

Analyse sociologique de données d'enquête par questionnaire

Olivier Martin

Cours de l'UE « Méthodes d'investigation sociologique »

L3 de sciences sociales – Université Paris Descartes – 2011-2012

Séance 6 :

Comment étudier les relations
entre deux variables ?

Deux variables qualitatives (χ^2)

Rappels

→ Comment apprécier la relation entre deux variables ?

→ Cas le plus fréquent pour nous :

- Variables **qualitatives**
- Deux variables

→ Cas correspondant à un tableau croisé...

Tableau croisé

= représentation de la distribution des réponses selon deux critères (en ligne, en colonne)

L'attachement aux livres selon le sexe

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	504	312	37	11	864
homme	186	146	30	7	369
Total	690	458	67	18	1233

Rappels :

Quel outil pour apprécier cette relation ?

→ Le test du Khi^2 = test d'indépendance

→ La situation d'indépendance est la situation où les caractères en ligne n'ont aucune relation avec les caractères en colonne.

→ A la vue d'un tableau croisé, doit-on plutôt croire à l'hypothèse d'une indépendance ? Ou à l'hypothèse d'une dépendance entre les deux variables ?

Rappels : Comparaison de deux tableaux

Il s'agit de comparer deux tableaux :

- le tableau observé ou empirique

ie : le tableau obtenu par l'enquête

- le tableau d'indépendance

ie : le tableau théorique

Rappels : Comparaison de deux tableaux

Les différences entre les deux tableaux :

- 1) Résultent-elles de non-indépendance (dépendance) entre les deux variables ?
- 2) Résultent-elles des aléas de l'échantillonnage ?

La comparaison... et ses difficultés

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	485	320	45	14	864
homme	205	138	22	4	369
Total	690	458	67	18	1233

Proches...

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	483,5	320,9	46,9	12,6	864
homme	206,5	137,1	20,1	5,4	369
Total	690	458	67	18	1233

La comparaison... et ses difficultés

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	480	325	50	9	864
homme	210	133	17	9	369
Total	690	458	67	18	1233

Un peu moins proches...

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	483,5	320,9	46,9	12,6	864
homme	206,5	137,1	20,1	5,4	369
Total	690	458	67	18	1233

La comparaison... et ses difficultés

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	580	256	10	18	864
homme	110	202	57	0	369
Total	690	458	67	18	1233

Très différents...

	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	483,5	320,9	46,9	12,6	864
homme	206,5	137,1	20,1	5,4	369
Total	690	458	67	18	1233

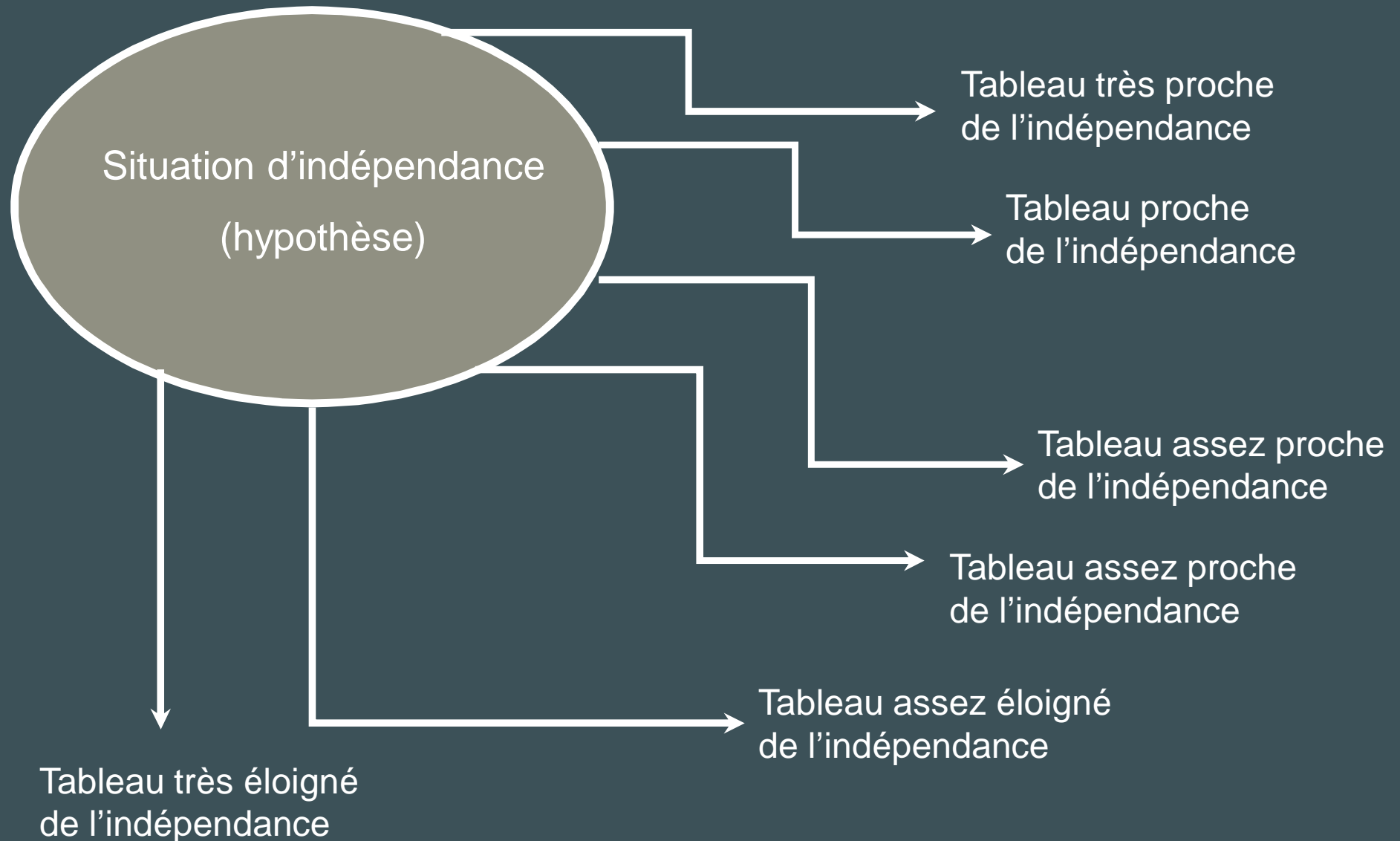
Comment parvenir à évaluer les effets des aléas ?

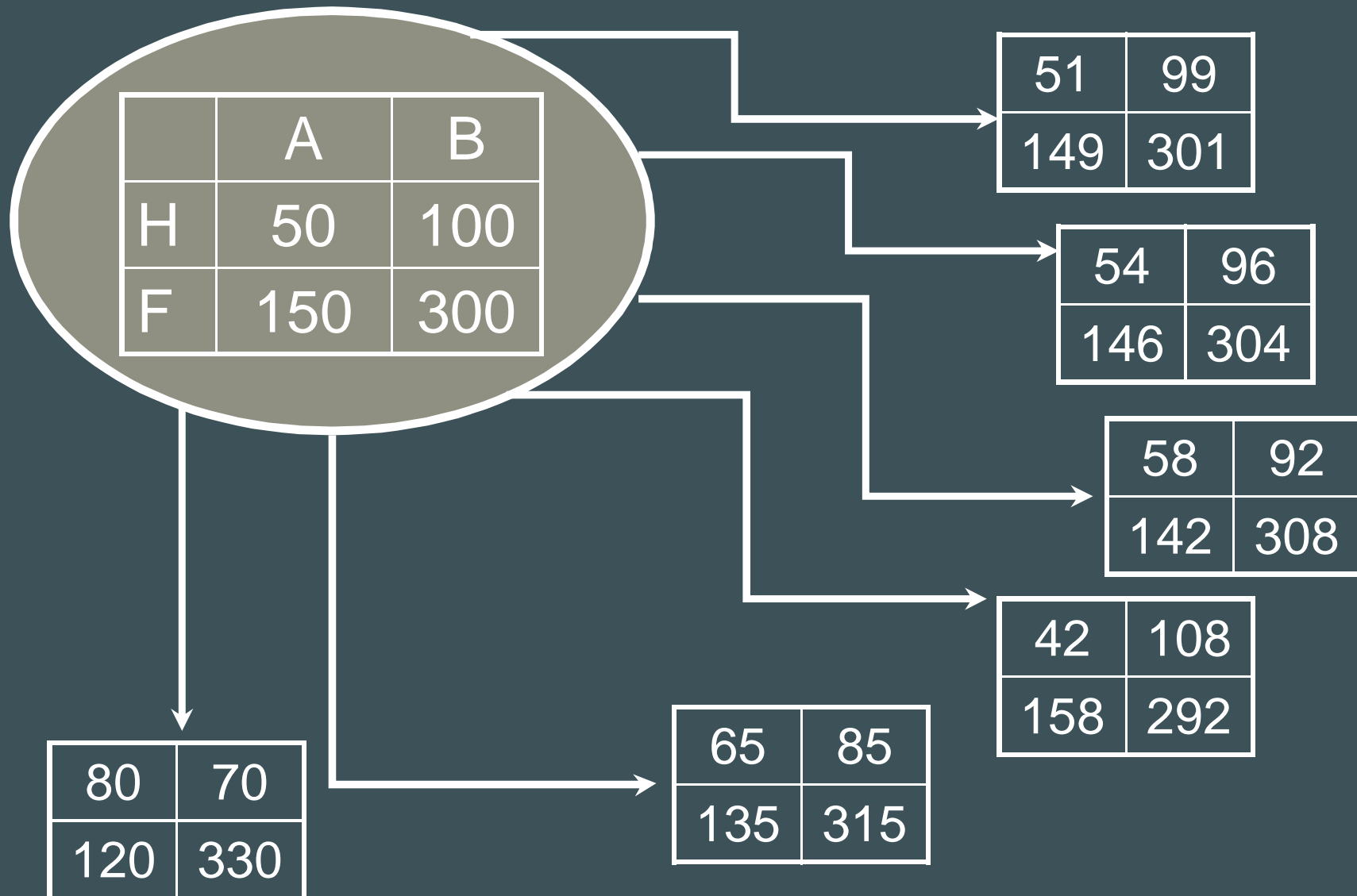
La science statistique permet de « résoudre » cette difficulté :

- Les tests statistiques...
- en l'occurrence, le test du khi-deux, (chi-deux, khi², chi², χ^2)

Raisonnement :

- 1) Faire l'hypothèse que les deux variables n'entretiennent aucune relation de dépendance.
- 2) En déduire quelles sont les formes possibles des tableaux d'individus choisis au sein d'une population où il aurait indépendance entre les deux variables.
- 3) Pour chacune de ces formes, déterminer si elle est très probable, assez probable, moyennement probable, peu probable, pas du tout probable...
- 4) Voir si notre tableau empirique fait parti des formes de tableaux qu'il est très probable, assez probable, peu probable, pas du tout probable...
- 5) Conclure sur la validité ou, plus exactement, la plausibilité (crédibilité) de cette hypothèse.





Intuitivement :

« plus le tableau sera éloigné de la situation d'indépendance plus faible sera sa probabilité de l'obtenir (si l'hypothèse d'indépendance est vérifiée) »

Probabilité :

- quelle est la probabilité d'obtenir ce tableau dans une population où les deux variables sont indépendantes ?

51	99
149	301

→ calcul → assez forte

- quelle est la probabilité d'obtenir ce tableau dans une population où les deux variables sont indépendantes ?

80	70
120	330

→ calcul → très faible

Autre présentation de ce raisonnement :

- 1) Faire l'hypothèse que les deux variables n'entretiennent aucune relation de dépendance.
- 2) En déduire quels sont les écarts possibles entre le tableau d'indépendance et les différents tableaux obtenus aléatoirement à partir d'une population où il aurait indépendance entre les deux variables.
- 3) Pour chacun de ces écarts, déterminer s'il est très probable, assez probable, moyennement probable, peu probable, pas du tout probable... Quelle est sa probabilité ?
- 4) Voir si l'écart entre notre tableau empirique et le tableau d'indépendance fait parti des écarts qu'il est très probable, assez probable, peu probable, pas du tout probable...
- 5) Conclure sur la validité ou, plus exactement, la plausibilité (crédibilité) de cette hypothèse.

	A	B
H	50	100
F	150	300

51	99
149	301

- Ecart faible
- Probable

54	96
146	304

- Ecart faible
- Probable

58	92
142	308

- Ecart assez faible
- Assez probable

42	108
158	292

- Ecart assez faible
- Assez probable

80	70
120	330

- Ecart très élevé
- Très peu probable

65	85
135	315

- Ecart assez élevé
- Peu probable

Difficultés « pratiques » de chacun de ces points :

- 1) Faire l'hypothèse que les deux variables n'entretiennent aucune relation de dépendance... ..
... Construire le tableau d'indépendance.

Difficultés « pratiques » de chacun de ces points :

2) En déduire quels sont les écarts possibles entre le tableau d'indépendance et les différents tableaux obtenus aléatoirement à partir d'une population où il aurait indépendance entre les deux variables... ..

... Pouvoir calculer ou simuler les écarts possibles entre le tableau d'indépendance et les tableaux obtenus par échantillonnage d'une population où il aurait indépendance entre les deux variables.

Pour calculer l'écart, on utilise une distance : la distance du khi². Cette distance est définie comme :

$$\text{distance du khi}^2 = \sum_{\text{cellules du tableau}} \frac{(\text{effectif observé} - \text{effectif théorique})^2}{\text{effectif théorique}}$$

Exemple d'un calcul de distance :

Tableau empirique					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	504	312	37	11	864
homme	186	146	30	7	369
Total	690	458	67	18	1233

Tableau d'indépendance					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	483,5036	320,9343	46,9489	12,6131	864
homme	206,4964	137,0657	20,0511	5,3869	369
Total	690	458	67	18	1233

Exemple d'un calcul de distance (suite) :

Tableau des différences					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	20,4964	-8,9343	-9,9489	-1,6131	0,0000
homme	-20,4964	8,9343	9,9489	1,6131	0,0000
Total	0,0000	0,0000	0,0000	0,0000	0,0000

Tableau des contributions à la distance du χ^2					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	0,8689	0,2487	2,1083	0,2063	
homme	2,0344	0,5824	4,9364	0,4831	
Total					

Exemple d'un calcul de distance (suite et fin) :

Tableau des contributions à la distance du χ^2					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	0,8689	0,2487	2,1083	0,2063	
homme	2,0344	0,5824	4,9364	0,4831	
Total					

$$\text{Distance} = 0,8689 + 0,2487 + 2,1083 + 0,2063 + \dots + 2,0344 + 0,5824 + 4,9364 + 0,4831 = 11,4684$$

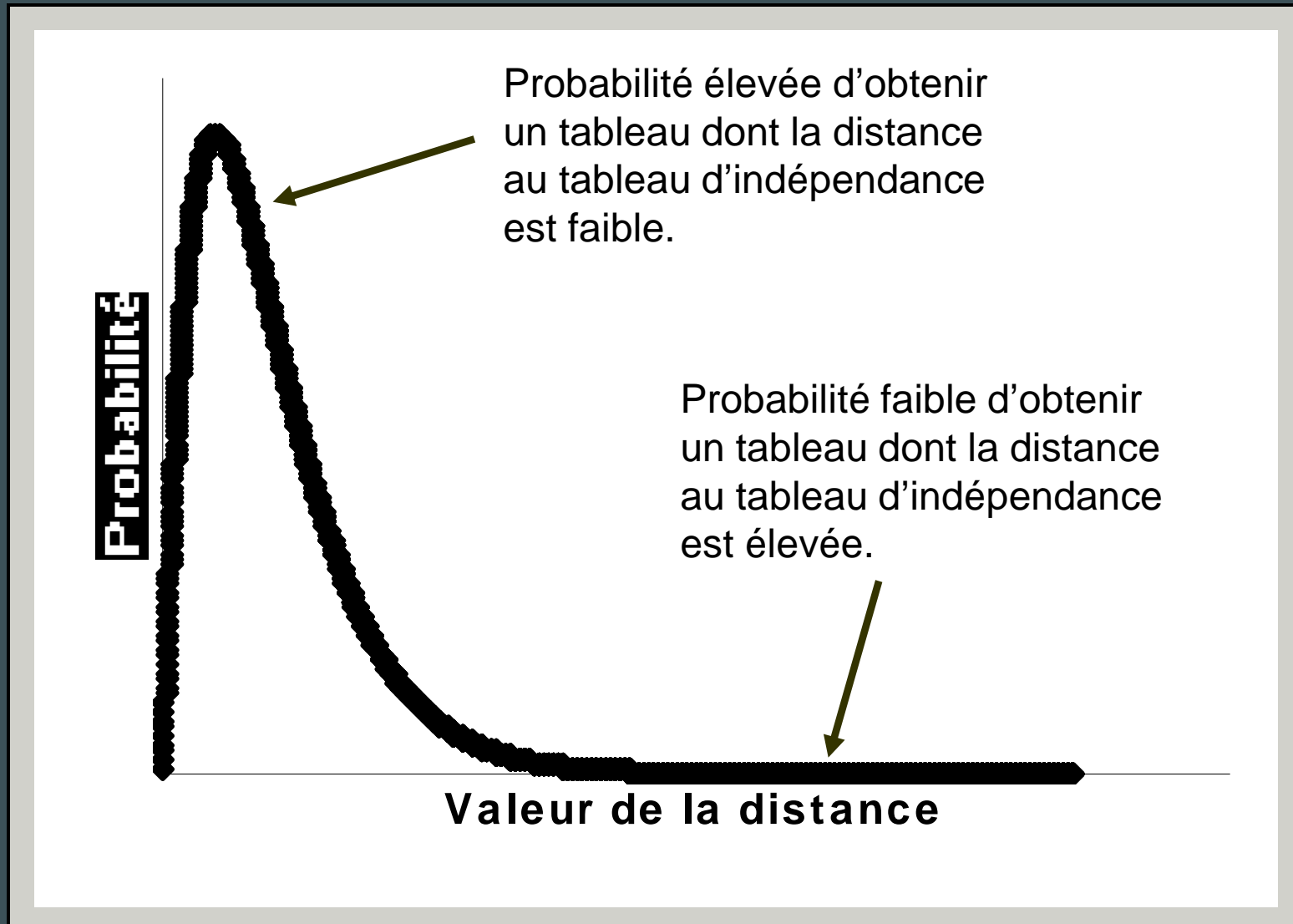
Difficultés « pratiques » de chacun de ces points :

- 3) Pour chacun de ces écarts, déterminer s'il est très probable, assez probable, moyennement probable, peu probable, pas du tout probable. Quelle est sa probabilité ?
... .. Pour chacun de ces écarts, calculer sa probabilité.

La simulation et plus sûrement le calcul probabiliste permet de faire ce calcul

Pour chaque tableau possible on peut calculer :

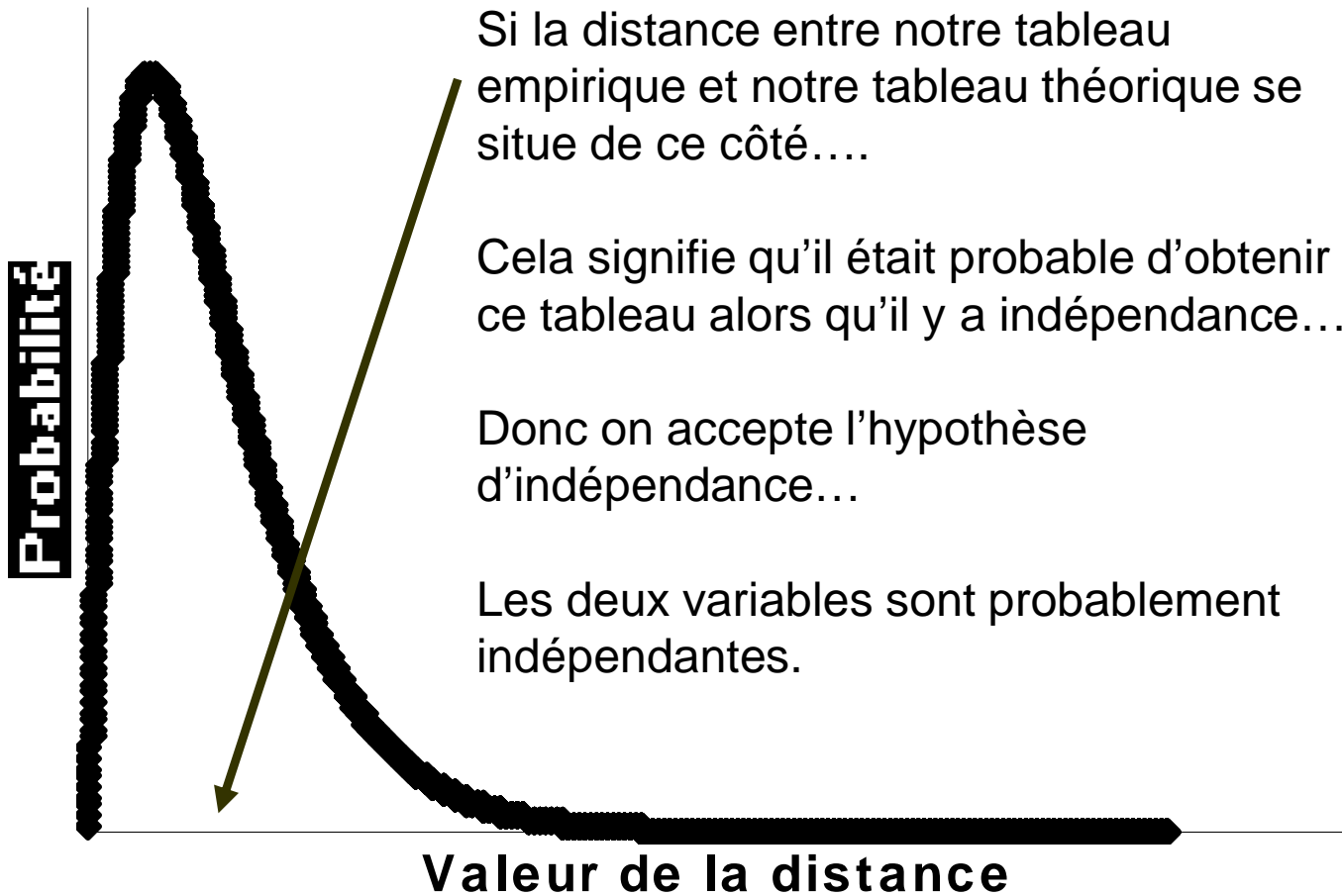
- **Son écart / sa distance au tableau d'indépendance...**
- **Sa probabilité (probabilité de l'obtenir s'il y a indépendance)...**
- **On peut donc tracer la courbe indiquant le niveau de probabilité en fonction de l'écart/distance...**



Difficultés « pratiques » de chacun de ces points :

- 4) Voir si l'écart entre notre tableau empirique et le tableau d'indépendance fait parti des écarts qu'il est très probable, assez probable, peu probable, pas du tout probable... ..
... Voir à quelle probabilité est associé l'écart entre notre tableau empirique et le tableau théorique.

**Où se situe la valeur de la distance
entre notre tableau empirique
et le tableau théorique d'indépendance ?**

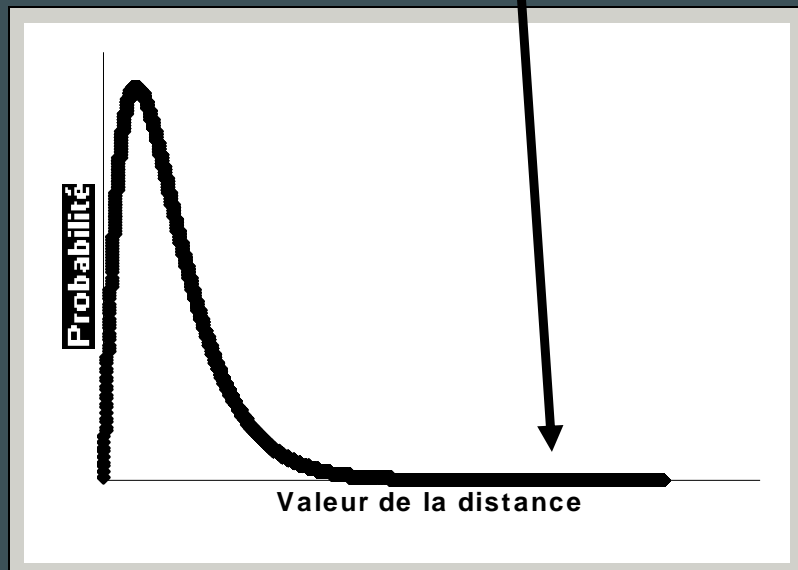


Si la distance entre notre tableau empirique et notre tableau théorique se situe de ce côté...

Cela signifie qu'il est peu probable d'obtenir ce tableau alors qu'il y a indépendance...

Donc... soit « on n'a pas eu de chance »... soit l'hypothèse d'indépendance n'est pas plausible...

Les deux variables sont probablement non-indépendantes : on considère les deux variables comme dépendantes.



Difficultés « pratiques » de chacun de ces points :

5) Conclure sur la validité ou, plus exactement, la plausibilité (crédibilité) de cette hypothèse. ...

Selon que la probabilité est très faible, faible, assez faible, assez forte, forte ou très forte, en déduire la validité ou, plus exactement, la plausibilité (crédibilité) de l'hypothèse d'indépendance.

Plusieurs cas :

- Probabilité faible : hypothèse d'indépendance peu plausible
→ dépendance
- Probabilité forte : hypothèse d'indépendance plausible
→ indépendance

Revenons à notre exemple :

Tableau empirique					
	tu aimes beaucoup les livres	tu aimes plutôt les livres	tu es indifférent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré tout	Total
femme	504	312	37	11	864
homme	186	146	30	7	369
Total	690	458	67	18	1233

$$\text{Distance} = 0,8689 + 0,2487 + 2,1083 + 0,2063 + \dots + 2,0344 + 0,5824 + 4,9364 + 0,4831 = 11,4684$$

Probabilité associée à cette distance = $0,009444708\dots \approx 0,01 = 1 \%$

Tableau peu probable... donc rejet de l'hypothèse d'indépendance... donc « dépendance » des variables.

En résumé...

- 1) On fait l'hypothèse que les deux variables sont indépendantes.
- 2) Sous cette hypothèse, pour chacun des écarts possibles à l'indépendance, on détermine si chacun de ces écarts est très probable, assez probable, moyennement probable, peu probable, pas du tout probable...
- 3) On en déduit la valeur de la probabilité pour la distance entre notre tableau empirique et le tableau d'indépendance...
- 5) On en conclut la valeur de plausibilité (crédibilité) de l'hypothèse initiale (indépendance).

En pratique...

1) On fait l'hypothèse que les deux variables sont indépendantes.

2) ORDINATEUR – STATISTICIEN – LOGICIEL
probable, pas du tout probable...

3) ORDINATEUR – STATISTICIEN – LOGICIEL

5) On en conclut la valeur de plausibilité (crédibilité) de l'hypothèse initiale (indépendance).

En pratique...

- 1) On réalise le tri croisé souhaité.
- 2) On calcule... ou plutôt on fait calculer la distance du khi² et surtout la probabilité associée.
- 3) Si la probabilité est faible (< 5 % ou < 10 %) on en déduit que les deux variables ne sont probablement pas indépendantes... Elles sont probablement dépendantes.

Exemple (sous Modalisa)

The screenshot shows a statistical software window titled "Tris croisés / ***** Les étudiants et leur livres - ...". The window displays a contingency table for the relationship between "45. Q37 sexe" and "43. Q35 Degré 'd'attachement' au livre". The table includes columns for response categories and a "Total" column. Below the table, the chi-square test results are shown: "Khi2=10,9 ddl=3 p=0,012 (Très significatif)".

	tu aimes beaucoup les livres	tu aimes plutot les livres	tu es indifferent aux livres	tu n'aimes pas les livres mais tu dois t'en servir malgré to	Total
femme	443	282	32	11	768
homme	173	131	29	7	340
Total	616	413	61	18	1108

Khi2=10,9 ddl=3 p=0,012 (Très significatif)

Distance

Probabilité

Probabilité faible ? Probabilité élevée ?

- La réponse dépend de la discipline...
- En sciences sociales / sociologie :
 - $p < 1\% = 0,01 \rightarrow$ Dépendance très probable
 - $p < 5\% = 0,05 \rightarrow$ Dépendance probable
 - $p < 10\% = 0,1 \rightarrow$ Dépendance assez probable
- On parle de significativité et on utilise des **** (étoiles).
 - $p < 1\% = 0,01 \rightarrow$ Très significatif = ***
 - $p < 5\% = 0,05 \rightarrow$ Significatif = **
 - $p < 10\% = 0,1 \rightarrow$ Assez significatif = *
- Attention, c'est une convention : elle peut varier d'une publication à l'autre, d'un auteur à l'autre, d'un échantillon à l'autre...

Le Khi² : une solution miracle ?

- La réponse est évidemment négative
- Limites du khi²
- Forces du khi²
- Conseils d'utilisation