TRAITEMENT ET INTERPRÉTATION DES DONNÉES

Contenu du cours

- Les données chiffrées (sondages, statistiques sur l'emploi, etc.) envahissent l'espace public où elles sont un argument majeur des débats.

- Dans les sciences humaines et les sciences du langage, l'utilisation de données quantitatives est de plus en plus courante.

- Ce cours est une initiation « sans mathématiques » aux méthodes destinées à recueillir et traiter ce type de données. Il propose également une réflexion sur leur interprétation et leur utilisation dans la société. Il prendra appui sur des exemples concrets et l'usage direct de logiciels de statistique.

1

TRAITEMENT ET INTERPRÉTATION DES DONNÉES

Contenu- Partie 1 : Utilisation des données chiffrées

Illustration 1 : Fumer tue ?

- Comment les données chiffrées ont participé à la « guerre du tabac »

- Débats scientifiques qui ont commencé dans les années 1950 pour établir la nocivité du tabac.

- Poursuite des enquêtes :
  - Tabagisme passif tue ?
  - Fumer du tabac à rouler tue ?
  - Fumer moins de 5 cigarettes/jour tue ?

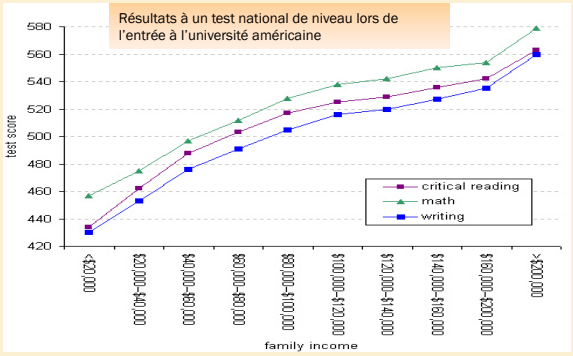
2

TRAITEMENT ET INTERPRÉTATION DES DONNÉES

Contenu– Partie 1 (suite)

Illustration 2 : quelles sont les facteurs de l'intelligence et de la réussite scolaire ? (si on a le temps)

- Comment les données chiffrées ont participé au débat ?
- Rôle de l'éducation et de l'hérédité



http://economix.blogs.nytimes.com/2009/08/27/sat-scores-and-family-income/

Blog *Economix* du New York Times – 9/09/10

3

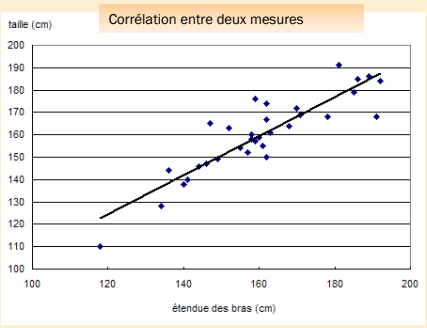
TRAITEMENT ET INTERPRÉTATION DES DONNÉES

Contenu– Partie 2 : Panorama des méthodes

Quelles sont les méthodes qui permettent de recueillir, représenter, traiter des données chiffrées en SHS ?

- à partir d'exemples de recherches qui sont menées dans l'UFR de SDL
- souvent des travaux sur l'acquisition du langage.

Juste un panorama !



http://www.statcan.gc.ca/pub/12-593-x/2007001/figures/4183030-fra.htm- 8/09/10

4

TRAITEMENT ET INTERPRÉTATION DES DONNÉES

Contenu- Partie 3 : TD sur le logiciel StatView 5.0

Logiciel simple, facile d'accès

- Décrire une série de données
- Généraliser un résultat de l'échantillon à la population
- Analyser un tableau
- Comparer des moyennes
- Estimer une relation entre deux séries de données

Statview « free download » ?

5

Partie 1.
Utilisation des données chiffrées
Usages sociaux

1 . UTILISATION DES DONNÉES CHIFFRÉES DANS LA SOCIÉTÉ
1.1 La « guerre du tabac »
1.1.1 La lutte contre le tabac est devenue une guerre de chiffres

A/ La prévention contre la consommation de tabac est fondée sur des grandes enquêtes qui estiment le nombre de décès « dus au tabac »

> « Les conséquences du tabagisme sur la santé ont été formellement démontrées dès les années 1950 et sont maintenant bien connues. **Le nombre de décès dus au tabac est estimé aujourd'hui à 548 000 par an dans l'Union européenne et à 60 000 en France, soit plus d'un décès sur neuf**. Au vu des tendances passées et actuelles de consommation, des prévisions pour 2025 évaluent le nombre de morts liées au tabac, pour la France, à 160 000 par an dont 50 000 chez la femme, soit 10 fois plus qu'aujourd'hui. »

http://www.sante.gouv.fr/pdf/actu/tabac.pdf - 9/09/10

B/ A partir de ces enquêtes, on estime le risque pour tout fumeur de mourir de son tabagisme

> « La lutte contre le tabagisme est une priorité de santé publique : un fumeur régulier sur deux meurt du tabac »

http://www.tabac.gouv.fr/ - 9/09/10

7

C/ Le tabac est présenté comme la cause de certaines maladies

« Le tabac est la principale cause de décès liés au cancer en France, avec plus de 35 000 décès par an. »

« 83 % des cancers des poumons chez les hommes et 69 % chez les femmes sont attribuables au tabac. »

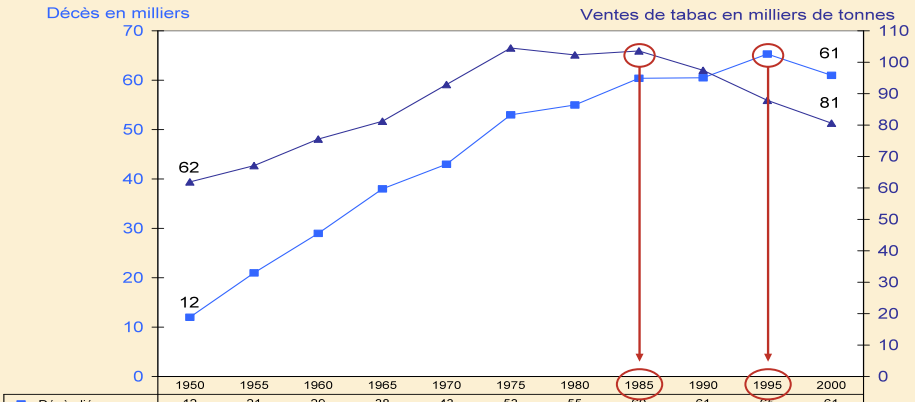
http://www.gouvernement.fr/gouvernement/le-tabac-premier-facteur-de-risque-de-cancer - 09/09/10

> « En 2000, on estime que 33 000 décès annuels par cancer seraient consécutifs au tabac, dont environ 20 500 cancers du poumon. Le tabac causerait encore 5000 décès par bronchite chronique obstructive, 11 000 décès par maladies cardio-vasculaires et 11 600 décès à la suite d'autres pathologies. »

http://www.statcan.gc.ca/pub/12-593-x/2007001/figures/4183030-fra.htm - 08/09/10
Observatoire français des drogues et des toxicomanies

8

D/ On met en relation nombre de décès liés au tabac et volume des ventes de tabac



Sources : Registre national des causes de décès, CépiDC, exploitation CTSU, université of Oxford et Ventes de tabac, Altadis

« La courbe des décès liés au tabac suit avec 10 à 20 ans de retard celle des ventes de tabac ».

http://www.statcan.gc.ca/pub/12-593-x/2007001/figures/4183030-fra.htm - 08/09/10

e/ Conclusion : caractéristiques de la lutte anti-tabac depuis les 60's

Dimension quantitative

- Utilisation de données chiffrées
- Idée que le tabac est responsable d'un « meurtre de masse »
→ Termes utilisés dans les articles scientifiques : The Big Kill (*la grande tuerie*), The growing brown plague (*l'extension de la peste brune)*

Consensus des scientifiques sur la nocivité du tabac et autorité de la science

Le tabagisme n'est plus un vice mais une maladie
« Fumer, c'est mal » (un vice) → « Fumer, ça fait mal » (une maladie)

Utilisation du vocabulaire de la médecine
- « Epidémie mondiale de tabagisme »
- « Incidence du tabagisme »
Taux d'incidence = nouveaux cas d'une maladie durant une période donnée rapportés à la population

Ambiguïté : les fumeurs sont-ils responsables ou victimes ?

1.1.2 Une distinction importante : association, cause, facteur

Discipline des sciences médicales impliquée : l'épidémiologie

Branche de la science médicale qui s'occupe de l'étude des **facteurs** influant sur la santé et les maladies des populations humaines.

Facteur : notion importante qu'on retrouve dans toutes les sciences
A distinguer de :
- **association statistique**
- **cause**

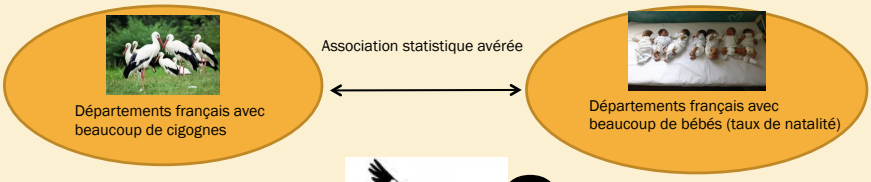
11

1.1.2 Une distinction importante : association, cause, facteur

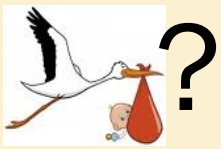
A/ Association statistique entre deux évènements : deux évènements arrivent généralement en même temps ou l'un après l'autre.

Chute du 2e étage de la Tour Eiffel et décès
Chant du coq puis lever du soleil
Consommation de cigarettes et cancer du poumon
Consommation de cigarette et consommation de café

L'association de deux événements ne signifie pas que l'un est la cause et l'autre la conséquence (lien de causalité)



Relation cause-conséquence
Cigognes → naissances ?

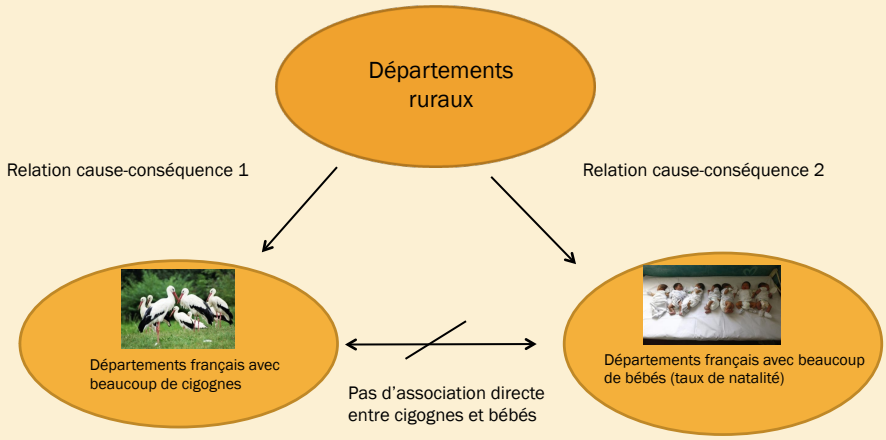


Indice de la croyance populaire : les cigognes amènent les bébés…

12

1.1.2 Une distinction importante : association, cause, facteur

La relation entre cigognes et bébés passe par un troisième élément….



13

B/ relation cause-conséquence catégorique entre deux évènements : trois cas de figure

b1 – L'événement X est une condition suffisante de l'événement Y

- il suffit que X arrive/soit vrai pour que Y arrive/soit vrai
- impossible d'avoir X sans Y, on peut toutefois avoir Y sans X

Exemple :

X = Avoir moins de 10 à chacune de ses évaluations ; Y = rater son semestre

b2 – L'événement X est une condition nécessaire de l'événement Y

- il est nécessaire que X arrive/soit vrai pour que Y arrive/soit vrai
- impossible d'avoir Y sans X, on peut toutefois avoir X sans Y

Exemple :

X = ne pas avoir de note éliminatoire à ses évaluations; Y = réussir son semestre

b3 – L'événement X est une condition nécessaire et suffisante de l'événement Y

- Si X arrive/est vrai, alors Y arrive/est vrai, et vice versa.

Exemple :

X = (Ne pas avoir de note éliminatoire + une moyenne de 10) ; Y =réussir son semestre

14

C/ La notion de facteur (l'événement X est un facteur de l'événement Y)

Par exemple :
- facteurs de risque / facteur de protection pour une maladie : tabac → maladies
- facteurs de l'échec scolaire
- facteurs des conduites à risque chez l'adolescent

Le facteur X augmente/diminue la probabilité de l'événement Y
- Fumer augmente le risque d'avoir certaines maladies
- Faire du sport protège contre certaines de ces maladies

Un facteur X n'est une condition ni nécessaire ni suffisante de l'événement Y
- On peut avoir X (fumer) sans Y (cancer) : certains fumeurs n'ont pas le cancer des poumons
- On peut avoir Y (cancer) sans X (fumer) : certaines personnes atteintes du cancer du poumon n'ont jamais fumé

Questions : Quels sont les critères pour décider qu'une association entre deux événements est de type cause-conséquence (pour décider que X et un facteur causal de Y) ?

15

Critères pour décider d'un lien causal (cause-conséquence) entre X et Y

Austin Bradford Hill, "The Environment and Disease: Association or Causation?," Proceedings of the Royal Society of Medicine, 58 (1965), 295-300.

1/ Existence d'une association statistique

Davantage de cancers du poumons chez les fumeurs vs. Les non-fumeurs

2/ L'association doit être forte

Risque d'avoir un cancer du poumon, comparaison non fumeurs et fumeurs de plus de 25 cigarettes / jour

3/ Existence d'une relation de type "dose-effet"

Plus on fume (beaucoup, longtemps), plus le risque est important

4/ X doit précéder Y dans le temps

En général, la cause doit précéder la conséquence !

16

5/ Concordance entre les résultats d'études menées avec des méthodes différentes, dans des régions ou des populations différentes

Lien entre tabac et maladies étudié dans de nombreuses situations

6/ Diminution ou disparition de Y quand X est supprimé ou diminué

Arrêt de la cigarette → diminution du risque

7/ On doit pouvoir expliquer la relation entre X et Y par un mécanisme plausible

Trouver le mécanisme biologique qui va de l'inhalation de fumée à la transformation des cellules saines en cellules cancéreuses

17

Mini-TD numéro 1

Pour les évènements X et Y suivants dire si l'association entre eux est de type :
- fortuite (de type cigogne → bébé)
- X condition nécessaire de Y
- X condition suffisante de Y
- X condition nécessaire et suffisante de Y
- X facteur causal de Y (X augmente la probabilité de Y)

X	Y
Chant du coq	Lever du soleil
Chute du 2e étage de la Tour Eiffel	Décès
Absentéisme aux cours	Echec pour l'obtention de la licence
Consommation de cigarette (fumeur)	Consommation de café (amateur de café)
Revenus parentaux	Réussite scolaire

18

Correction mini-TD 1

X	Y	Type d'association
Chant du coq	Lever du soleil	**Y est une condition suffisante de X** *« Le lever du soleil suffit à faire chanter le coq »*
Chute du 2e étage de la Tour Eiffel	Décès	**X est une condition suffisante de Y** *« Chuter du 2ème étage de la TE est suffisant pour décéder »* **facteur causal (hauteur chute)** *« Chuter du 2ème étage augmente le risque de décès »*
Absentéisme aux cours	Echec pour l'obtention de la licence	**X facteur causal de Y** *«L'absentéisme augmente les risques d'échec pour l'obtention de la licence »*
Consommation de cigarette (fumeur)	Consommation de café (amateur de café)	**X facteur causal de Y** *« La consommation de cigarette augmente la consommation de café chez un fumeur »* = des addictions **Association fortuite** *« Le fumeur ne consomme pas obligatoirement du café »*
Revenus parentaux	Réussite scolaire	**X facteur causal de Y** *« La catégorie socio-professionnelle des parents augmente/diminue la réussite scolaire de leur enfant »*

1 . UTILISATION DES DONNÉES CHIFFRÉES DANS LA SOCIÉTÉ
1.1 La « guerre du tabac »
1.1.1 La lutte contre le tabac est devenue une guerre de chiffres
1.1.2 Une distinction importante : association, cause, facteur
1.1.3 Exemple d'enquête : fumer 1 à 4 cigarettes par jour, conséquences pour la santé ?

Bjartveit K, Tverdal A. (2005). Health consequences of smoking 1-4 cigarettes per day. *Tobacco Control* 14: 315-320.

A/ La question de recherche :

Y a t-il une valeur de seuil en dessous de laquelle la cigarette n'est pas nocive ?

➢ Les travaux disponibles font apparaître une forte relation dose-effet (= lien entre nbr de cigarettes fumées/jour et l'apparition ou la présence de maladies)

➢ En général, le groupe de faible consommation regroupe 1-9 ou 1-15 cigarettes…

➢ Et si on diminue encore la dose, par exemple de 1-4 cigarettes, y a-t-il moins ou pas de risque pour la santé ?

20

B/ Les participants à l'enquête

1/ Utilisation d'un échantillon pour le dépistage des maladies cardiovasculaires constitué en 1972-78		
- Tous les sujets de 40-49 ans - et 7 à 10 % des 20-39 ans, - habitant Oslo et 3 provinces norvégiennes	- mesure poids, taille, tension, cholestérol, triglycérides, glucose - auto-estimation nombre cigarettes/jour : 0, 1- 4, 5-9,...., 25 et + - auto-estimation activité physique : échelle 1-4	
2/ On garde les 35-49 ans (en 1972-1978) et on enlève ceux qui, au départ, présentaient certaines caractéristiques		
- maladie cardiaque, AVC, hypertension, artériosclérose - fumeurs de pipe et de cigares, ex-fumeurs		
3/ Il reste	23 521 hommes	19 201 femmes

21

C/ Recueil des données sur les décès et les maladies

L'examen initial des participants a eu lieu entre 1972 et 1978 et la cohorte est suivi jusqu'en 2002.

Grace au *Registre national des causes de décès*, on note deux informations pour chaque individu suivi

- si l'individu est décédé
  - et si c'est le cas, pour quelle(s) raison(s) il est décédé
- le nombre d'années entre l'examen initial et :
  - le décès
  - le 31 décembre 2002
  - la date de son émigration (sortie du registre national)

22

Mini-TD numéro 2

En utilisant les informations des deux diapos précédentes, calculer :

1) l'âge minimum des sujets concernés par l'étude ;
2) l'âge maximum des sujets concernés par l'étude.

Faire ce calcul :
- d'une part pour les sujets décédés pendant le suivi
- et d'autre part pour les sujets toujours vivants en 2002.

23

Mini-TD numéro 2

Sujets décédés pendant l'étude			
	Âge d'entrée	Âge décès	Nombre d'années de survie pendant le suivi
Plus jeune			
Plus vieux			

Sujets toujours en vie à la fin de l'étude			
	Âge d'entrée	Âge en 2002	Nombre d'années de survie pendant le suivi
Plus jeune			
Plus vieux			

24

Mini-TD numéro 2 **CORRECTION**

Sujets décédés pendant l'étude			
	Âge d'entrée	**Âge décès**	**Nombre d'années de survie pendant le suivi**
Plus jeune	35 ans en 1978	35 ans	0 année
Plus vieux	49 ans en 1972	79 ans en 2002	30 années
Sujets toujours en vie à la fin de l'étude			
	Âge d'entrée	**Âge en 2002**	**Nombre d'années de survie pendant le suivi**
Plus jeune	35 ans en 1978	59 ans	24 années
Plus vieux	49 ans en 1972	79 ans	30 années

25

Mini-TD numéro 2 **CORRECTION**

L'âge minimal des sujets concernés par l'étude est 35 ans.

Sujets décédés pendant l'étude			
	Âge d'entrée	Âge décès	Nombre d'années de survie pendant le suivi
Plus jeune	35 ans	à 35 ans	0 année
Plus vieux	49 ans en 1972	à 79 ans en 2002	30 années
Sujets toujours en vie à la fin de l'étude			
	Âge d'entrée	Âge en 2002	Nombre d'années de survie pendant le suivi
Plus jeune	35 ans en 1978	59 ans	24 années
Plus vieux	49 ans en 1972	79 ans en 2002	30 années

Dans tous les cas, vivant ou mort, l'âge maximal des sujets concernés est 79 ans.

26

D/ Résultat 1 – Caractéristique des sujets lors de l'examen initial *(baseline)*

Table 1 Baseline characteristics of 23521 male and 19201 female participants.* Mean values, by cigarette consumption recorded at screening†

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Age (years)	42.3	43.5	43.2	43.0	42.6	42.6	42.4
Duration of smoking (years)		18.7	21.6	22.7	22.8	23.3	23.9
Systolic BP (mm Hg)	134.9	134.8	135.2	135.2	135.4	135.2	135.2
Diastolic BP (mm Hg)	85.6	84.9	84.2	84.1	84.1	85.0	85.5
Total serum cholesterol (mmol/l)	6.65	6.67	6.97	7.04	7.04	7.14	7.13
Serum triglycerides (mmol/l)	2.16	2.28	3.31	2.36	2.39	2.46	2.44
Serum glucose (mmol/l)	5.79	5.77	5.78	5.81	5.82	5.80	5.82
Physical activity leisure‡	2.22	2.13	2.09	2.04	1.97	1.91	1.79
Physical activity work‡	2.20	2.17	2.43	2.39	2.27	2.23	2.15
BMI (kg/m^2)	24.9	24.7	24.5	24.5	24.6	24.9	25.2
Height (cm)	176.4	176.3	175.0	175.4	175.9	175.9	176.4
Females							
Age (years)	42.2	42.3	42.0	41.5	41.4	41.5	41.0
Duration of smoking (years)		12.5	16.3	18.3	19.5	19.5	20.1
Systolic BP (mm Hg)	132.0	129.3	130.0	129.0	128.9	128.6	128.7
Diastolic BP (mm Hg)	81.4	80.1	80.0	79.5	79.7	80.5	81.4
Total serum cholesterol (mmol/l)	6.63	6.74	6.88	6.92	6.95	6.99	7.00
Serum triglycerides (mmol/l)	1.55	1.64	1.74	1.76	1.76	1.84	1.91
Serum glucose (mmol/l)	5.76	5.69	5.71	5.65	5.69	5.71	5.65
Physical activity leisure‡	1.91	1.93	1.89	1.85	1.77	1.74	1.75
Physical activity work‡	2.22	2.18	2.12	2.10	2.01	2.03	2.08
BMI (kg/m^2)	25.0	24.7	24.0	23.7	23.7	23.7	24.5
Height (cm)	162.5	162.7	162.1	162.6	162.8	163.5	163.0

*Participants not reporting cardiovascular disease, diabetes or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.
†Number of participants within the consumption groups (see table 2).
‡Physical activity during leisure and at work was graded 1–4, with 4 as the heaviest activity.
BMI, body mass index; BP, blood pressure.

Question :
Essayez de comprendre comment le tableau est construit…
- **Titre**
- **Intitulés lignes et colonnes**
- **Intitulés subdivisions**

27

D/ Résultat 1 – Caractéristique des sujets lors de l'examen initial *(baseline)*

Table 1 Baseline characteristics of 23521 male and 19201 female participants.* Mean values, by cigarette consumption recorded at screening†

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Age (years)	42.3	43.5	43.2	43.0	42.6	42.6	42.4
Duration of smoking (years)		18.7	21.6	22.7	22.8	23.3	23.9
Systolic BP (mm Hg)	134.9	134.8	135.2	135.2	135.4	135.2	135.2
Diastolic BP (mm Hg)	85.6	84.9	84.2	84.1	84.1	85.0	85.5
Total serum cholesterol (mmol/l)	6.65	6.67	6.97	7.04	7.04	7.14	7.13
Serum triglycerides (mmol/l)	2.16	2.28	3.31	2.36	2.39	2.46	2.44
Serum glucose (mmol/l)	5.79	5.77	5.78	5.81	5.82	5.80	5.82
Physical activity leisure‡	2.22	2.13	2.09	2.04	1.97	1.91	1.79
Physical activity work‡	2.20	2.17	2.43	2.39	2.27	2.23	2.15
BMI (kg/m^2)	24.9	24.7	24.5	24.5	24.6	24.9	25.2
Height (cm)	176.4	176.3	175.0	175.4	175.9	175.9	176.4
Females							
Age (years)	42.2	42.3	42.0	41.5	41.4	41.5	41.0
Duration of smoking (years)		12.5	16.3	18.3	19.5	19.5	20.1
Systolic BP (mm Hg)	132.0	129.3	130.0	129.0	128.9	128.6	128.7
Diastolic BP (mm Hg)	81.4	80.1	80.0	79.5	79.7	80.5	81.4
Total serum cholesterol (mmol/l)	6.63	6.74	6.88	6.92	6.95	6.99	7.00
Serum triglycerides (mmol/l)	1.55	1.64	1.74	1.76	1.76	1.84	1.91
Serum glucose (mmol/l)	5.76	5.69	5.71	5.65	5.69	5.71	5.65
Physical activity leisure‡	1.91	1.93	1.89	1.85	1.77	1.74	1.75
Physical activity work‡	2.22	2.18	2.12	2.10	2.01	2.03	2.08
BMI (kg/m^2)	25.0	24.7	24.0	23.7	23.7	23.7	24.5
Height (cm)	162.5	162.7	162.1	162.6	162.8	163.5	163.0

*Participants not reporting cardiovascular disease, diabetes or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.
†Number of participants within the consumption groups (see table 2).
‡Physical activity during leisure and at work was graded 1–4, with 4 as the heaviest activity.
BMI, body mass index; BP, blood pressure.

Commentaires (tels que donnés dans le tableau)

Les fumeurs « tempérés » (*light smokers* : 1-4 cigarettes/ jour) fument depuis moins longtemps.

Chez les sujets des deux sexes, **augmentation du cholestérol et des triglycérides** lorsqu'augmente la consommation journalière de tabac.

Chez les sujets des deux sexes, **diminution de l'activité physique** estimée pendant les loisirs lorsqu'augmente la dose journalière de cigarettes.

28

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Table 2 Number of participants and person years; deaths from all causes, ischaemic heart disease, all cancer, and lung cancer, number and per 100000 person years, by number of cigarettes recorded at screening. 23521 male and 19201 female participants aged 35–49*

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Number of participants	8308	627	2526	4941	3401	2599	1119
Number of person years	219699	16367	63464	122128	82836	61967	26310
All causes							
Number of deaths	1248	161	760	1697	1245	1084	499
Per 100000 person years	568	984	1198	1390	1503	1749	1897
Ischaemic heart disease							
Number of deaths	287	63	233	521	406	311	128
Per 100000 person years	131	385	367	427	490	502	487
All cancer							
Number of deaths	446	40	226	525	392	349	171
Per 100000 person years	203	244	356	430	473	563	650
Lung cancer							
Number of deaths	17	4	61	168	126	146	70
Per 100000 person years	8	24	96	138	152	236	266
Females							
Number of participants	11077	796	2759	3005	1008	477	79
Number of person years	288178	20537	70509	76158	25196	11720	2036
All causes							
Number of deaths	956	98	427	530	222	113	16
Per 100000 person years	332	477	606	696	881	964	786
Ischaemic heart disease							
Number of deaths	85	17	72	79	36	13	2
Per 100000 person years	29	83	102	104	143	111	98
All cancer							
Number of deaths	579	46	197	264	100	58	9
Per 100000 person years	201	224	279	347	397	495	442
Lung cancer							
Number of deaths	14	5	43	65	33	14	3
Per 100000 person years	5	24	61	85	131	119	147

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

Question 1
Essayez de comprendre comment le tableau est construit…

Question 2
Pourquoi y a-t-il plus de décès par cancer du poumon chez les hommes qui fument 10-14 cigarettes/ jours que chez ceux qui en fument 25+ ?

29

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Réponse à la question 2 : début

Plus de cancer du poumon chez les hommes qui fument 10-14 c/jour que chez ceux qui fument 25+
Mais quatre fois plus d'hommes qui fument 10-14 c/jour que d'hommes qui fument 25+

→ Pour comparer sur la même base : exprimer le nombre de décès comme un pourcentage de l'effectif des groupes.

	Hommes Fumeurs 10-14 c/jour	Hommes Fumeurs 25+ c/jour
Nombre de décès par cancer du poumon	168	70
Nombre de personnes	4941	1119
Pourcentage Décès / personne	(168 / 4941) x100 = **3,4 %**	(70 / 1119) x100 = **6,2 %**

30

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Réponse à la question 2 : suite car autre problème pour comparer les groupes....

L'importance des cas de décès dans un groupe dépend de la longueur du suivi des sujets		
Exemple 1	100 sujets suivis pendant **1 an**	5 décès
Exemple 2	100 sujets suivis pendant **10 ans**	5 décès
A quel groupe préférez-vous appartenir ?		

Problème : la longueur du suivi diffère selon les sujets, pour deux raisons :
- certains sont entrés dans la cohorte en 1972, d'autres en 1978
- certains sont décédés lors du suivi, d'autres pas : plus un groupe est à risque, plus le suivi est court en moyenne !

Question

Si on ne tient pas compte de la longueur du suivi, quelles en sont les conséquences pour la comparaison entre « gros » fumeurs et non fumeurs ?

31

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Solution au problème

On rapporte le risque au nombre de sujets X nombres d'années de suivi
= le **nombre de personnes années** (*number of person years*)

Calcul du nombre de personnes années

Sujets	Durée du suivi en années
S1	30
S2	30
S3	30
S4	28
S5	28
S6	28
S7	15
S8	15
S9	10
...	...
Sn	etc.

⟹ (3 x 30) + (3 x 28) + (2 x 15) + (1 x 10) + ...

Calcul de la probabilité de décès

	Hommes Fumeurs 10-14 c/jour	Hommes Fumeurs 25+ c/jour
Nombre de personnes.années	122 128	26 310
Nombre de décès par cancer du poumon	168	70
Probabilité décès / personnes.années	168 / 122128 = 0.00138 Soit 0.137 %	70 / 26310 = 0.00266 Soit 0.266 %
Ramené à 100 000 pesonnes années	0.00138 X 100 000 = **138 pour 100 000**	0.00266 X 100 000 = **266 pour 100 000**

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Commentaires du tableau 2 (tels que donnés dans l'article)

Table 2 Number of participants and person years; deaths from all causes, ischaemic heart disease, all cancer, and lung cancer, number and per 100000 person years, by number of cigarettes recorded at screening. 23521 male and 19201 female participants aged 35–49*

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Number of participants	8308	627	2526	4941	3401	2599	1119
Number of person years	219699	16367	63464	122128	82836	61967	26310
All causes							
Number of deaths	1248	161	760	1697	1245	1084	499
Per 100000 person years	568	984	1198	1390	1503	1749	1897
Ischaemic heart disease							
Number of deaths	287	63	233	521	406	311	128
Per 100000 person years	131	385	367	427	490	502	487
All cancer							
Number of deaths	446	40	226	525	392	349	171
Per 100000 person years	203	244	356	430	473	563	650
Lung cancer							
Number of deaths	17	4	61	168	126	146	70
Per 100000 person years	8	24	96	138	152	236	266
Females							
Number of participants	11077	796	2759	3005	1008	477	79
Number of person years	288178	20537	70509	76158	25196	11720	2036
All causes							
Number of deaths	956	98	427	530	222	113	16
Per 100000 person years	332	477	606	696	881	964	786
Ischaemic heart disease							
Number of deaths	85	17	72	79	36	13	2
Per 100000 person years	29	83	102	104	143	111	98
All cancer							
Number of deaths	579	46	197	264	100	58	9
Per 100000 person years	201	224	279	347	397	495	442
Lung cancer							
Number of deaths	14	5	43	65	33	14	3
Per 100000 person years	5	24	61	85	131	119	147

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

Conclusion 1

Chez les hommes et les femmes, quelle que soit la cause du décès, les probabilités de décès sont plus grandes pour les fumeurs de 1-4 cigarettes que pour les non-fumeurs

33

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Commentaires du tableau 2 (tels que donnés dans l'article)

Table 2 Number of participants and person years; deaths from all causes, ischaemic heart disease, all cancer, and lung cancer, number and per 100000 person years, by number of cigarettes recorded at screening. 23521 male and 19201 female participants aged 35–49*

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Number of participants	8308	627	2526	4941	3401	2599	1119
Number of person years	219699	16367	63464	122128	82836	61967	26310
All causes							
Number of deaths	1248	161	760	1697	1245	1084	499
Per 100000 person years	568	984	1198	1390	1503	1749	1897
Ischaemic heart disease							
Number of deaths	287	63	233	521	406	311	128
Per 100000 person years	131	385	367	427	490	502	487
All cancer							
Number of deaths	446	40	226	525	392	349	171
Per 100000 person years	203	244	356	430	473	563	650
Lung cancer							
Number of deaths	17	4	61	168	126	146	70
Per 100000 person years	8	24	96	138	152	236	266
Females							
Number of participants	11077	796	2759	3005	1008	477	79
Number of person years	288178	20537	70509	76158	25196	11720	2036
All causes							
Number of deaths	956	98	427	530	222	113	16
Per 100000 person years	332	477	606	696	881	964	786
Ischaemic heart disease							
Number of deaths	85	17	72	79	36	13	2
Per 100000 person years	29	83	102	104	143	111	98
All cancer							
Number of deaths	579	46	197	264	100	58	9
Per 100000 person years	201	224	279	347	397	495	442
Lung cancer							
Number of deaths	14	5	43	65	33	14	3
Per 100000 person years	5	24	61	85	131	119	147

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

Conclusion 2

Chez les hommes et les femmes, quelle que soit la cause du décès, les probabilités augmentent avec la consommation de cigarettes.

34

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Commentaires du tableau 2 (tels que donnés dans l'article)

Table 2 Number of participants and person years; deaths from all causes, ischaemic heart disease, all cancer, and lung cancer, number and per 100000 person years, by number of cigarettes recorded at screening. 23521 male and 19201 female participants aged 35–49*

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Number of participants	8308	627	2526	4941	3401	2599	1119
Number of person years	219699	16367	63464	122128	82836	61967	26310
All causes							
Number of deaths	1248	161	760	1697	1245	1084	499
Per 100000 person years	568	984	1198	1390	1503	1749	1897
Ischaemic heart disease							
Number of deaths	287	63	233	521	406	311	128
Per 100000 person years	131	385	367	427	490	502	487
All cancer							
Number of deaths	446	40	226	525	392	349	171
Per 100000 person years	203	244	356	430	473	563	650
Lung cancer							
Number of deaths	17	4	61	168	126	146	70
Per 100000 person years	8	24	96	138	152	236	266
Females							
Number of participants	11077	796	2759	3005	1008	477	79
Number of person years	288178	20537	70509	76158	25196	11720	2036
All causes							
Number of deaths	956	98	427	530	222	113	16
Per 100000 person years	332	477	606	696	881	964	786
Ischaemic heart disease							
Number of deaths	85	17	72	79	36	13	2
Per 100000 person years	29	83	102	104	143	111	98
All cancer							
Number of deaths	579	46	197	264	100	58	9
Per 100000 person years	201	224	279	347	397	495	442
Lung cancer							
Number of deaths	14	5	43	65	33	14	3
Per 100000 person years	5	24	61	85	131	119	147

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

Conclusion 3

Quelle que soit la consommation de cigarettes et la cause du décès, les femmes ont moins de risque de mourir que les hommes !

35

E/ Résultat 2 – Causes de décès dans les différents groupes de consommation

Commentaires du tableau 2 (tels que donnés dans l'article)

Table 2 Number of participants and person years; deaths from all causes, ischaemic heart disease, all cancer, and lung cancer, number and per 100000 person years, by number of cigarettes recorded at screening. 23521 male and 19201 female participants aged 35–49*

	Number of cigarettes smoked daily						
	0	1–4	5–9	10–14	15–19	20–24	25+
Males							
Number of participants	8308	627	2526	4941	3401	2599	1119
Number of person years	219699	16367	63464	122128	82836	61967	26310
All causes							
Number of deaths	1248	161	760	1697	1245	1084	499
Per 100000 person years	568	984	1198	1390	1503	1749	1897
Ischaemic heart disease							
Number of deaths	287	63	233	521	406	311	128
Per 100000 person years	131	385	367	427	490	502	487
All cancer							
Number of deaths	446	40	226	525	392	349	171
Per 100000 person years	203	244	356	430	473	563	650
Lung cancer							
Number of deaths	17	4	61	168	126	146	70
Per 100000 person years	8	24	96	138	152	236	266
Females							
Number of participants	11077	796	2759	3005	1008	477	79
Number of person years	288178	20537	70509	76158	25196	11720	2036
All causes							
Number of deaths	956	98	427	530	222	113	16
Per 100000 person years	332	477	606	696	881	964	786
Ischaemic heart disease							
Number of deaths	85	17	72	79	36	13	2
Per 100000 person years	29	83	102	104	143	111	98
All cancer							
Number of deaths	579	46	197	264	100	58	9
Per 100000 person years	201	224	279	347	397	495	442
Lung cancer							
Number of deaths	14	5	43	65	33	14	3
Per 100000 person years	5	24	61	85	131	119	147

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

Conclusion 4

Les femmes qui fument 25+ cigarettes meurent quand même un peu plus que les hommes qui n'ont jamais fumé… Mais même pas pour les maladies cardiaques !

36

F/ Résultat 3 – Augmentation du risque relatif à fumer *n* cigarettes/jour

Risque Relatif (RR) = nombre par lequel le risque de maladie est multiplié en présence de l'exposition.

Rapport entre le risque dans un groupe exposé et le risque dans un groupe non exposé

Exemple issu du tableau 2

Risque relatif de décès par maladie cardiaque chez les hommes fumant 25+ c/jour	
Probabilité chez les hommes fumeurs 25+	487 pour 100 000 personnes années
Probabilité chez les hommes non fumeurs	131 pour 100 000 personnes années
RR chez les fumeurs 25+	487/ 131 = **3.72**

Les hommes fumeurs de 25+ c/jour ont 3.72 plus de risque de mourir d'une maladie cardiaque

37

F/ Résultat 3 – Augmentation du risque relatif à fumer *n* cigarettes/jour

Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
Males							
All causes							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
Ischaemic heart disease							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
All cancer							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
Lung cancer							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
Females							
All causes							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
Ischaemic heart disease							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
All cancer							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
Lung cancer							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

†Adjusted for age.

‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

Question

Trouve-t-on la valeur de **3.72** pour le RR des hommes fumeurs 25+ dans la table de l'article ? Pourquoi ?

38

F/ Résultat 3 – Augmentation du risque relatif à fumer n cigarettes/jour

Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
Males							
All causes							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
Ischaemic heart disease							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
All cancer							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
Lung cancer							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
Females							
All causes							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
Ischaemic heart disease							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
All cancer							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
Lung cancer							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.
†Adjusted for age.
‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

Réponse

Les valeurs du risque relatif ne sont pas des valeurs brutes mais des valeurs ajustées.
Les auteurs proposent deux calculs :
- valeurs ajustées pour l'âge
- valeurs ajustées pour l'âge, la tension, le cholestérol, les triglycérides, l'activité physique pendant les loisirs, indice de masse corporelle, la taille.

But des ajustements

Eviter les *biais* ou les *facteurs de confusion* : variable qui peut modifier (augmenter, diminuer, créer, supprimer) la force de l'association entre deux variables

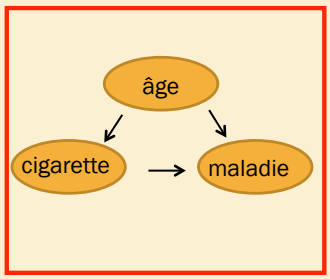
39

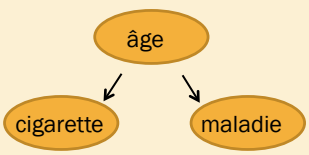
F/ Résultat 3 – Augmentation du risque relatif à fumer n cigarettes/jour

Exemple de biais

- Il se peut que les fumeurs de 25+ c/jour soient en moyenne plus vieux que les non fumeurs, pour des raisons culturelles et générationnelles.
- On sait que les maladies cardiaques surviennent davantage chez les personnes plus avancées en âge.
- Comment savoir si le RR chez les fumeurs de 25+ c/ jour vient de leur consommation de cigarette, de leur âge, des deux ?

L'âge pourrait-donc jouer comme un biais dans la relation cigarette-maladie
Quel est le bon schéma causal ?





40

F/ Résultat 3 – Augmentation du risque relatif à fumer n cigarettes/jour

Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
Males							
All causes							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
Ischaemic heart disease							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
All cancer							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
Lung cancer							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
Females							
All causes							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
Ischaemic heart disease							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
All cancer							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
Lung cancer							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

†Adjusted for age.

‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

Dernière info à connaître : **le niveau de significativité des résultats**

- **L'échantillon** : La cohorte des 23521 norvégiens et 19201 norvégiennes de 35 à 79 ans suivis de 1972/78 à 2002.
- **La population** = l'ensemble des norvégiens de 35 à 79 ans.

Le taux de significativité répond à la question :
Est-ce que le résultat obtenu pour l'échantillon peut être généralisé à la population ?
Quel risque prend-on à généraliser ce résultat ?

La population = la soupe	L'échantillon = la cuillerée prélevée pour savoir si la soupe est assez salée…
	
- La cuillerée représente-t-elle bien toute la soupe ? - Quel risque prend-on de tomber sur un échantillon qui est plus/moins salé que le reste de la soupe ?	

41

F/ Résultat 3 – Augmentation du risque relatif à fumer n cigarettes/jour

Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
Males							
All causes							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
Ischaemic heart disease							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
All cancer							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
Lung cancer							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
Females							
All causes							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
Ischaemic heart disease							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
All cancer							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
Lung cancer							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

†Adjusted for age.

‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

Dans toutes les sciences, on admet un résultat lorsque le seuil de significativité est inférieur à 0.05 (5%)

Les taux donnés dans la table 3 estiment si la relation entre **l'augmentation du risque (RR)** et **l'augmentation de la consommation de cigarette (1-4, 5-9, 10-14,…)** peut être généralisée.

➔ est-ce qu'il est VRAI que les fumeurs qui fument plus sont davantage exposés à des risques (maladies notamment) ?

Question
En général, la relation est significative (généralisable), sauf dans un seul cas. **Lequel ?**

42

F/ Résultat 3 – Augmentation du risque relatif à fumer n cigarettes/jour

Commentaires du tableau 3 (comme donnés dans l'article)

Table 3 - Adjusted relative risk of death from all causes, ischaemic heart disease, all cancer, and lung cancer, by number of cigarettes daily recorded at screening, with never smokers as reference.

	Number of cigarettes smoked daily						
	1–4	5–9	10–14	15–19	20–24	25+	Level of significance
Males							
All causes							
RR†	1.56	2.03	2.47	2.78	3.35	3.71	<0.001
RR‡	1.57	1.94	2.36	2.66	3.19	3.42	<0.001
Ischaemic heart disease							
RR†	2.65	2.67	3.24	3.89	4.10	4.07	<0.001
RR‡	2.74	2.47	3.09	3.70	3.75	3.60	<0.001
All cancer							
RR†	1.09	1.69	2.14	2.45	3.03	3.57	<0.001
RR‡	1.08	1.63	2.05	2.37	2.93	3.41	<0.001
Lung cancer							
RR†	2.84	11.94	17.98	20.77	33.48	38.64	<0.001
RR‡	2.79	11.30	16.73	19.36	31.69	36.22	<0.001
Females							
All causes							
RR†	1.44	1.90	2.31	3.01	3.29	2.67	<0.001
RR‡	1.47	1.90	2.29	2.97	3.14	2.61	<0.001
Ischaemic heart disease							
RR†	2.81	3.69	4.03	5.80	4.51	3.89	0.064
RR‡	2.94	3.55	3.78	5.28	4.25	3.53	0.114
All cancer							
RR†	1.11	1.44	1.87	2.18	2.73	2.44	<0.001
RR‡	1.14	1.44	1.85	2.22	2.47	2.43	<0.001
Lung cancer							
RR†	5.02	13.06	19.19	30.37	27.68	34.02	<0.001
RR‡	5.03	11.85	17.62	28.83	23.85	31.95	<0.001

*Participants not reporting cardiovascular disease, diabetes, or treatment for hypertension, nor symptoms of angina pectoris and atherosclerosis obliterans.

†Adjusted for age.

‡Adjusted for age, systolic blood pressure, total serum cholesterol, serum triglycerides, physical activity during leisure, body mass index, and height.

1 - Pas de différences importantes entre les deux façons d'ajuster les RR.

2 - En moyenne, le RR des fumeurs « légers » (1-4 c./ j), par rapport aux non fumeurs est :
- multiplié par 1.5 pour toutes les causes de décès confondues
- multiplié par un peu moins de 3 pour les problèmes cardiaques
- multiplié par 5 pour les femmes en ce qui concerne le cancer des poumons
- pas significativement plus élevé pour les hommes en ce qui concerne le cancer du poumon.

Conclusion générale de l'article
"The results from this and other studies imply that smoking control policymakers and health educators should emphasise more strongly that light smokers are also endangering their health."

1 . UTILISATION DES DONNÉES CHIFFRÉES DANS LA SOCIÉTÉ

1.1 La « guerre du tabac »

1.1.1 La lutte contre le tabac est devenue une guerre de chiffres

1.1.2 Une distinction importante : association, cause, facteur

1.1.3 Exemple d'enquête : fumer 1 à 4 cigarettes par jour, conséquences pour la santé ?

1.1.4 Simplifier pour mieux convaincre

Toutes les études s'accordent sur la nocivité du tabac, qu'il s'agisse de fumer beaucoup, peu, des cigarettes industrielles ou des cigarettes roulées, de fumer soi-même ou d'être exposé à la fumée des autres.

Ce qui n'empêche pas de remarquer que les données scientifiques sont simplifiées par les politiques publiques pour mieux convaincre.

Trois façons de simplifier :

1/ Les politiques publiques mettent en avant le rôle du tabac et parlent peu d'autres facteurs des maladies liées au tabac, notamment au cancer des poumons
- mise en évidence d'un facteur génétique : les fumeurs ne sont pas tous égaux devant le risque de cancer bronchique
- rôle de la pollution atmosphérique
- autres facteurs…

44

2/ Les estimations (RR, nombre de décès liés au tabac) reposent toujours sur des comparaisons entre fumeurs et non fumeurs.

Postulat – L'excès de mortalité chez les fumeurs par rapport aux non fumeurs est la conséquence de leur tabagisme.

- De nombreux autres facteurs peuvent jouer comme facteurs de confusion.

Etude comparant 9000 fumeurs, ex-fumeurs, non fumeurs, fumeurs passifs sur 33 facteurs de « style de vie » connus pour leur effet néfaste sur la santé.
Thornton, A., Lee, P. & Fry J. (1994). Differences between smokers, ex-smokers, passive smokers and non-smokers. J of Clinical Epidemiology 47, pp.1143-1162.

- Sur 27 des 33 facteurs, les fumeurs vivent dans des conditions plus dangereuses: revenus plus bas, niveau d'éducation plus bas, travail plus dangereux, plus de consommation d'alcool, absence de souci d'être en forme, plus de dépressions et de maladies mentales, repas sautés, faible consommation fruits, légumes, céréales, forte consommation friture, café et thé (avec plus de sucre), extraversion, etc.

Les fumeurs meurent plus… de tout : cirrhose, ulcère, empoisonnement, suicide, mort violente, etc.

45

3/ Les valeurs des chiffres avancés par les politiques publiques sont souvent des valeurs « extrapolées »

"Second-hand tobacco smoke kills 600 000 people each year."

Organisation Mondiale de la Santé , REPORT ON THE GLOBAL TOBACCO EPIDEMIC, 2009

- Valeur extrapolée à partir des enquêtes disponibles sur le tabagisme passif dans les pays « développés »
- Mais les données sont défaillantes dans de nombreux pays (qui ont d'autres urgences).
- Est-ce valide d'extrapoler sur le plan scientifique, même si c'est légitime sur le plan de la santé publique ?

46

Partie 2.
Méthodes de recueil et de traitement des données

2 . MÉTHODES DE RECUEIL ET DE TRAITEMENT DES DONNEES
2.1 Panorama général des méthodes de recueil

A/ Pour traiter des données, il faut qu'elles aient été recueillies lors d'observations

On recueille des données avec un objectif :

1/ Explorer une question de recherche « large »

- En construisant leur lexique (=vocabulaire), quelles unités les enfants de 2-6 ans apprennent-ils en premier : verbes, noms, unités communicatives (*merci, au revoir*) ?

2/ En se basant sur des résultats déjà acquis, tester une hypothèse plus précise

Etant donné que l'acquisition des unités linguistiques par l'enfant dépend de leur fréquence dans les énoncés qui lui sont adressés, les noms et les verbes fréquemment produits par la mère devraient être acquis en premier.

Différence entre 1 et 2 ⇒

Question de recherche : aborder un domaine peu exploré
Hypothèse : confirmer/infirmer un aspect d'un domaine déjà exploré

48

➔ Le type de données recueillies dépend du domaine de recherche et de la question ou de l'hypothèse qui est posée au départ.

Une fois ces données recueillies, il faut les « traiter » :

1/ Les trier ou les coder pour séparer ce qui est intéressant pour la question ou l'hypothèse initiale.

2/ Les synthétiser
Soit forme de tableaux contenant des pourcentages, des moyennes, etc.
➔ analyse quantitative

Soit en les résumant sous forme de textes et de schémas
➔ analyse qualitative

Dans les deux cas, un seul but ⇒
- Faire apparaître des tendances générales
- Et éventuellement les phénomènes qui s'écartent de cette tendance.

49

B/ Les méthodes de recueil et de traitement des données sont innombrables

Les méthodes de recueil et de traitement des données sont très variées.
Chaque science à ses méthodes d'observation privilégiées:

- **La sociologie** : questionnaires, entretiens
- **L'épidémiologie** : suivi de cohortes, comparaison malades / sains
- **La linguistique** : l'étude des corpus
- **La physique et la psychologie expérimentale** : l'expérimentation.

Les méthodes pour traiter (statistiques) et représenter (graphiques, etc.) les données sont aussi très nombreuses : présentation exhaustive impossible

Un panorama organisé selon quatre critères (cf. document 1)

- Document organisé en 5 colonnes
- La première colonne liste les types de méthode
- Les colonnes 2 -5 listent les critères pour différencier les types de méthodes

50

1/ Le critère *Structuration* : le plus compliqué !

Question de recherche - Ecriture des formes verbales en /E/ (ais, ait, aient, é, és, ée, ées, er) à l'école primaire et au collège. Des progrès ? Quelles erreurs subsistent ? Pourquoi ?

a/ Situation non structurée de recueil des données : on récupère les écrits « naturels » des élèves (prise de note, rédactions, brouillons, etc.).

b/ Situation structurée de recueil : on met en place une épreuve standardisée (la même pou tous)

Carnet avec une phrase à trou par page	**Quentin .....son copain**	**Quentin .....le train**
L'expérimentateur dicte les phrases en entier et les sujets remplissent le pointillé	« Quentin **a aidé** son copain »	« Quentin **va manquer** le train »
12 phrases avec cible -é de type sujet + *avoir* + participe passé +COD « *Les filles ont dans**é** un tango* » 12 phrases avec cible -er de type sujet + *devoir/aller* + V-infinitif + COD « *Les garçons doivent gagn**er** le match* » contrôle nombre de syllabes et fréquence des verbes.		
6 sujets m.sg (QUENTIN), 6 sujet m.pl. (LES GARCONS), 6 sujets f.sg (CAPUCINE), 6 sujets f.pl.(LES FILLES)		
600 élèves du CE2 à la 4ème collège		

Brissaud, C. et Chevrot, J.-P. (à paraître, 2011). The late acquisition of a major difficulty of French inflectional orthography: the homophonic /E/ verbal endings, Writing System Research

51

Structuré / non structuré : avantages et inconvénients

	Avantages	Inconvénients
Recueil **structuré** des données	- On cible le recueil sur le point étudié - On contrôle mieux les variables de confusions (les biais) - Facile à traiter *après* le recueil	- On perd des informations « inattendues » - On n'est pas certain de pouvoir généraliser les résultats aux situations « naturelles » - Long à élaborer *avant* le recueil
Recueil **non structuré** des données	- On recueille le point étudié en contexte (information riche) - On peut généraliser les résultats car situation « naturelle » - Recueil rapide à mettre en œuvre	- On doit trier les informations pertinentes et les non pertinentes - Difficile de contrôler les variables de confusion (les biais) - Long à traiter *après* le recueil

	Observation Structurée **Mesure orientation de la tête + temps de regard** 	**Observation non structurée** **Enregistrement vidéo** 
Validité « écologique » (conditions analogues au milieu naturel)	-	+
Validité « interne » (contrôle des biais)	+	-

52

2/ Le critère Taille

C'est la quantité de données qu'on recueille :
- Le nombre de mots dans un corpus
- Le nombre de sujets observés → les extrêmes :

Grande enquête	Plusieurs dizaines, centaines, milliers de sujets	- Suivi d'une cohorte d'enfants pour étudier les progrès en lecture en fonction des méthodes d'enseignement au CP
Etude de cas	Un seul sujet, suivi pendant X années ou enregistré pendant de nombreuses heures	- Enregistrement de trois enfants, 1 heure par semaine entre 2 et 6 ans, dans le milieu familial: étude de la relation entre énoncés adressés à l'enfant et progrès dans un domaine particulier (phonologie, lexique, morphosyntaxe, pragmatique)

	Grande enquête	Etude de cas
Richesse et précision des données	-	+
Généralisation des résultats	+	-

53

3/ Le critère Temps

- *synchronique* signifie qu'on photographie un phénomène à un moment donné ;
- *diachronique* signifie qu'on s'intéresse aussi à son évolution dans le temps : un processus d'apprentissage, une évolution historique.

Dans le domaine de l'acquisition du langage et du développement		
Étude longitudinale	Le même groupe de sujets ou le même sujet sont observés à différents âges	Ex : 30 sujets observés à 2 ans, puis 3 ans, puis 4 ans.
Etude transversale	Observation de différents groupes ou différents sujets d'âges différents	Ex : on observe 30 sujets de 2 ans, 30 autres sujets de 3 ans, 30 autres de 4 ans. Puis comparaison

	longitudinale	transversale
On peut voir l'évolution pour *chaque* sujet	oui	non
Les sujets sont les même entre les différents temps d'observation	oui	non
Risque de perdre des sujets en route (déménagement....)	Risque	Pas de risque
L'observation répétée modifie les comportements observés (les fumeurs suivis stoppent plus le tabac !)	Risque	Pas de risque

54

4/ Le critère quantitatif et qualitatif

Correspond au traitement des données :
- *Quantitatif* : l'analyse porte sur des données numériques (pourcentage, scores, moyennes, etc.)
- *Qualitatif* : l'analyse exclut les chiffres.

Compte-rendu écrit et schémas qui font ressortir les tendances générales, la cohérence d'un phénomène, différents types.

	quantitatif	qualitatif
Phase d'exploration d'un domaine inconnu	Reste possible	Plus souvent utilisé
Méthode d'observation structurée	oui	non
Méthode d'observation peu structurée	Reste possible	Plus souvent utilisé
Généralisation des résultats	oui	Difficile

Querelle des méthodes entre ceux qui pensent que les sciences humaines et sociales doivent
- utiliser les mêmes méthodes que les sciences « dures »
- créer leurs méthodes spécifiques.

55

mini-TD 2

Cf. les deux documents
« six méthodes types de recueil de données en SHS »
+
« Mini-TD2 sur les types de méthode »

56

2 . MÉTHODES DE RECUEIL ET DE TRAITEMENT DES DONNES
2.1 Panorama général des méthodes de recueil
2.2 Un exemple d'expérimentation : Tâche d'apostrophe de figurines

1/ Objectifs

Répondre à la question : à quel mot l'enfant attache-t-il les consonnes de liaison ?
Quelques travaux antérieurs suggèrent que l'enfant (< 5 ans) attache les CL au nom : *nours, zoreille, tâne, etc.*
→ Expérience destinée à amener un indice proche de la preuve

2/ Intérêt théorique (et applicatif)

La liaison « brouille » la frontière entre les mots	
Une liaison apparaît entre un mot1 et un mot2	**les** / ânes → les-z-ânes **un** / ami → un-n-ami
Elle n'apparaît pas si le mot2 commence par une consonne	**les** garçons **un** garçon
Chez l'adulte, elle n'apparait pas si le mot2 est au début d'un énoncé	Âne, gare à toi ! Ami !
Elle n'apparaît pas si le mot1 et à la fin d'un énoncé	Donne-**les** ! J'en veux **un**.

La liaison fait apparaître les procédés que l'enfant utilise pour segmenter des mots nouveaux dans la parole continue qu'il entend.

57

3/ Logique de l'expérience

Si la liaison est au début des mot2 chez les enfants, ils devraient la produire lorsqu'un nom (commençant par une voyelle chez l'adulte) est au début d'un énoncé

→ tâche d'apostrophe de figurine d'animaux : appeler des animaux en plastique pour les faire avancer : *Ours, viens ici !* , *Ane, viens ici !* , *Ecureuil, viens ici !* , etc.



On attend des réponse du type :

Nours viens ici !
OU zours viens ici !

Pour en savoir plus :

Chevrot, J.-P., Dugua, C. & Fayol, M. (2009). Liaison acquisition, word segmentation and construction in French: A usage based account, *Journal of Child Language*, 36, 557-596.
Dugua, C. (2006). Liaison, segmentation lexicale et schémas syntaxiques entre 2 et 6 ans. Un modèle développemental basé sur l'usage. Thèse de doctorat. Grenoble: Université Stendhal Grenoble 3.

58

4/ Tâche et matériel

- 4 figurines correspondant à des noms commençant par une voyelle chez l'adulte : *âne, écureuil, éléphant, ours*.
- 3 figurines correspondant à des noms commençant par une consonne chez l'adulte : *cochon, chien, perroquet*

Distracteur : éléments introduit dans une tâche afin d'éviter que les sujets centrent (trop) leur conscience sur le point à l'étude

- Chaque enfant appelle deux fois chaque figurine.

5/ Echantillon de sujets : 200 enfants de 2 ans 4 mois (2;4) à 6 ans 1 mois (6;1)

Nous avons formé
4 groupes d'âge disjoints

Age Group	Number of children	Age range	Mean age
Age 2-3	49 children	2;4-3;1	2;9
Age 3-4	50 children	3;2-4;1	3;6
Age 4-5	52 children	4;2-5;1	4;7
Age 5-6	49 children	5;2-6;1	5;7

59

6/ Eventail des réponses et calcul des scores individuels

Pour chaque enfant, on travaille uniquement sur le 8 essais correspondant aux noms commençant par une voyelle. Des 8 essais, on enlève :
- les non réponses : l'enfant se tait ou parle d'autre chose
- les réponses où l'enfant se trompe de nom : *Cheval, viens ici !*
- les réponses où l'enfant supprime la voyelle initiale : *Cureil, viens ici !*

Une fois ce tri effectué, il reste :
a- Les productions avec liaison initiale : ***n**ours, **z**âne, **t**âne*
b- Les productions avec voyelle initiale conformes à ce que ferait un adulte : *ours !*
c- Les productions où l'enfant fait précéder le nom d'un déterminant : *l'ours, viens ici ! Un ours, viens ici !*

60

Pour chaque enfant, on calcule les pourcentages que représente chacune des réponses a, b et c parmi toutes ses réponses

Non réponse (silence)	**1**
Erreur (chat au lieu de ours)	1
a - Liaison initiale (**n**ours)	2
b - Conforme à l'adulte (ours)	3
c - Avec déterminant (**l'**ours)	1

Réponses de Matthieu B., 3;6

Calculez ce que représentent en pourcentages les productions a, b et c dans ses réponses.

-

CORRECTION

a - Liaison initiale (**n**ours)	2	(2/6) x 100	= **33,33 %**
b - Conforme à l'adulte (ours)	3	(3/6) x 100	= **50 %**
c - Avec déterminant (**l'**ours)	1	(1/6) x 100	= **16,66 %**

61

7/ Résultat 1: Evolution avec l'âge des pourcentages moyens pour chaque type de réponses

Age	Avec liaison initiale (nours !)	A voyelle initiale, conforme à l'adulte (ours!)	Avec déterminant (un ours, l'ours)
2-3 ans	35.4%	40.6%	24%
	p=0.01	p=0.40	
3-4 ans	19%	47.5%	33.4%
	p=0.03	p<0.0001	
4-5 ans	8%	83.3%	8.7%
	p=0.02	p=0.6694	
5-6 ans	2.8%	81.5%	15.8%

Taux de significativité

p=0.01 Probabilité de généraliser à tort la différence de moyennes entre deux âges

Comment interpréter les évolutions ?

62

8/ Conclusions sur l'expériences (telles qu'énoncées dans l'article de J. of Child Language)

Description des tendances

➢ Jusqu'à 4 ans, la grande majorité des enfants produisent des variantes (= erreurs) de type *nours* ou *zours*. Bien que leur fréquence diminue de façon continue entre 2 et 6 ans, près de la moitié des enfants de 5-6 ans continuent à en produire un petit nombre.

➢ Les productions de type *ours (adulte)* sont observées chez un peu moins de la moitié des sujets à 2-3 ans, âge auquel elles ne sont pas plus fréquentes que les variantes de type *nours*. Leur fréquence augmente subitement entre 3 et 4 ans et elles deviennent alors plus fréquentes que les variantes de type *nours*, tous les enfants les produisant à 5-6 ans.

Interprétation des tendances

➢ Ces résultats constituent un indice fort que les enfants segmentent les noms suivant une liaison en attachant la liaison à l'initiale de ce nom.

➢ Ces chercheurs citent un autre article où les auteurs disent avoir recueilli 41 productions similaires en situation naturelle : une fillette de 2;10 voit un âne dans un pré et elle crie : *nâne !*

63

A retenir…

Dans le commentaire des résultats, bien distinguer

Description des tendances des valeurs

→ Décrire les tendances générales des chiffres : quelles différences de moyennes sont significatives ? Augmentation ou diminution ?
→ Indiscutable !

Interprétation des tendances des valeurs

→ Donner du sens aux tendances générales : l'hypothèse est-elle vérifiée ? Existe-t-il une relation cause-conséquence entre les variables (âge et évolution des deux types de production)?
→ Sujet à la discussion, à la critique, à l'atténuation.

Importance des tests statistiques qui permettent de décider si un résultat (une différence de moyennes) est significatif ou pas.

64

2 . MÉTHODES DE RECUEIL ET DE TRAITEMENT DES DONNES
2.1 Panorama général des méthodes de recueil
2.2 Un exemple d'expérimentation : Tâche d'apostrophe de figurines
2.3 La logique des tests de significativité (ou tests d'inférence statistique)

A/ Le problème : présentation imagée et intuitive…

Soupe A

Soupe B

- **Question de recherche** : laquelle des deux soupes est la plus salée ?
- **Procédure** : sans mélanger la soupe auparavant, on prélève un échantillon de 8 cuillerées de soupe dans chacune des soupière. Pour chaque cuillère, on mesure la quantité de sel en gramme/litre
- Résultats en g./litre

	Soupe A	Soupe B
Cuillerée 1	2.2	3.1
Cuillerée 2	1.8	2.9
Cuillerée 3	2.4	3.08
Cuillerée 4	1.6	2.92
Cuillerée 5	2.9	3.2
Cuillerée 6	1.1	2.8
Cuillerée 7	1.0	3.0
Cuillerée 8	3.0	3.0
Moyenne	2.0 g/litre	3.0 g/litre

1. Quel risque prend-on en affirmant, à la vue des données et des moyennes, que la soupe B est plus salée que la soupe A (différence de moyennes de 1 g/ litre constatée dans les deux échantillons de 8 cuillerées) ?

2. Et si, par hasard, on était tombé sur huit cuillerées de la soupe B beaucoup plus salées que tout le reste de la soupe B ? (biais d'échantillonnage)

65

Pour être au clair…

Population ≠ Echantillon

Ensemble d'objets ou d'individus ayant des caractéristiques qui leurs sont propres

Les étudiants de l'Université Stendhal

Sous-ensemble d'une population
Un « bon » échantillon doit être représentatif de la population dont il est issu

Ech. 1 : étudiants de SdL
Ech.2 : étudiants de Lettres

1/ On prend **une** soupe (une population). La quantité moyenne de sel dans cette soupe est évidemment unique (2g/litre)

2/ On prélève au hasard un très grand nombre d'échantillons de 8 cuillerées (disons 1000 paires)

	Moyenne sel dans chaque échantillon (g/ litre)
Echantillon 1	2 .1
Echantillon 2	1.8
Echantillon 3	2
Echantillon 4	2
Echantillon 5	1.9
.....	.....
Echantillon 252	2.1
Echantillon 253	4.6
.....	.....
Echantillon 998	0.1
Echantillon 999	4.5
Echantillon 1000	0.5

67

A/ Le problème : solution imagée et intuitive

	1/ Moyenne sel échantillon (g/ litre)
Echantillon 1	2
Echantillon 2	1.8
Echantillon 3	2
Echantillon 4	2
Echantillon 5	1.9
.....	.....
Echantillon 250	2
Echantillon 251	2.3
Echantillon 252	2.9
.....	.....
Echantillon 599	3.5
.....	.....
Echantillon 998	3.9
Echantillon 997	2.1
Echantillon 998	5.5
Echantillon 999	4.6
Ecahntillon 1000	4

250 échantillons sur 1000 avec une moyenne ≤ 2 g/litre →

Conclusion

J'ai **25 %** ((250/1000)x100) de chances de dire vrai si j'affirme que toute ma soupière contient ≤ 2 g de sel /litre

OU

75 % de chances de dire faux si j'affirme que toute ma soupière contient ≤ 2 g de sel /litre

3 échantillons sur 1000 avec une moyenne ≥ 4 g/litre →

Conclusion

J'ai **0,3 %** ((3/1000)x100) de chances de dire vrai si j'affirme que toute ma soupière contient ≥4 g de sel /litre

OU

99,7 % de chances de dire faux si j'affirme que toute ma soupière contient ≥4 g de sel / litre

68

A/ Le problème : conclusion

Les tests d'inférence servent à calculer la significativité statistique d'un résultat obtenu sur un ou des échantillons.

La significativité d'un résultat, c'est :

➢Le risque de généraliser (inférer) à tort un résultat à partir d'un échantillon.

➢ Le risque de **ne pas** trouver le même résultat si on recommence l'observation sur un autre échantillon de la même population.

La significativité d'un résultat, ce n'est pas :

➢ La probabilité qu'un résultat soit vrai
➢ La preuve qu'une tendance est forte

➔ On exprime ce risque par une probabilité p, qui varie de 0 à 1 et peut être convertie en pourcentage. Le maximum admis est 0.05 (cette valeur est un usage, pas une décision scientifique)

69

Mini-TD 3 (à vous de compléter)

Proba.	%	Conclusion	Proba.	%	Conclusion	Proba.	%	Conclusion
p = 0.01	1 %	significatif	p=0.23	?	?	p=0.305	?	?
p=0.05	5 %	significatif	p=0.45	?	?	p=.0634	?	?
p=0.15	15 %	non significatif	p=0.0001	?	?	p=0.0507	?	?
P=0.107	10.7 %	non significatif	p=0.99	?	?	p <0 .03	?	?

CORRECTION

Proba.	%	Conclusion	Proba.	%	Conclusion	Proba.	%	Conclusion
p =0.01	1 %	significatif	p=0.23	23 %	Non sign.	p=0.305	30.5 %	Non sign.
p=0.05	5 %	significatif	p=0.45	45 %	Non sign.	p=0.0634	6.34 %	Non sign.
p=0.15	15 %	non significatif	p=0.0001	0.01 %	Sign.	p=0.0507	5.07 %	Non sign.
P=0.107	10.7 %	non significatif	p=0.99	99 %	Non sign.	p <0 .03	< 3 %	Sign.

B/ La notion de **distribution normale** : approche intuitive

Dans les faits, les statisticiens qui élaborent les tests de significativité (ou d'inférence) ne passent pas leur temps à faire des tirages aléatoires d'échantillons dans des populations fictives.

↳ Ils utilisent des **lois statistiques** qui estiment comment se distribuent des valeurs de mesure dans une population.

Les valeurs des mesures qui sont influencées par de nombreux facteurs suivent une distribution dite **normale** ou **« de Gauss »**.

➢ On observe ces distributions normales dans des domaines variés : physique, biologie, SHS : taille des souris, taille des éléphants, taux de cholestérol, revenus, QI, etc.

➢ L'idée générale qui sous-tend une distribution normale :

1/ la probabilité d'avoir une mesure proche de la moyenne est importante
2/ le nombre de mesures décroit au fur et à mesure qu'on s'éloigne de la moyenne

↳ Allure générale de courbe en cloche

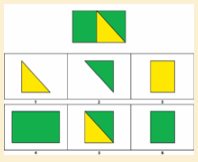
71

B/ La notion de distribution normale : l'exemple du Quotient Intellectuel

On mesure le QI par des tests comportant une ou plusieurs épreuves mettant en jeu la « logique »

Ex : Wechsler Adult Intelligence Scale :
mallette contenant plusieurs épreuves

Épreuves de raisonnement à partir perceptions visuelles :
- lesquelles de ces figures faut-il assembler pour obtenir celle du haut ?



Épreuves des similitudes :
capacité d'abstraction de concepts verbaux, de raisonnement et de compréhension verbale :
- en quoi table/chaise, se ressemblent ?
- manger/dormir ?
- sédentaire/ nomade ?
- etc.

1/ les sujets passent l'ensemble des épreuves

2/ obtention d'un score par épreuve

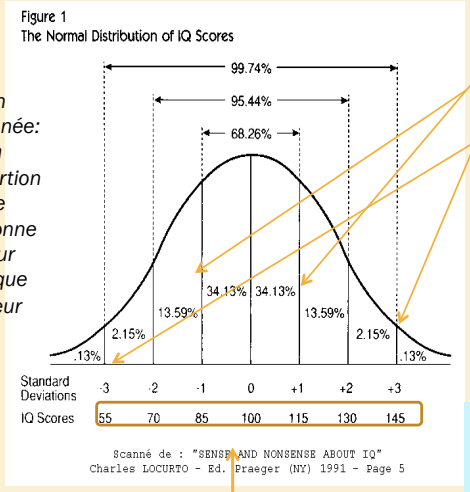
3/ addition des scores

Normalisation pour que le QI moyen soit de 100 pour chaque groupe d'âge (étalonnage)

72

→ **Représentation graphique de la distribution des QI : répartition commune à toutes les distributions normales**

En ordonnée: la proportion de personne pour chaque valeur



En abscisse : les valeurs du QI

- 68,26 % des gens ont un QI autour de la moyenne, entre 85 et 115 (34.13 + 34.13)
- 99.74 % des gens ont un QI autour de la moyenne, entre 55 et 145 (2.15 + 13.59 + 34.13 + 34.13 + 13.59 + 2.15)
- 2.28 % des gens ont un QI supérieur à 130 (2.15+0.13)
- 0.13 % des gens ont un QI supérieur à 145

Quelle proportion des sujets ont un QI supérieur à la moyenne ?

Nota bene

Exemple de bêtise trouvée sur le web http://www.test-de-qi.eu/test-de-qi.html

115-124 - Au dessus de la moyenne (étudiant à l'université)
125-134 - Doué (étudiant en grande école)
135-144 - Très doué
145-154 - Génie (professeur)
155-160 - Génie (prix Nobel)

73

2 . MÉTHODES DE RECUEIL ET DE TRAITEMENT DES DONNES
2.1 Panorama général des méthodes de recueil
2.2 Un exemple d'expérimentation : Tâche d'apostrophe de figurines
2.3 La logique des tests de significativité (ou tests d'inférence statistique)
2.4 Différents types de problèmes traités par les tests de significativité

EN BREF : les tests d'inférence statistiques permettent d'estimer le risque d'inférer un résultat d'un échantillon à une population et de décider si on « prend le risque » (si ≤ 0.05 ou 5 %)

Un résultat ? Mais de quel type ?

Une différence de moyennes

- Le revenu moyen est-il plus élevé dans la tranche 30-35 ans que dans la tranche 25-29 ans ?

Une association entre des classifications

- La répartition fumeurs/non fumeurs est-elle identique chez les hommes et les femmes ?

Une association entre des ordres

- La consommation de cigarettes (nbr de cig./jour) est-elle reliée au revenu (en euros /an) ?

La possibilité de poser une des ces trois questions dépend du **type de variable** : on ne peut pas calculer une moyenne sur des variables « étiquettes » comme le genre (homme/femme)

74

Les différents types de variables

	Variables catégorielles ou nominales	Variables quantitatives	Variables ordinales
Définition	Collent une étiquette, un nom aux objets	Attribuent une valeur numérique aux objets	Attribuent un ordre mais pas de valeur numérique
Exemples	-Le genre : **deux modalités** : masculin et féminin -Le lieu d'habitation : **trois modalités** : centre ville, campagne, périphérie urbaine -Le statut par rapport à la cigarette : **deux modalités** : fumeurs, non fumeur	-Le revenu mensuel net : en euros -La moyenne au bac -Le QI -La taille : en mètres ou centimètres	-Le niveau d'étude : **trois modalités** : pré-bac, de bac à L3, au delà de L3 - L'orientation politique : **quatre modalités** : sympathie pour NPA, PS, UMP, FN
Critères			
L'ordre à un sens	Non	Oui	Oui
Il existe une origine (zéro) et une unité de mesure	Non	Oui	Non

75

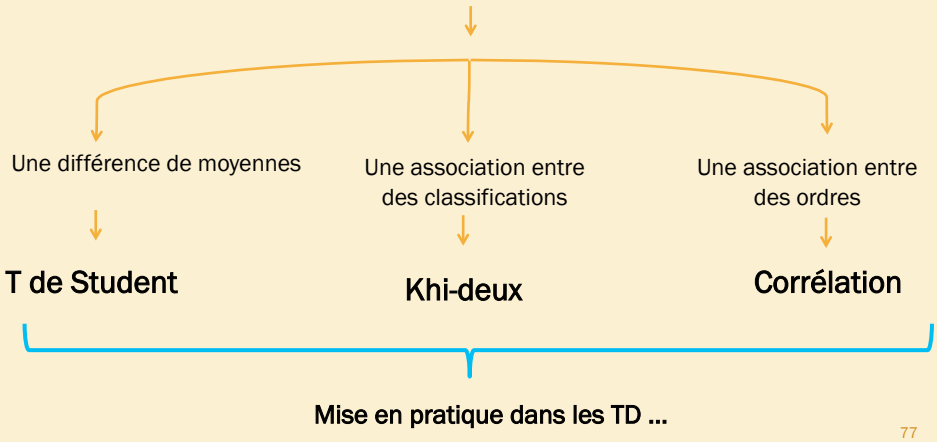
Types de problèmes traités par les tests de significativité et types de variables

	Différence de moyennes	Association entre des classifications	Association entre des ordres
Variables impliquées	- Une variable quantitative et plusieurs variables nominales	- Plusieurs variables nominales	- Plusieurs variables impliquant une relation d'ordre → donc ordinales ou quantitatives.
Exemples	- La moyenne au bac (*variable quantitative*) est-elle différente chez les filles et les garçons (*variable nominale*) ? - La moyenne au bac est-elle différente chez des enfants d'ouvriers, d'employés, de cadres ?	-La proportion de fumeurs et de non fumeurs est-elle différente chez les hommes et les femmes ? (fumeur/ non fumeur et homme/femme : deux variables nominales)	<u>quantitative / quantitative</u> -Le revenu mensuel (en euros) augmente-t-il avec le nombre d'années d'études ? <u>quantitative / ordinale</u> -Le revenu mensuel (en euros) est-il lié à l'orientation politique à droite (sur une échelle NPA, PS, UMP,FN) ? <u>qrdinale / ordinale</u> L'orientation politique à droite (échelle NPA, PS, UMP, FN) est-elle liée au niveau d'étude (pré-bac, bac à L3, supérieur à L3) ?

76

2 . MÉTHODES DE RECUEIL ET DE TRAITEMENT DES DONNES : EN BREF

Les tests d'inférence statistiques permettent d'estimer le risque d'inférer un résultat d'un échantillon à une population et de décider si on « prend le risque » (si ≤ 0.05 ou 5 %)



77