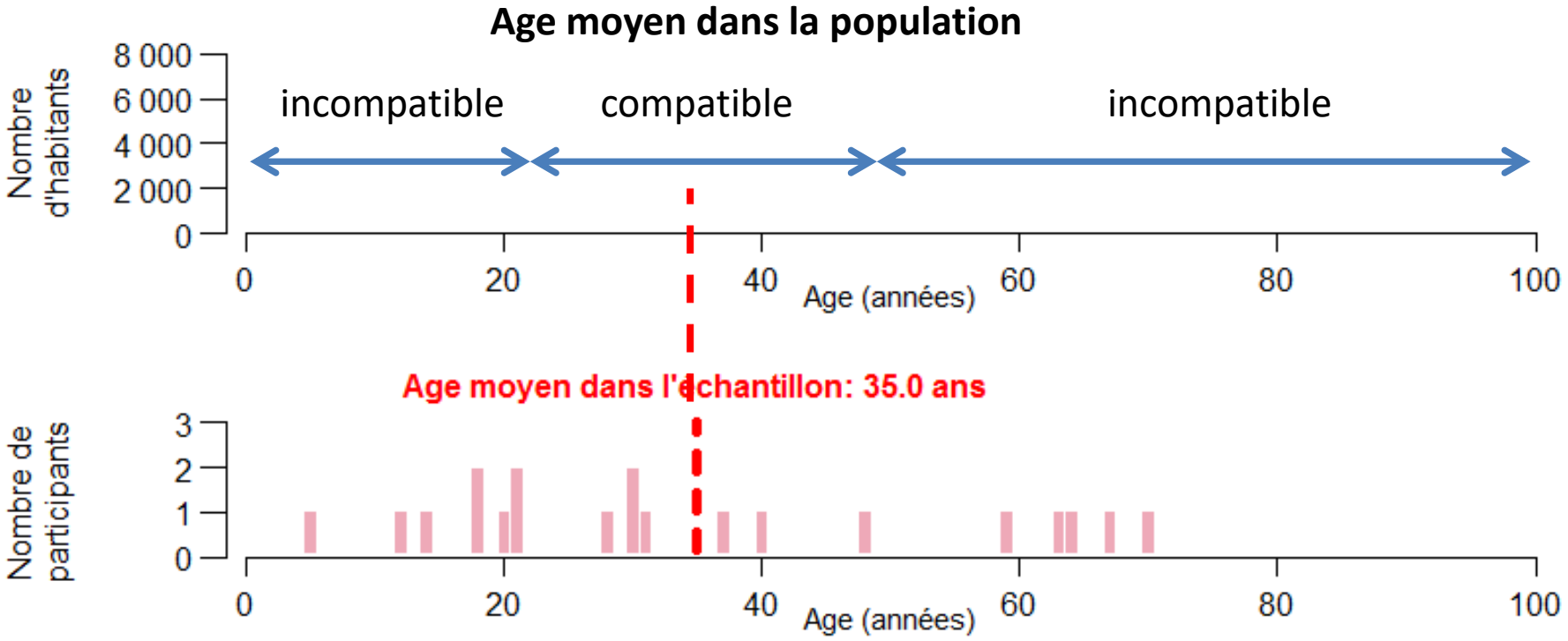
Un âge moyen de **70 ans** **dans la population** **n'est pas compatible** avec les données observés dans l'échantillon car:
l'âge moyen **dans l'échantillon** est en dehors de l'intervalle contenant 95% des échantillons si l'âge moyen était **70 ans** dans la population



Un âge moyen de **40 ans** **dans la population** **est compatible** avec les données observées dans l'échantillon car l'âge moyen **dans l'échantillon** est dans l'intervalle contenant 95% des échantillons si l'âge moyen était **40 ans** dans la population



Un âge moyen de **20 ans dans la population** **n'est pas compatible** avec les données observés dans l'échantillon car l'âge moyen **dans l'échantillon** est en dehors de l'intervalle contenant 95% des échantillons si l'âge moyen était **20 ans** dans la population



<u>Intervalle de confiance à 95%</u>:
ensemble des valeurs du paramètre (dans la population) qui sont compatibles avec les données observées dans l'échantillon

Effect of tai chi versus aerobic exercise for fibromyalgia: comparative effectiveness randomized controlled trial

Chenchen Wang,[1] Christopher H Schmid,[2] Roger A Fielding,[3] William F Harvey,[1] Kieran F Reid,[3] Lori Lyn Price,[4] Jeffrey B Driban,[1] Robert Kalish,[5] Ramel Rones,[6] Timothy McAlindon[1]

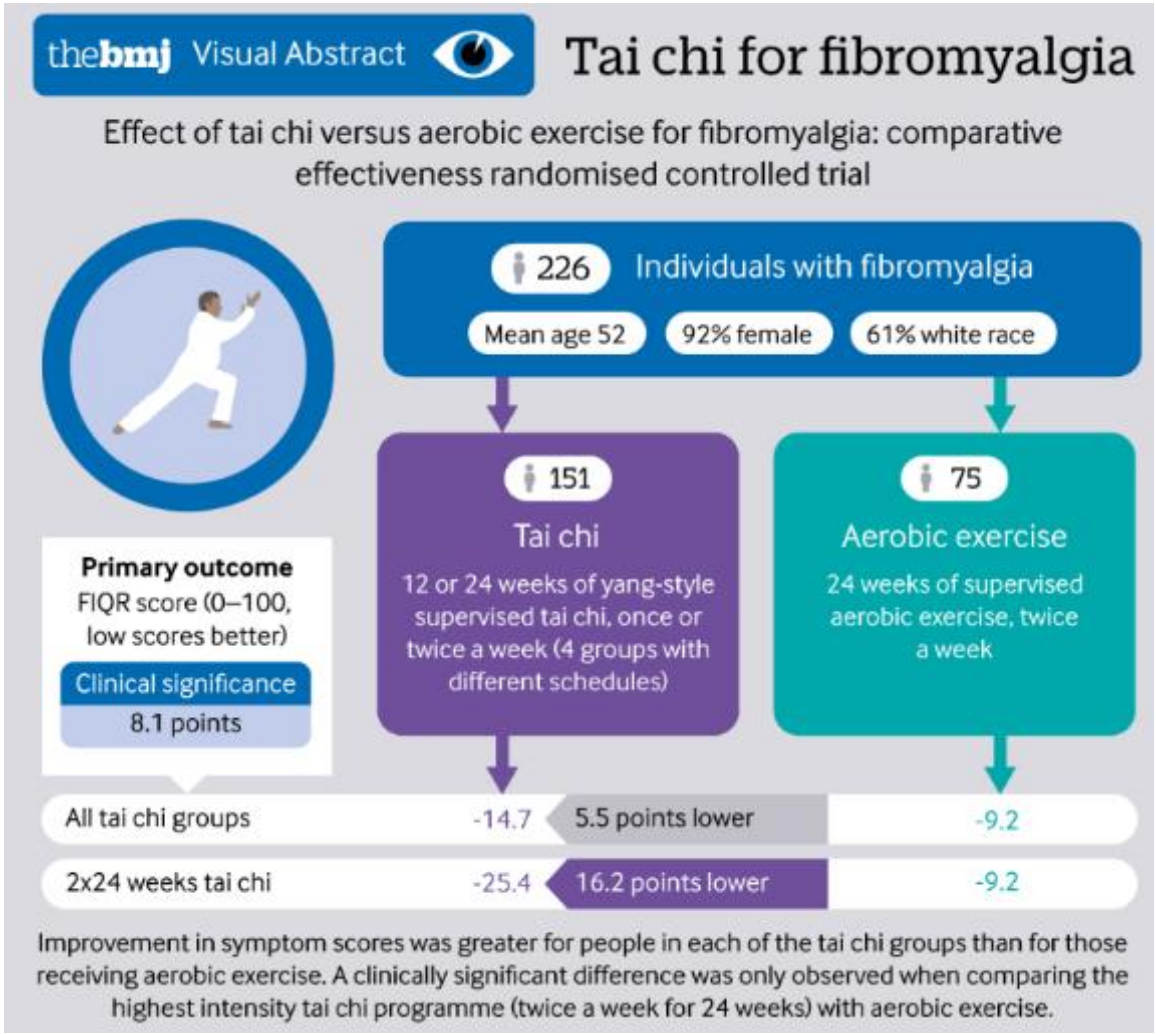


Table 3 | Between group differences at all weeks*

Variables	Aerobic exercise *v* tai chi groups combined Mean (95% CI)	Aerobic exercise 2×24 weeks *v* tai chi 2×24 weeks Mean (95% CI)
FIQR score (range 0-100)†:		
Week 12	5.4 (0.6 to 10.1)	10.9 (3.4 to 18.5)
Week 24	5.5 (0.6 to 10.4)	16.2 (8.7 to 23.6)
Week 52	2.7 (−2.3 to 7.7)	11.1 (2.7 to 19.6)

L'effet estimé est +5.5 en faveur du tai chi (sur le score FIQR)

L'intervalle de confiance à 95% va de 0.6 à 10.4: les données de l'échantillon permettent de dire que, en réalité, l'effet du tai chi peut être 0.6 ou 10.4.

Significativité statistique: l'hypothèse d'absence d'effet dans la population (vraie différence de moyenne = 0) est rejetée

L'intervalle de confiance à 95% contient 8.1 (la plus petite différence cliniquement pertinente): les données de l'échantillon ne permettent pas d'exclure que le véritable effet est plus faible que la plus petite différence cliniquement pertinente.

Table 3 | Between group differences at all weeks*

Variables	Aerobic exercise *v* tai chi groups combined Mean (95% CI)	Aerobic exercise 2×24 weeks *v* tai chi 2×24 weeks Mean (95% CI)
FIQR score (range 0-100)†:		
Week 12	5.4 (0.6 to 10.1)	10.9 (3.4 to 18.5)
Week 24	5.5 (0.6 to 10.4)	16.2 (8.7 to 23.6)
Week 52	2.7 (−2.3 to 7.7)	11.1 (2.7 to 19.6)

L'effet estimé est +16.2 en faveur du tai chi (**2x24 semaines**)

L'intervalle de confiance à 95% va de 8.7 à 23.6: les données de l'échantillon permettent de dire que, en réalité, l'effet du tai chi peut être 8.7 ou 23.6.

Significativité statistique: l'hypothèse d'absence d'effet dans la population (vraie différence de moyenne = 0) est rejetée

L'intervalle de confiance à 95% ne contient pas la valeur 8.1 (la plus petite différence cliniquement pertinente): les données de l'échantillon permettent d'exclure que le véritable effet est plus faible que la plus petite différence cliniquement pertinente.

H_1: la pièce est tombe plus souvent sur face que pile (paramètre > 0.5)
H_0: la pièce a autant de chance de tomber sur pile ou face (paramètre = 0.5)

#	côté	probabilité	95%CI[1]	p-value[2]
1	face	1/2 = 0.500	0.05 to 1	0.500
2	face	1/4 = 0.250	0.22 to 1	0.250
3	face	1/8 = 0.125	0.37 to 1	0.125
4	face	1/16 = 0.062	0.47 to 1	0.062
5	face	1/32 = 0.031	0.55 to 1	0.031
6	face	1/64 = 0.016	0.61 to 1	0.016
7	face	1/128 = 0.008	0.65 to 1	0.008
8	face	1/256 = 0.004	0.69 to 1	0.004
9	face	1/512 = 0.002	0.72 to 1	0.002
10	face	1/1024=0.001	0.74 to 1	0.001

[1]méthode de Clopper-Pearson

[2]test binomial exact unilatéral

Table 3 | Between group differences at all weeks*

Variables	Aerobic exercise *v* tai chi groups combined		Aerobic exercise 2×24 weeks *v* tai chi 2×24 weeks	
	Mean (95% CI)	P value	Mean (95% CI)	P value
FIQR score (range 0-100)†:				
Week 12	5.4 (0.6 to 10.1)	**0.03**	10.9 (3.4 to 18.5)	**0.005**
Week 24	5.5 (0.6 to 10.4)	**0.03**	16.2 (8.7 to 23.6)	**<0.001**
Week 52	2.7 (−2.3 to 7.7)	0.29	11.1 (2.7 to 19.6)	**0.01**

L'effet estimé est +5.5 en faveur du tai chi (sur le score FIQR)

L'intervalle de confiance à 95% va de 0.6 à 10.4: les données de l'échantillon permettent de dire que, en réalité, l'effet du tai chi peut être 0.6 ou 10.4.

Significativité statistique: l'hypothèse d'absence d'effet dans la population (« vraie différence de moyenne = 0) est rejetée

L'intervalle de confiance à 95% contient 8.1 (la plus petite différence cliniquement pertinente): les données de l'échantillon ne permettent pas d'exclure que le véritable effet est plus faible que la plus petite différence cliniquement pertinente.

L'effet estimé est +16.2 en faveur du tai chi (**2x24 semaines**)

L'intervalle de confiance à 95% va de 8.7 à 23.6: les données de l'échantillon permettent de dire que, en réalité, l'effet du tai chi peut être 8.7 ou 23.6.

Significativité statistique: l'hypothèse d'absence d'effet dans la population (vraie différence de moyenne = 0) est rejetée

L'intervalle de confiance à 95% ne contient pas la valeur 8.1 (la plus petite différence cliniquement pertinente): les données de l'échantillon permettent d'exclure que le véritable effet est plus faible que la plus petite différence cliniquement pertinente.

- Un intervalle de confiance (ou p-value) sert à inférer l'ensemble des valeurs de paramètre compatibles avec les observations.
- Cette inférence n'est pertinente que pour le paramètre que l'on cherche à estimer.

Dans un RCT cherchant à démontrer une différence de niveau moyen de douleur entre un groupe expérimental et un groupe placebo:

- Paramètre à inférer: différence des moyennes de douleur des deux groupes
  → 95%CI (p-value)

- On ne cherche pas à inférer la paramètre d'âge moyen des deux groupes
  → pas de 95%CI (p-value) sur les moyennes d'âge des deux groupes

Les recommandations

The CONSORT statement (CONsolidated Standards Of Reporting Trials)

Tableau 1 (*Suite*)

Section/Topic	Item N°	Checklist item
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)

STROBE Statement—checklist of items that should be included in reports of observational studies

Results		
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included

PRISMA 2009 Checklist

RESULTS		
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.

CASP Critical Appraisal Skills Programme

CASP Checklist: 11 questions to help you make sense of a Randomised Controlled Trial

7. How large was the treatment effect?

HINT: Consider

- what outcomes were measured
- Is the primary outcome clearly specified
- what results were found for each outcome

8. How precise was the estimate of the treatment effect?

HINT: Consider

- what are the confidence limits

Consider:

– *if a confidence interval were reported. Would your decision about whether or not to use this intervention be the same at the upper confidence limit as at the lower confidence limit?*

Conclusion (1/2)

- L'intervalle de confiance à 95% est l'ensemble des valeurs du paramètre compatibles avec les données de l'échantillon
- L'estimation est toujours incluse dans l'IC95%
- La largeur de l'IC95% diminue lorsque la taille d'échantillon augmente
- L'intervalle de confiance à 95% existe pour toute estimation (moyenne, proportion, taux d'incidence, odds ratio, risque relatif, différence de moyennes, hazard ratio…)

Conclusion (2/2)

- L'IC 95%:
  - mesure la précision de l'estimation
  - capte l'incertitude liée à l'aléa de l'échantillonnage
  - ne mesure pas l'impact d'éventuel biais sur l'estimation
  - est utile pour tirer des conclusions sur la population
  - n'est pas pertinent pour décrire les données
  - ne contient pas toujours la valeur du paramètre (mais vous ne saurez pas si c'est le cas ou non dans votre étude!)

- Un·e chercheur·se doit
  - le comprendre
  - le rapporter
  - l'interpréter

L'IC95% est essentiel à l'interprétation des résultats d'une étude clinique

PLOS BIOLOGY

META-RESEARCH ARTICLE

Analysis of 567,758 randomized controlled trials published over 30 years reveals trends in phrases used to discuss results that do not reach statistical significance

Willem M. Otte[1,2], Christiaan H. Vinkers[3], Philippe C. Habets[3], David G. P. van IJzendoorn[4], Joeri K. Tijdink[5,6]*

Abstract

The power of language to modify the reader's perception of interpreting biomedical results cannot be underestimated. Misreporting and misinterpretation are pressing problems in randomized controlled trials (RCT) output. This may be partially related to the statistical significance paradigm used in clinical trials centered around a *P* value below 0.05 cutoff. Strict use of this *P* value may lead to strategies of clinical researchers to describe their clinical results with *P* values approaching but not reaching the threshold to be "almost significant." The question is how phrases expressing nonsignificant results have been reported in RCTs over the past 30 years.

Principales phrases associées à une valeur p > 0.05 (parmi 29'000 phrases étudiées)

Phrase	Total RCTs
Marginally significant	7,735
All but significant	7,015
A nonsignificant trend	3,442
Failed to reach statistical signifi[illegible]nc	2,5[illegible]
A strong trend	70[illegible]
Nearly significant	1,391
A clear trend	1,372
An increasing trend	1,202
Only marginally signif[illegible]an	1,[illegible]9
A significant trend	1,[illegible]4
Potentially significant	1,104
Significant tendency	1,064
A positive trend	1,055
A decreasing trend	[illegible]2
Marginal significance	[illegible]7
A slight trend	885
Almost significant	813
A statistical trend	[illegible]11
Approaching significance	7[illegible]
Nominally significant	7[illegible]
Quite significant	547
Near significant	546
An overall trend	445
Likely to be significant	425
Difference was apparent	409
Uncertain significan[illegible]	383
Di[illegible]ot quit[illegible] reach sta[illegible]st[illegible]r si[illegible]nce	379
A w[illegible]k [illegible]en[illegible]	343
Ma[illegible]inally statistically significant	314
Tended to be significant	293
Possible significance	286
No[illegible]e s[illegible]i[illegible]nt	26[illegible]
A[illegible]vorabl[illegible] trend	261
Just failed to reach statistical significance	252
A negative trend	225
Almost reac[illegible]d statistical significance	219
A[illegible]p[illegible]ibl[illegible]n[illegible]	21[illegible]
F[illegible]l shor[illegible] sig[illegible]fi[illegible]nc[illegible]	1[illegible]
Not as significant	204
A small trend	185
A numerical trend	184
[illegible]ig[illegible]tly sig[illegible]fi[illegible]n[illegible]	182
[illegible]eache[illegible] borde[illegible]n[illegible] s[illegible]gnifica[illegible]ce	165
Near significance	156
Weakly significant	147
Moderately significant	146
An apparent trend	145
Barely significant	135
Practically significant	135
A definite trend	131
An interesting trend	129
Almost statistically significant	126
Marginally nonsignificant	101
Possibly significant	100
Significantly significant	100

Interprétez les intervalles de confiance à 95% !