HEALTH SCIENCE ECOSYSTEMS
GRADUATE SCHOOL AND RESEARCH

Risques Epidémiologie Territoire INformations Education et Santé

ECUE 5
BIOSTATISTIQUE ET STATISTIQUES APPLIQUEES

Présentation de l'enseignement

G Maignant, L Lupi, C Pradier, P Staccini (responsable ECUE)

UNIVERSITÉ CÔTE D'AZUR

Sources et crédits

-1- Méthodes statistiques à l'usage des médecins et des biologistes
D. Schwartz.
Flammarion, Médecine Sciences.

-2- Biostatistique.
Beuscart, R. (Dir.) (2009).
Paris : Omniscience.

-3- Probabilités et statistiques
Alain-Jacques Valleron
Masson

Préliminaires

Benjamin Disraeli, premier ministre britannique (1804 – 1881) :

« Il y a trois sortes de mensonges : les petits mensonges, les gros mensonges et les statistiques!"

Statistiques, enquêtes, sondages, moyennes, indices... sont diffusés dans les journaux écrits et télévisés.. Le vocabulaire des probabilités et des statistiques est couramment employé dans la vie quotidienne : **espérance de vie**, salaire **moyen**, **fréquence** des bus…

Mal comprise, ou mal utilisée, cette pratique peut conduire à des conclusions surprenantes, voire absurdes.

L'usage de la statistique devient abusif. Le grand public reste perplexe et pense "on fait dire ce que l'on veut aux chiffres".

Préliminaires

Il y a 1,5 milliards de personnes dans le monde qui ne connaissent pas ou très peu de problèmes d'hémorroïde, ou de diverticulite (inflammation des parois de l'intestin).
Une étude statistique a prouvé que ces gens, majoritairement, faisaient leurs besoins accroupis.. Et non pas assis sur une cuvette comme par exemple les Européens, lesquels présentent très fréquemment des problèmes d'hémorroïde, ou de diverticulite..

Conclusion « évidente » : la position assise « à l'européenne », favorise l'apparition de problèmes d'hémorroïde, ou de diverticulite.

En fait, **une attitude critique** doit nous amener à nous poser d'autres questions : au-delà des constatations précédentes tous les autres paramètres sont ils identiques par ailleurs ? Alimentation, bien sûr, hygiène, pollution, stress, …

Attitude critique, indispensable dans le monde d'aujourd'hui : Conclusions statistiques, info, réseaux sociaux, …

1 - Biostatistique

Application des théories statistiques au domaine du vivant

⟹ **Le domaine de la Santé Publique**

Décrire l'état de santé de populations

Evaluer des traitements, des techniques, des coûts

Mettre en place des observations épidémiologiques, en tirer des conclusions

⟹ **Analyse descriptive**

⟹ **Analyse déductive**

Objectifs pédagogiques :

- **Notion de variabilité**
- **Définitions essentielles**
- **Les types de variables**

1 - Biostatistique

Statistique = art de collecter, d'analyser et d'interpréter des « données ».

Données = résultat de l'observation d'un individu, par l'utilisation d'un instrument de mesure, ou par les sens de l'observateur (signes cliniques, biologiques,..)

Cette donnée n'est intéressante que si on peut l'observer/la comparer sur plusieurs individus. Elle ne sera pas strictement équivalente d'un individu à l'autre. On parle donc de **variable**. On dira que la **variable** prend une valeur pour un individu, une autre valeur pour un autre individu, etc ..

On observe une grande **variabilité** des données dans le domaine biologique (taille, poids, groupe sanguin, température corporelle,..)

VARIABILITE → due au hasard
→ physiologique (intra ou inter sujets)

Paramètre = grandeur apportant une information résumée (ou synthétisée) sur la variable étudiée.

Exemple : moyenne d'une série de valeurs (notion détaillée plus loin..)

Les 2 domaines de la Statistique

- **Statistique descriptive :** Description d'une situation à l'aide de **paramètres**.

- **Statistique déductive (explicative ou inductive) :** Conclusions à partir d'observations et de mesures : **hasard ou autre explication** ?

Exemple : ***2 traitements anti cancéreux donnent à 5 ans une survie de 42 % pour l'un et de 48 % pour l'autre. Hasard ou efficacité plus grande pour l'un des deux ?***

La statistique = méthode scientifique
Les statistiques = collections de données, dénombrements (au 17ème siècle, les premiers dénombrements sont demandés par l'état ==> mot "statistiques")

1 - Biostatistique

- **Série statistique** **Collection d'objets de même nature, avec des caractéristiques différentes d'un objet à l'autre (variables).**
  - **Variables quantitatives >> mesurables (instr de mesure)**
  - **Variables qualitatives >> non mesurables (binaires, nominales…)**
- **Population** **Série exhaustive de TOUS les individus étudiés, sur lesquels on veut appliquer (inférer) des décisions.**

Exemple : *Population de la France, des étudiants en médecine de Nice, des patients opérés d'une certaine pathologie entre 2 dates précises dans un service donné,…*

- **Echantillon** **S/ensemble fini et d'effectif limité, extrait de la population.**

Pourquoi échantillonner ?

- Etude sur l'échantillon et « pari » sur l'application des résultats à la population.

- Population inaccessible dans son entier pour des raisons d'organisation et de moyens limités

Comment échantillonner ?

- L'échantillon doit être ***représentatif de la population***.

→ **Tirage au sort (randomisation)**

- **Echantillon connu, population inconnue**

Quelques variables

1 - La Biostatistique

1) Age
2) Sexe
3) Consommation tabac (nb cig/jour : 0 - 9 ; 10 – 19 ; >19)
4) Nb otites dans la dernière année
5) Déficit auditif moyen
6) Douleur articulaire (absente, modérée, intense)

Classer ces informations dans les différents types de variables définies :

Qualitative binaire, qualitative nominale, qualitative ordinale,

Quantitative discrète, quantitative continue

Quelques variables

1 - Biostatistique

A) Qualitative binaire **B**) Qualitative nominale **C**) Qualitative ordinale
D) Quantitative discrète **E**) Quantitative continue

Age	**D** : âges = **quantitative discrète**
Sexe	**A** : masculin ou féminin = **qualitative binaire**
Conso tabac	**C** : 3 niveaux de consommation (faible 0 - 9, moyenne 10 - 19, forte >19) = **qualitative ordinale**
Nb otites dans la dernière année	**D** : nb entier = **quantitative discrète**
Déficit auditif moyen :	**E :** précision 1/10ème(**11,5 dB**) = **quantitative continue**
Douleur articulaire (absente, modérée, intense)	**C** : 3 niveaux = **qualitative ordinale**

Quelques variables

1 - Biostatistique

Note Biostat	Rang Classement
12,4	210
4,9	555
18,1	6
5,4	445
19,4	5
16	14

Nature de ces variables ?

Quantitative ? Qualitative?

Quantitative continue **Qualitative ordinale**

Quelques variables

1 - Biostatistique

variable qualitative ordinale ⟹ variable pseudo quantitative

Variation ordinale

Ex : variation de la douleur sur une échelle de numérique, taux de satisfaction de 1 à 5, rang de classement …

5 = Très bon, 4 = Bon, 3 = Moyen, 2 = Mauvais, 1 = Très mauvais

Ou bien …1 = Aucune douleur, 2 = Supportable, 3 = Moyen, 4 = Douloureux, 5 = Très douloureux

Peuvent être considérées comme variables « pseudos quantitatives » pour certains tests

Attention : les nombres affectés aux modalités qualitatives n'ont pas de signification et ne peuvent faire l'objet d'opérations arithmétiques (calcul d'une somme ou d'une moyenne).

Etudié plus loin dans le cours

Représentation des variables qualitatives

1 - Biostatistique

Exemple : Quel est le degré de satisfaction des mères accouchant dans une certaine maternité ?

a) **Echantillon :** T.A.S mères ayant accouché dans cette maternité sur une période donnée (effectif n=154)

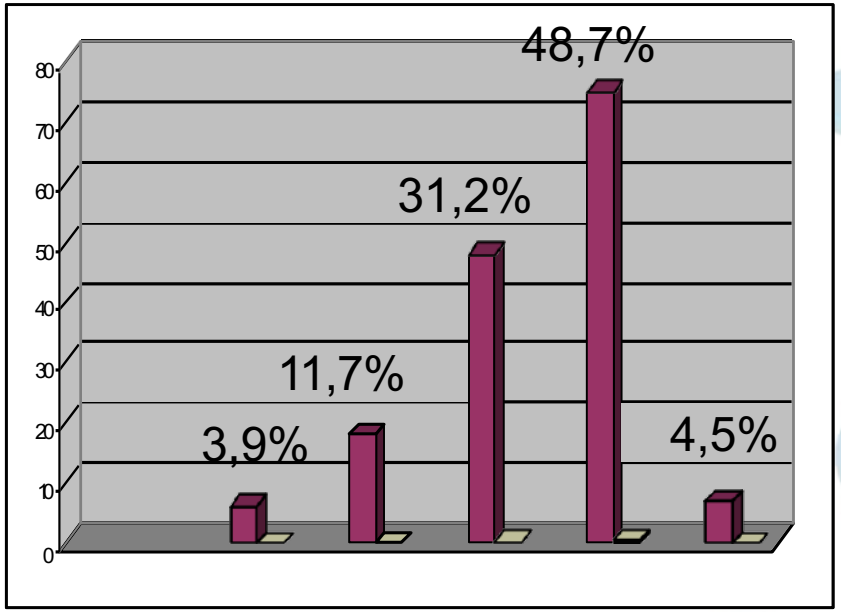
b) **Variable étudiée :** degré de satisfaction (très insatisfait, plutôt insatisfait, plutôt satisfait, très satisfait, pas d'opinion) : variables qualitatives ordinales

c) **Modalités de réponse = 5**

Représentation des variables qualitatives

1 - Biostatistique

Degré de satisfaction	Nb mères	%
Très insatisfait	6	3,9%
Plutôt insatisfait	18	11,7%
Plutôt satisfait	48	31,2%
Très satisfait	75	48,7%
Pas d'opinion	7	4,5%



Variables qualitatives : tableau de %
histogramme, secteurs …

Représentation des variables quantitatives

1 - Biostatistique

Exemple : on s'intéresse aux poids des nouveaux nés dans cette maternité

a) **Echantillon :** T.A.S mères ayant accouché dans cette maternité pendant une période donnée (effectif n=1165)

b) **Variable étudiée** : poids du nouveau né : variables quantitatives

c) **Données brutes** : n valeurs

Représentation des variables quantitatives

1 - Biostatistique

Poids (g)	Nb bébés
2200	2
2300	2
2400	4
2500	5
…	
3100	121
3200	150
3300	162
3400	170

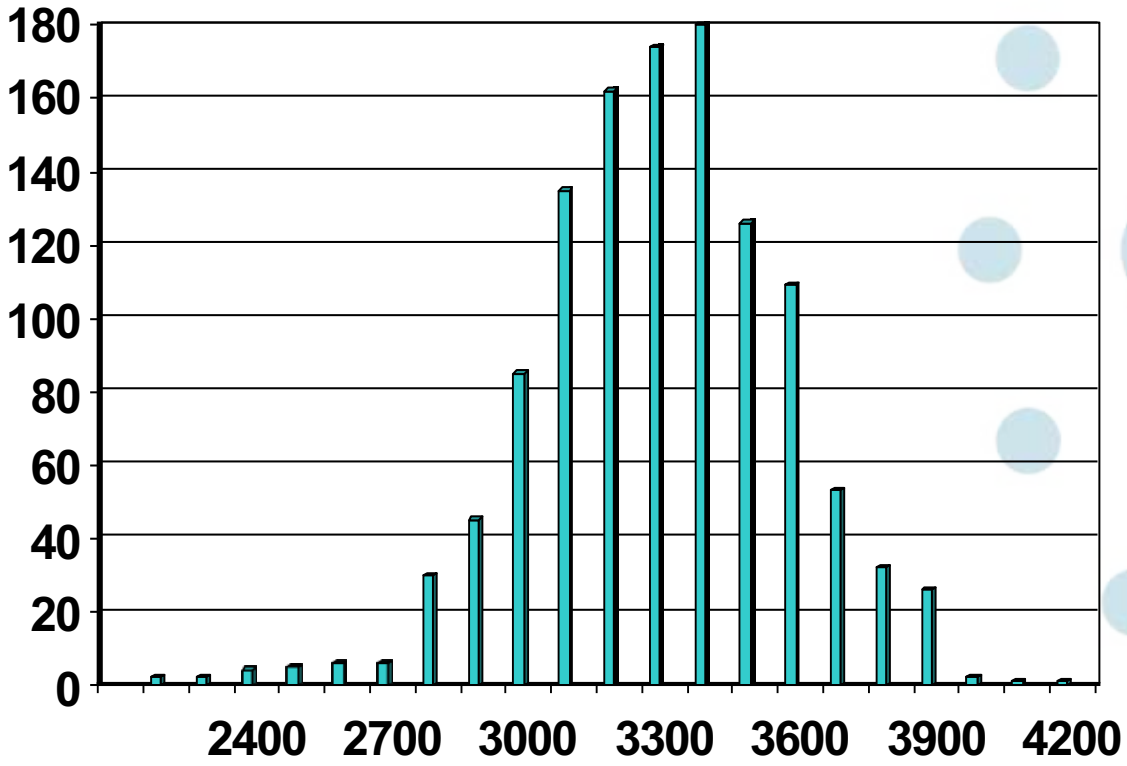


Tableau ou histogramme des effectifs
Mais pas seulement ……

Représentation des variables quantitatives

On peut « résumer » en quelques paramètres les caractéristiques de la série de données quantitatives.

Moyenne : cas d'une variable quantitative discrète $m = \frac{\sum_{1}^{n} x_i}{n}$

cas d'une variable quantitative continue $m = \frac{\sum_{1}^{n} n_i x_i}{n}$

Variance : paramètre indiquant la dispersion des données autour de la moyenne.

Médiane : valeur centrale si rangées par ordre croissant,

50% des valeurs < médiane et 50% des valeurs > médiane

(Exemple : {3,4,**6**,8,10}

Quartiles : Les quartiles partagent la série ordonnée en 4 groupes de même effectif (Exemple Q_{25} = 1er quartile : 25% de la série est < à cette valeur.)

Représentation des variables quantitatives

1 - Biostatistique

Exemple : Soit une série de poids de bébés (n = 15 valeurs)

1	3400
2	2570
3	3210
4	4070
5	3840
6	4180
7	3480
8	3990
9	2640
10	3000
11	3830
12	1890
13	2350
14	2140
15	2470

Valeurs rangées par ordre croissant

1	1890
2	2140
3	2350
4	2470
5	2570
6	2640
7	3000
8	**3210**
9	3400
10	3480
11	3830
12	3840
13	3990
14	4070
15	4180

Médiane : n=15 donc (n+1)/2 = 8

8ème valeur = 3210

Si n pair, médiane située entre les n° n/2 et n/2+1. **Moyenne des 2 valeurs correspondantes**

3ème quartile Q_{75}= 0,75 x 15 = 11,25

Le 3ème quartile est situé entre le n°11 et le n°12 soit (3830+3840)/2 = **3835**

Moyenne de la série = **3137,3**

Représentation des variables quantitatives

1 - Biostatistique

Comparaison des caractéristiques de la moyenne et de la médiane

Moyenne	Avantages	-Facile à calculer, se manipule facilement dans les tests statistiques (sera très utilisée) - Très significative si la répartition des données est assez symétrique et la dispersion faible
	Inconvénients	- Sensible aux valeurs anormales (mini ou maxi)
Médiane	Avantages	- Calcul facile, peu sensible aux valeurs anormales - Utilisable pour les valeurs ordinales, les classes, etc..
	Inconvénients	- Se prête moins aux calculs statistiques.

Ex : Salaire moyen = 2500€, salaire médian = 1800€….10% des salariés > 4000€
Durée moy grippe = 8j durée médiane = 6j
En fait moyenne et médiane sont complémentaires !

- Description de populations
- Estimation
- Notion de sondages

Objectifs pédagogiques :

- Variabilité
- Echantillonnage
- Intervalle de confiance
- Risque α

Exemple :
Glycémie d'une population de sujets normaux = 1g/L.
En **moyenne**, taux de sucre dans le sang chez les sujets normaux = 1g/L

A quoi sert cette information ?
Patient avec une glycémie de 1,2 g/L : glycémie normale ou anormale?

Quelle est la variabilité normale de la glycémie, liée aux variabilités individuelles ?

Variabilité non maitrisée ⇒ Biais

Variabilité maitrisée ⇒ Estimation

RÈGLES DE BASE

Variabilité en Biologie	……	Toujours
Individus	……	Tous différents
Groupes similaires	……	Résultats comparables mais non identiques.

Problème : déterminer un paramètre au niveau d'une population à partir d'observations réalisées sur un échantillon de cette population.

Exemples : - durée de séjour moyenne des patients hospitalisés en France, pour une pathologie donnée?

- durée moyenne d'attente aux urgences d'un hôpital ?

- **Estimation PONCTUELLE** — **Valeur, jugée la meilleure à un instant t (peu fiable)**
- **Estimation par INTERVALLE** — **Intervalle de valeurs contenant la valeur recherchée**

Intervalle de confiance autour de la valeur inconnue du paramètre

Méthodologie :

- **Détermination précise de la population étudiée = population cible**
- **Tirage au sort n sujets ➔ Echantillon représentatif**
- **Calcul de l'intervalle de confiance**

2.2 Estimation Statistique

a) Données quantitatives : Estimation de la moyenne

Echantillon
effectif = n
moyenne = m
écart type = s

ESTIMATION

Population cible
effectif = N
moyenne = μ
écart type = σ

$$m = \frac{\sum_{1}^{n} x_i}{n}$$

Estimateur de la moyenne vraie μ

$$s = \sqrt{\frac{\sum_{1}^{n} (x_i - m)^2}{n - 1}}$$

Estimateur de l'écart type vrai σ

= racine carrée de la variance

Données quantitatives : Estimation de la moyenne

NOTION D'INTERVALLE DE CONFIANCE

$$\mu \in [\, m \pm \frac{\varepsilon s}{\sqrt{n}} \,] \Rightarrow \text{Intervalle au risque } \alpha$$

α : Probabilité de se tromper dans l'estimation de μ

Plus α est petit, plus l'intervalle de confiance est grand : on réussi plus souvent. On s'expose aussi au risque de rater la "bonne" estimation. **Compromis universel** $\alpha = 5\%$

Intervalle de confiance à

$\alpha = 5\%$ $\varepsilon = 1{,}96$

$\alpha = 1\%$ $\varepsilon = 2{,}6$

Remarques

Différents échantillons ➔ Différentes estimations

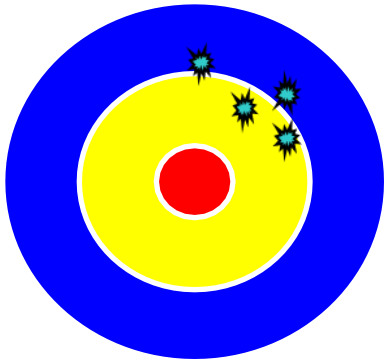
Taille échantillon augmente ➔ Estimation tend vers la moyenne vraie m.

C'est la méthode de calcul des normes des dosages biologiques

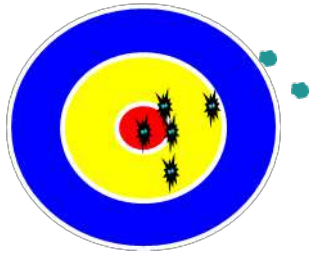
Précision de l'estimation

Intervalle de confiance peut être vu comme une cible

Large = plus de chances de l'atteindre, mauvaise précision de l'estimation



Resserré = risque de rater, meilleure précision de l'estimation



a) **Indice** qui permet de calculer **la précision** de l'estimation de μ: $i = \varepsilon \frac{s}{\sqrt{n}}$
Cette valeur est la largeur de l'intervalle de confiance.

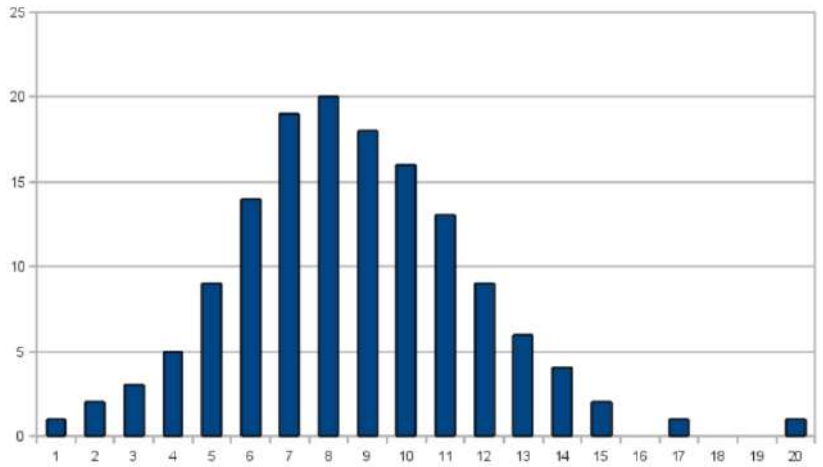
Indice petit = meilleure précision

b) Nombre de sujets nécessaires pour une précision donnée $n = \varepsilon^2 \frac{s^2}{i^2}$

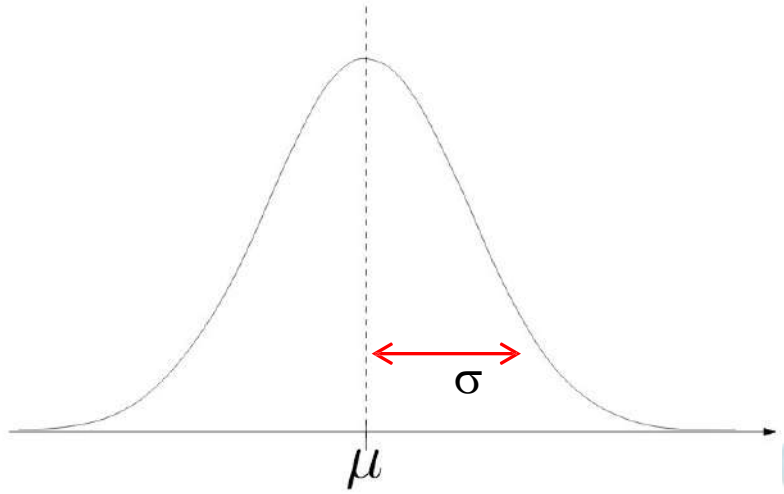
2.2 Estimation Statistique

En sciences humaines on observe souvent des distributions (X) plutôt symétriques autour de la moyenne avec une forme de cloche

Pour pouvoir faire des calculs, on va supposer que X suit une distribution « modèle », pour des **variables quantitatives continues** : **la Loi Normale**



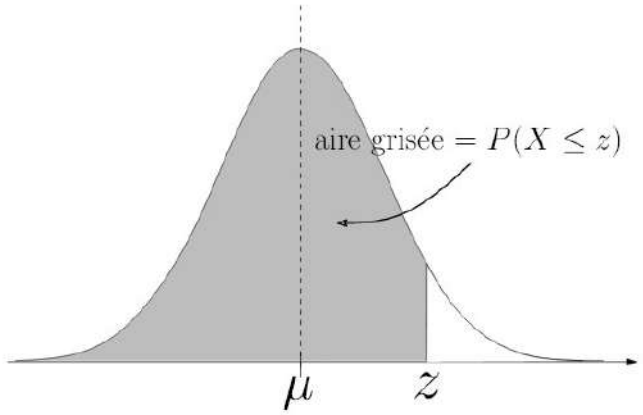
Par exemple : Répartition des notes à un examen



L'écart type caractérise la dispersion des données autour de la moyenne

A quoi çà sert ?

Pour chaque couple : moyenne, écart type (μ,σ)
il existe une loi normale de moyenne μ et d'écart type σ : $\mathcal{N}(\mu,\sigma)$



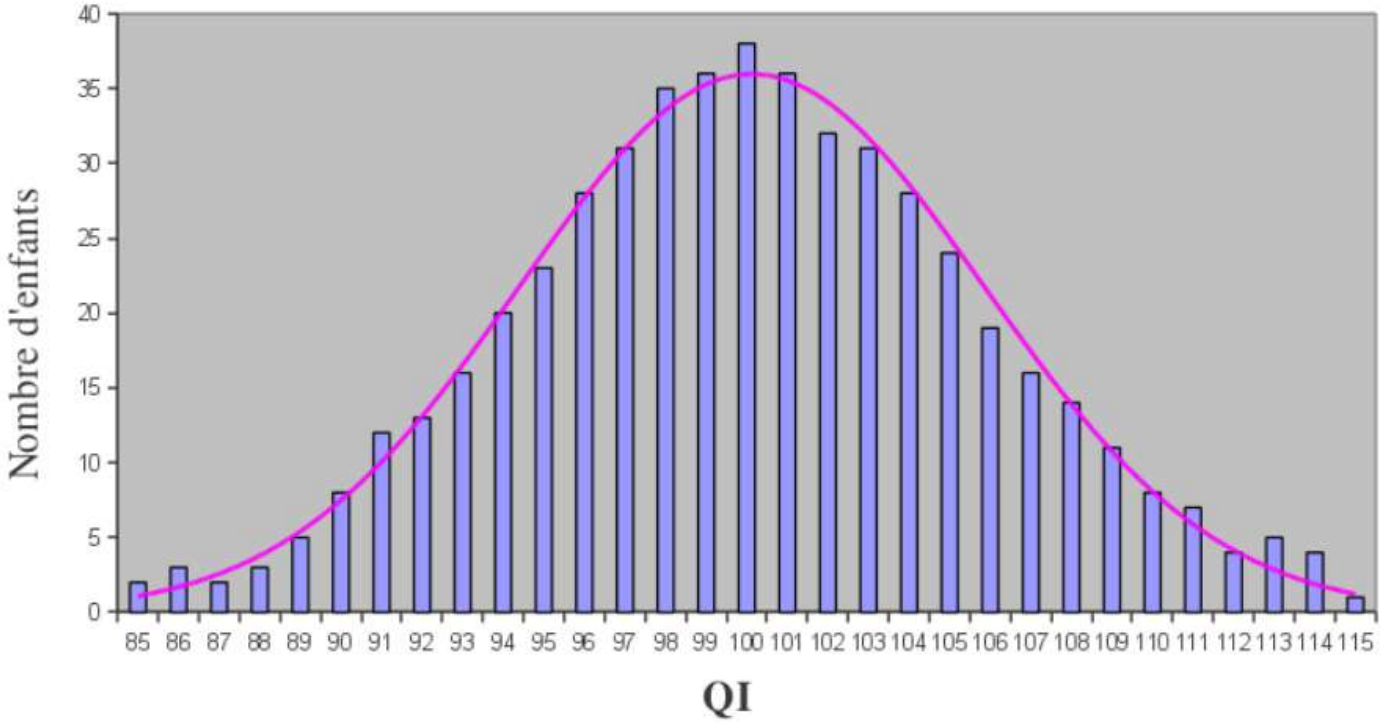
L'aire grisée représente une proportion cumulée : la probabilité que x $\leq$z donné.

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Cadeau pour les matheux !
Formule inutile pour ce cours !!

Echantillon effectif >30

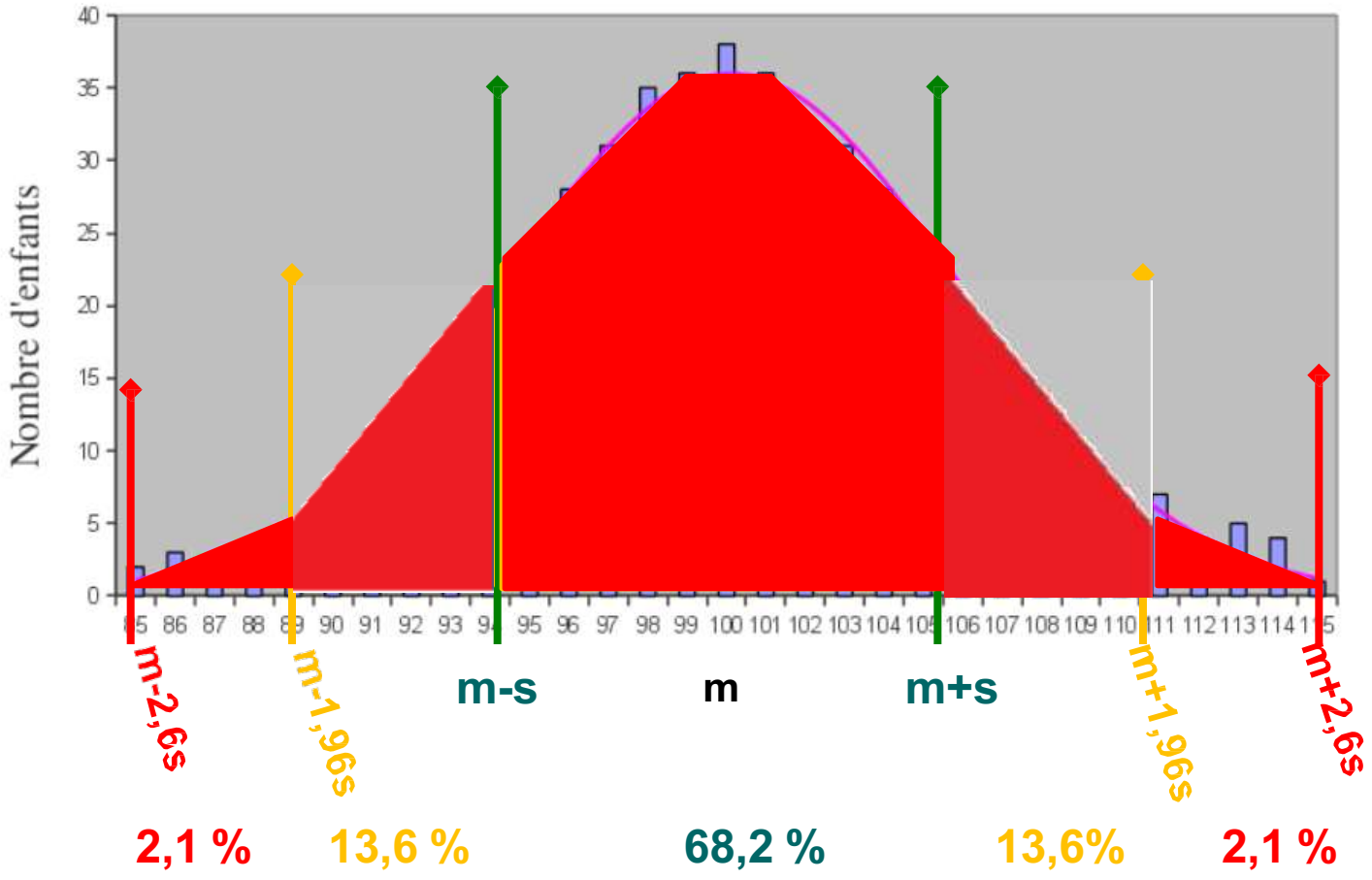
2.2 Estimation Statistique



Etude du QI de 515 enfants du même âge
Courbe rose = courbe de la loi normale $\mathcal{N}(\mu,\sigma)$
μ = 100, σ =5,7

2.2 Estimation Statistique

Loi Normale ou de GAUSS
Comment l'utiliser ?



2.2 Estimation Statistique

A partir de la Loi Normale ou de GAUSS, on précise :

Intervalles de confiance

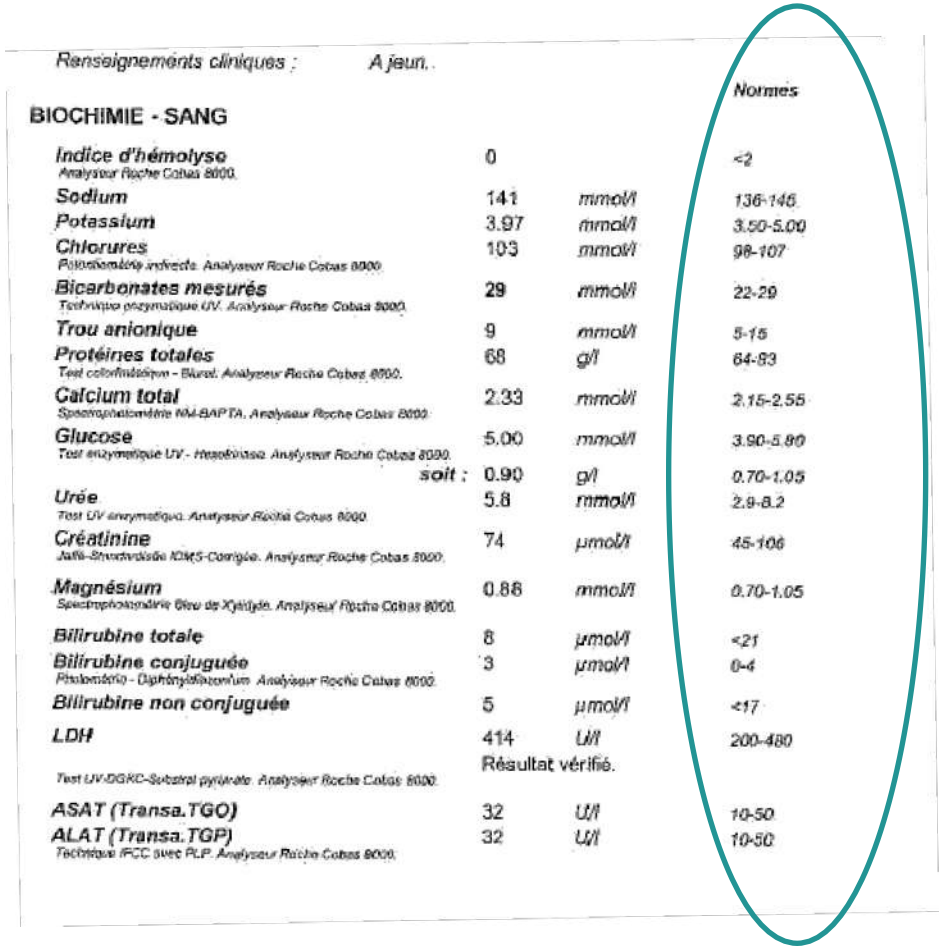
- [m - 1 s ; m + 1s] contient 68,2% de la population
- [m - 1,96 s ; m + 1,96s] contient 95,4% de la population
- [m - 2,6 s ; m + 2,6s] contient 99,6% de la population

Patient avec une glycémie de 1,2 g/l.
Quel est l'intervalle de confiance?

Intervalle de confiance au risque 5% = [0,90 ; 1,22]

On ne peut répondre au patient qu'en connaissant l'intervalle de confiance du dosage : NORMES

2.2 Estimation Statistique



Renseignements cliniques : *A jeun.*

BIOCHIMIE - SANG			Normes
Indice d'hémolyse *Analyseur Roche Cobas 8000.*	0		<2
Sodium	141	mmol/l	136-145
Potassium	3.97	mmol/l	3.50-5.00
Chlorures *Potentiométrie indirecte. Analyseur Roche Cobas 8000.*	103	mmol/l	98-107
Bicarbonates mesurés *Technique enzymatique UV. Analyseur Roche Cobas 8000.*	29	mmol/l	22-29
Trou anionique	9	mmol/l	5-15
Protéines totales *Test colorimétrique - Biuret. Analyseur Roche Cobas 8000.*	68	g/l	64-83
Calcium total *Spectrophotométrie NM-BAPTA. Analyseur Roche Cobas 8000.*	2.33	mmol/l	2.15-2.55
Glucose *Test enzymatique UV - Hexokinase. Analyseur Roche Cobas 8000.*	5.00	mmol/l	3.90-5.80
soit :	0.90	g/l	0.70-1.05
Urée *Test UV enzymatique. Analyseur Roche Cobas 8000.*	5.8	mmol/l	2.9-8.2
Créatinine *Jaffé-Standardisé IDMS-Corrigée. Analyseur Roche Cobas 8000.*	74	µmol/l	45-106
Magnésium *Spectrophotométrie Bleu de Xylidyle. Analyseur Roche Cobas 8000.*	0.88	mmol/l	0.70-1.05
Bilirubine totale	8	µmol/l	<21
Bilirubine conjuguée *Photométrie - Diphénylalbumine. Analyseur Roche Cobas 8000.*	3	µmol/l	0-4
Bilirubine non conjuguée	5	µmol/l	<17
LDH	414	U/l	200-480
Test UV-DGKC-Substrat pyruvate. Analyseur Roche Cobas 8000.	Résultat vérifié.		
ASAT (Transa.TGO)	32	U/l	10-50
ALAT (Transa.TGP) *Technique IFCC avec PLP. Analyseur Roche Cobas 8000.*	32	U/l	10-50

b) Données qualitatives : Fluctuation d'un pourcentage

Dans une population, quel % d'individus présentent un caractère donné ?

- Echantillon représentatif par tirage au sort (n sujets)
- Calcul d'un % qui tend vers la proportion cherchée, mais s'en écarte suivant une variabilité liée au hasard
- Autre échantillon ➔ autre %

p_{obs} ⟹ **Estimateur du pourcentage inconnu p**

$$s = \sqrt{\frac{p_0 q_0}{n}}$$ ⟹ **Estimateur de l'écart type inconnu σ**

avec $q_0 = 1-p_0$

INTERVALLE DE CONFIANCE

$$p \in [\, p_{obs} - \varepsilon s \; ; \; p_{obs} + \varepsilon s \,]$$

$\alpha = 5\%$ $\varepsilon = 1{,}96$ $\alpha = 1\%$ $\varepsilon = 2{,}6$

Exemple : **précision d'un sondage**

900 personnes ont été interrogées sur leur intention de vote à une élection présidentielle qui oppose 2 candidats A et B.

52% (p=0,52) ont déclaré qu'elles **voteraient A**.

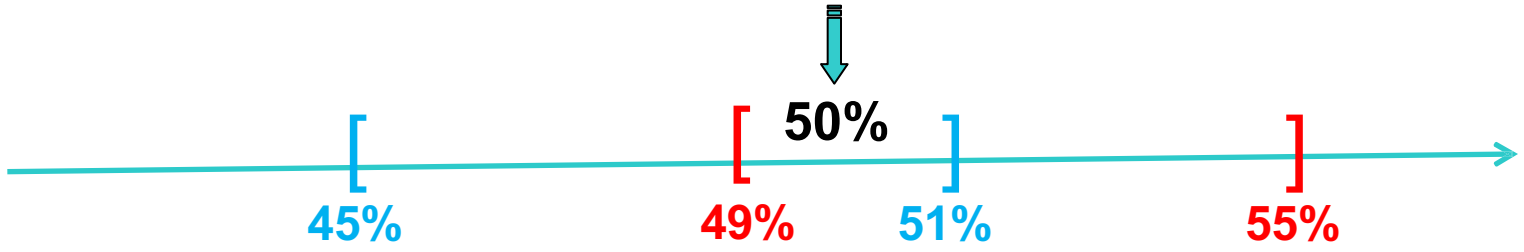
Les journaux annoncent que le candidat A arrive en tête.

Vérification statistique de cette affirmation

Pour A $IC_{0,95} = [0{,}52 \pm 1{,}96 \sqrt{\frac{0,52 \times 0,48}{900}}]$ = **[0,49 ; 0,55]**

Pour B $IC_{0,95} = [0{,}48 \pm 1{,}96 \sqrt{\frac{0,52 \times 0,48}{900}}]$ = **[0,45 ; 0,51]**

52% et 48 % possèdent des IC contenant 50%
Les 2 candidats peuvent être considérés comme à égalité !



Exemple de sondage

La popularité de XX et YY chute. Selon un sondage BVA diffusé ce vendredi, les deux hommes perdent 5 points. Avec respectivement 34% et 38% d'opinions positives, XX et YY sont à leur plus bas niveau dans ce baromètre depuis leur entrée en fonction.

A 34%, la cote de XX reste légèrement supérieure à celle de ZZ (32%) au même moment de son mandat, mais très en-dessous de celle de ZZZ (37%).

Commentaire absurde du point de vue statistique! Il est possible que les IC de ces pourcentages se recouvrent très largement, et que donc, finalement ces pourcentages soient considérés comme identiques!

2.2 Estimation Statistique

Comparaison estimation ponctuelle / estimation par intervalle

Soit un groupe de 220 patients, représentatif d'une population rhumatismale (R).
On observe 167 cas de rhumatismes inflammatoires.

Quel pourcentage de rhumatismes inflammatoires dans la population R?

1) Estimation ponctuelle p=167/220 = 0,76 soit **76%**

2) Estimation par intervalle

Nous choisissons le risque α **= 5%,** donc calcul de $IC_{0,95}$

p = 0,76 donc q = 0,24

$$IC_{0,95} = [0,76 \pm 1,96 \sqrt{\frac{0,76 \times 0,24}{220}}]$$

$$\mathbf{IC_{0,95} = [\ 0,70\ ;\ 0,82\ ]}$$

L'estimation par intervalle semble moins précise. Mais si l'on refait ce calcul sur un autre échantillon, cette nouvelle estimation recouvrira la première. Ce ne sera pas forcément vrai avec l'estimation ponctuelle.

2.2 Estimation Statistique

Précision

Soit P la population des ouvriers travaillant dans une usine

Nous voulons estimer le pourcentage **p** d'hommes dans cette population.

Considérons un échantillon TAS de 10 ouvriers : 7 hommes, soit $\mathbf{p_o=70\%}$

Estimation au niveau de P, au risque α=1% ?

$s= \sqrt{\frac{0,7 \times 0,3}{10}} = 0,144$ $\quad IC_{99\%} = [0,7 \pm 2,6 \times 0,144] = [\ 32,6\% ; 100\ \%]$

Considérons un échantillon de 1000 ouvriers : même % d'hommes $\mathbf{p_0 = 70\%}$

$s= \sqrt{\frac{0,7 \times 0,3}{1000}} = 0,014$ $\quad IC_{99\%} = [0,7 \pm 2,6 \times 0,014\ ] = [\ 66,4\% ; 73,6\%]$

Effectif n augmente ⟹ IC se resserre ⟹ Précision augmente

Un chirurgien écrit à 1000 de ses patients afin de connaitre leurs suites chirurgicales ➔ sur 100 réponses : 75 vont très bien, 25 ont des séquelles handicapantes.

Le chirurgien s'intéresse aux mauvaises suites chirurgicales, et veut estimer le % au niveau de l'ensemble de ses 1000 patients:

$$IC_{95\%} = [\ 0{,}25 \pm 1{,}96 \times \sqrt{\frac{0{,}25 \times 0{,}75}{100}}\ ]$$

Soit $IC_{95\%}$ = [25% ± 8%] de mauvais résultats soit **[17% ; 33%]**

Résultat : calcul statistique correct, mais non utilisable : il est faux !

1) Il y a eu 900 non-réponses. On ne peut pas préjuger de l'état de ces 900 patients. Ils sont peut être décédés des suites opératoires, ou bien très mécontents du chirurgien, ou tout au contraire sont très satisfaits et ne jugent pas utile de répondre .. Cet échantillon est BIAISÉ

2) Si on se trouve dans ce dernier cas, les échecs ne représentent que 25/1000 = 2,5%. La conclusion est toute autre !!

a) Données quantitatives

$$m = \frac{\sum_{1}^{n} x_i}{n}$$

m = moyenne calculée sur l'échantillon

$$s = \sqrt{\frac{\sum_{1}^{n} (x_i - m)^2}{n - 1}}$$

s = écart type calculé sur l'échantillon

Estimation de la moyenne inconnue dans la population cible

$$\mu \in [\, m \pm \varepsilon \frac{s}{\sqrt{n}} \,]$$

ε lu dans la table ⇒ risque d'erreur accepté

b) Données qualitatives

% au niveau de l'échantillon = p_o et $s = \sqrt{\frac{p_0 q_0}{n}}$ (avec $q_0 = 1-p_0$)

Estimation du % inconnu dans la population cible

$$p \in [p_0 \pm \varepsilon\, s]$$

ε lu dans la table ⇒ risque d'erreur accepté

- **Observations, mesures**
- **Conclusions**
- **Les tests**

Objectifs pédagogiques :
- **Notion d'hypothèses**
- **Risque de première espèce**
- **Choix du bon test**
- **Interprétation statistique et médicale**

Tirer des conclusions à partir d'observations

Exemple :
Comparer 2 groupes pour un caractère donné.

2 hypothèses :

- **H0 = Hypothèse nulle. Pas de différence observée entre les 2 groupes.**
- **H1 = Hypothèse alternative. Différence significative entre les 2 groupes.**

LES TESTS

Techniques permettant de décider si on garde ou repousse H0, en ayant fixé le risque d'erreur accompagnant cette décision.

LES ÉTAPES DE MISE EN ŒUVRE D'UN TEST D'HYPOTHÈSE

Question simple à propos d'un problème médical.

- **Etape 1** : Avant recueil des données définir H0 et H1. Les 2 hypothèses jouent des rôles **symétriques**

- **Etape 2** : Avant recueil des données **définir le test en fonction du type des données (qualitatives, quantitatives)**. Soit **Z** le paramètre qui sera calculé

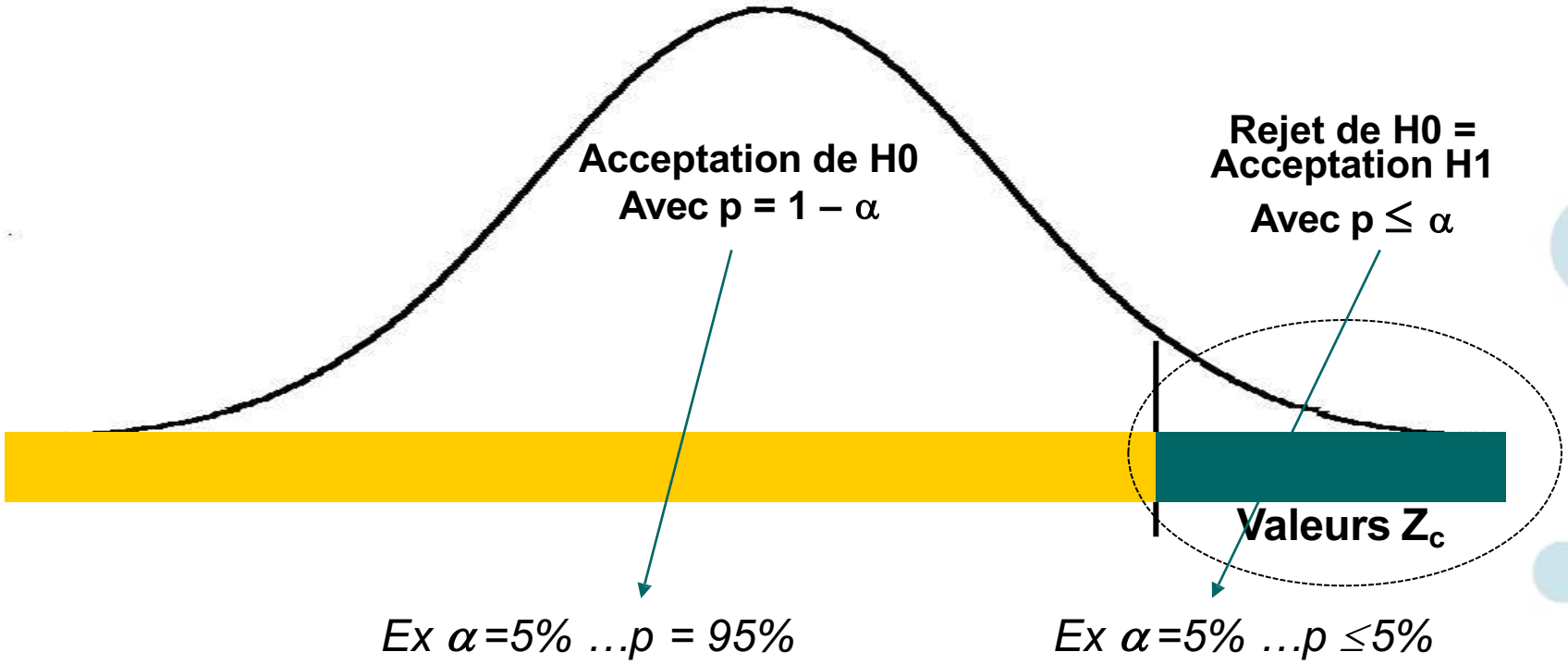
- **Etape 3** : Avant recueil des données on choisi **le risque** α (dans la pratique souvent 5%)

- **Etape 4** : Recueil des données.
  Calcul de Z.
  Règle de décision : examiner la position de cette valeur Z, par rapport à un modèle théorique dont on connait la distribution. Fixation du risque d'erreur attaché à la conclusion..

- **Etape 5** : Interprétation des résultats.

Interprétation graphique du risque α

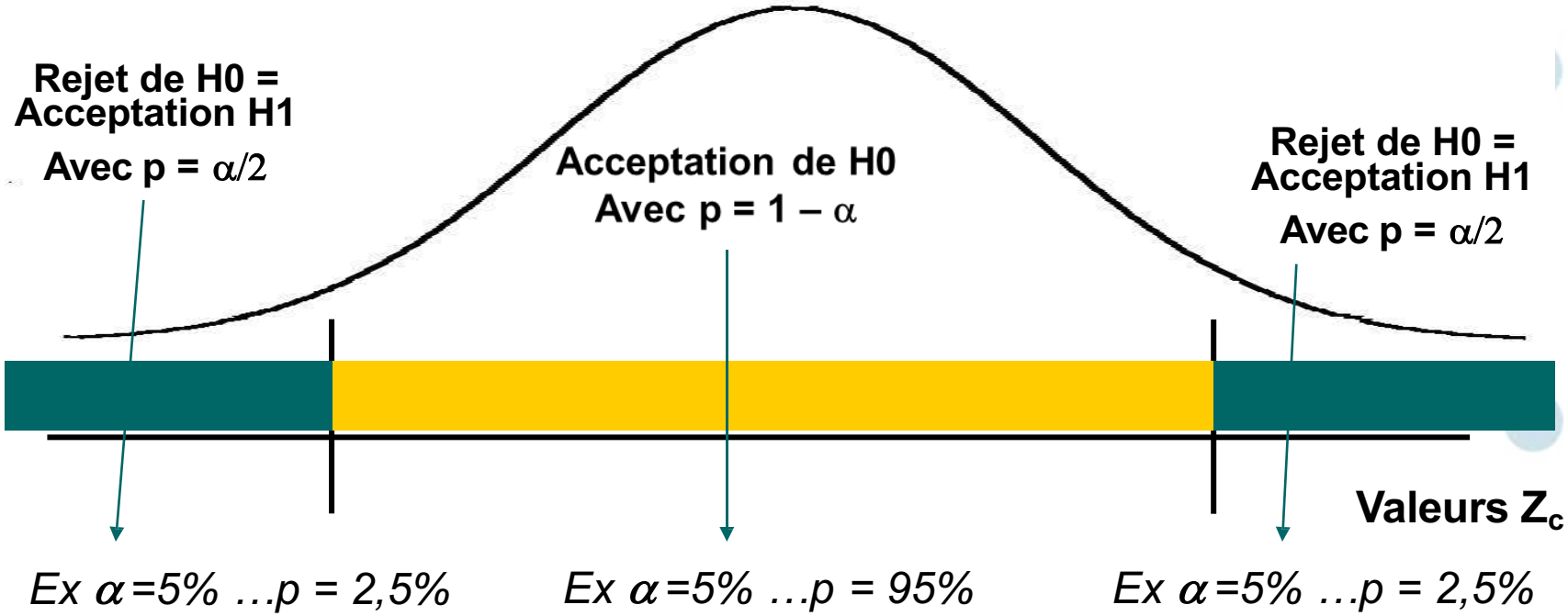
Le paramètre Zc résultat du test suit une distribution probabiliste en forme de **courbe de Gauss**. **Soit α, risque de 1ère espèce choisi = 5 %**



Situation unilatérale : **Les 2 situations observées sont elles différentes ?**

La seule réponse possible est OUI ou NON

Interprétation graphique du risque α



Ex α=5% …p = 2,5% *Ex α=5% …p = 95%* *Ex α=5% …p = 2,5%*

Situation bilatérale : **Les 2 situations observées sont elles différentes?**

Si OUI, il est possible d'affirmer laquelle est la meilleure des 2.

Interprétation graphique du risque α

Situation unilatérale :

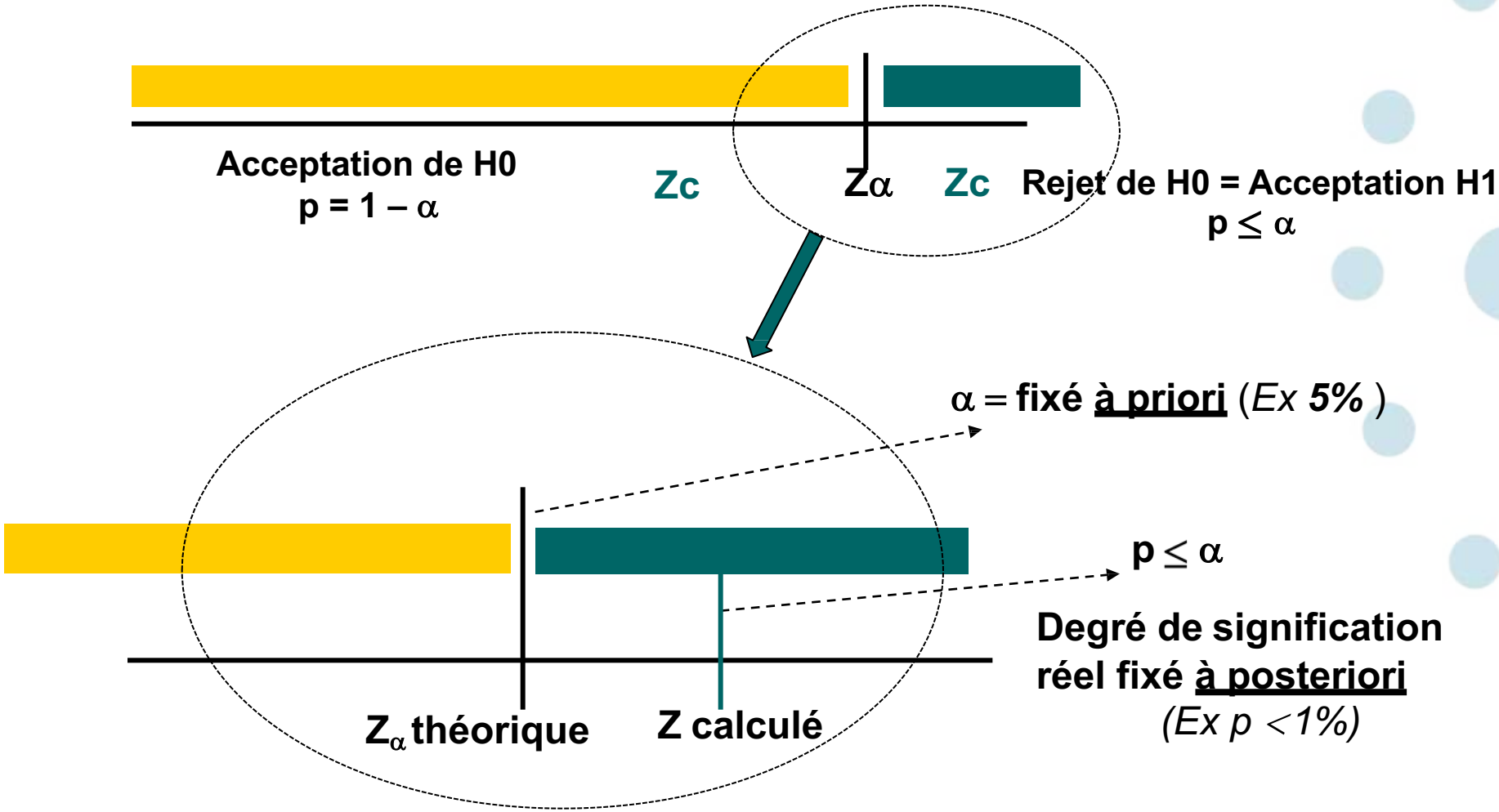


Table de l'écart réduit

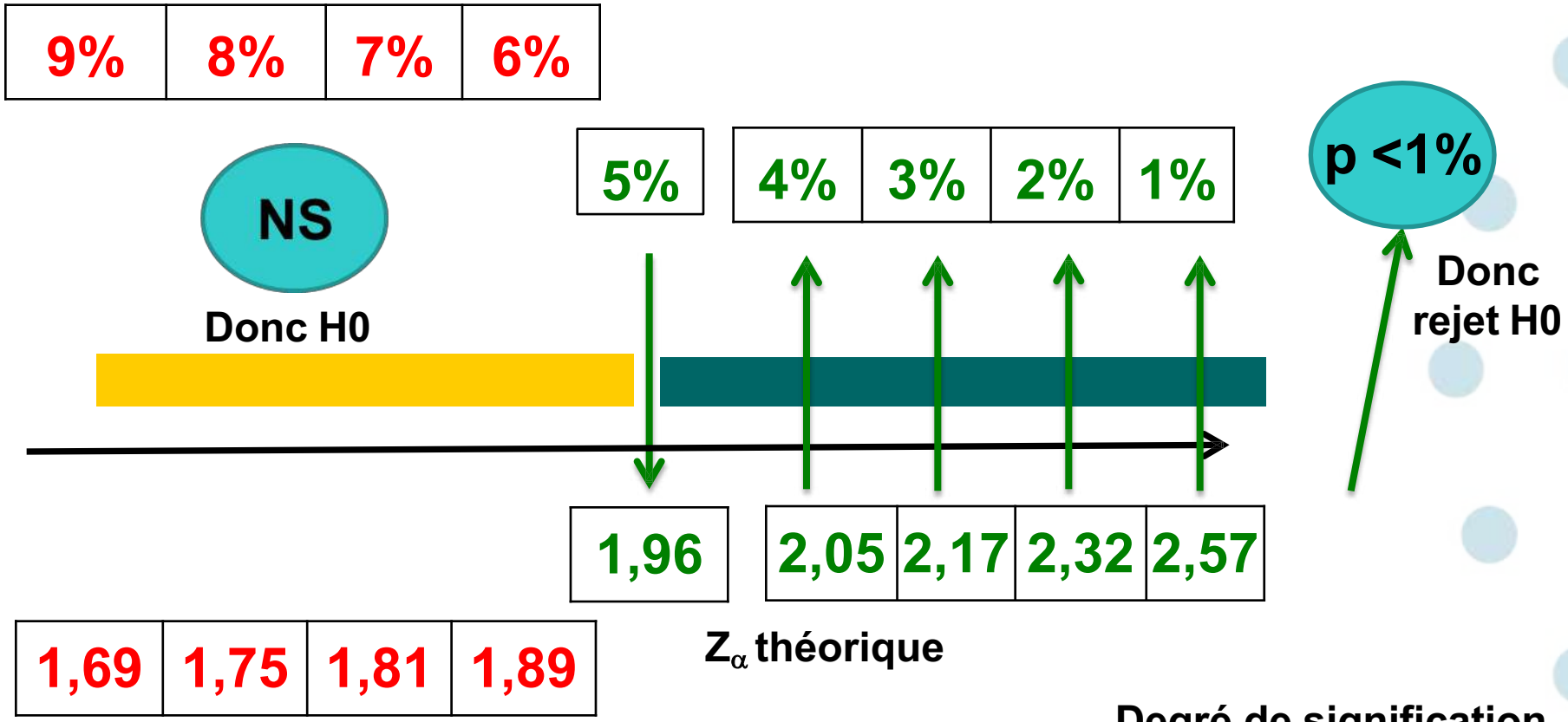
α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	**1,96**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

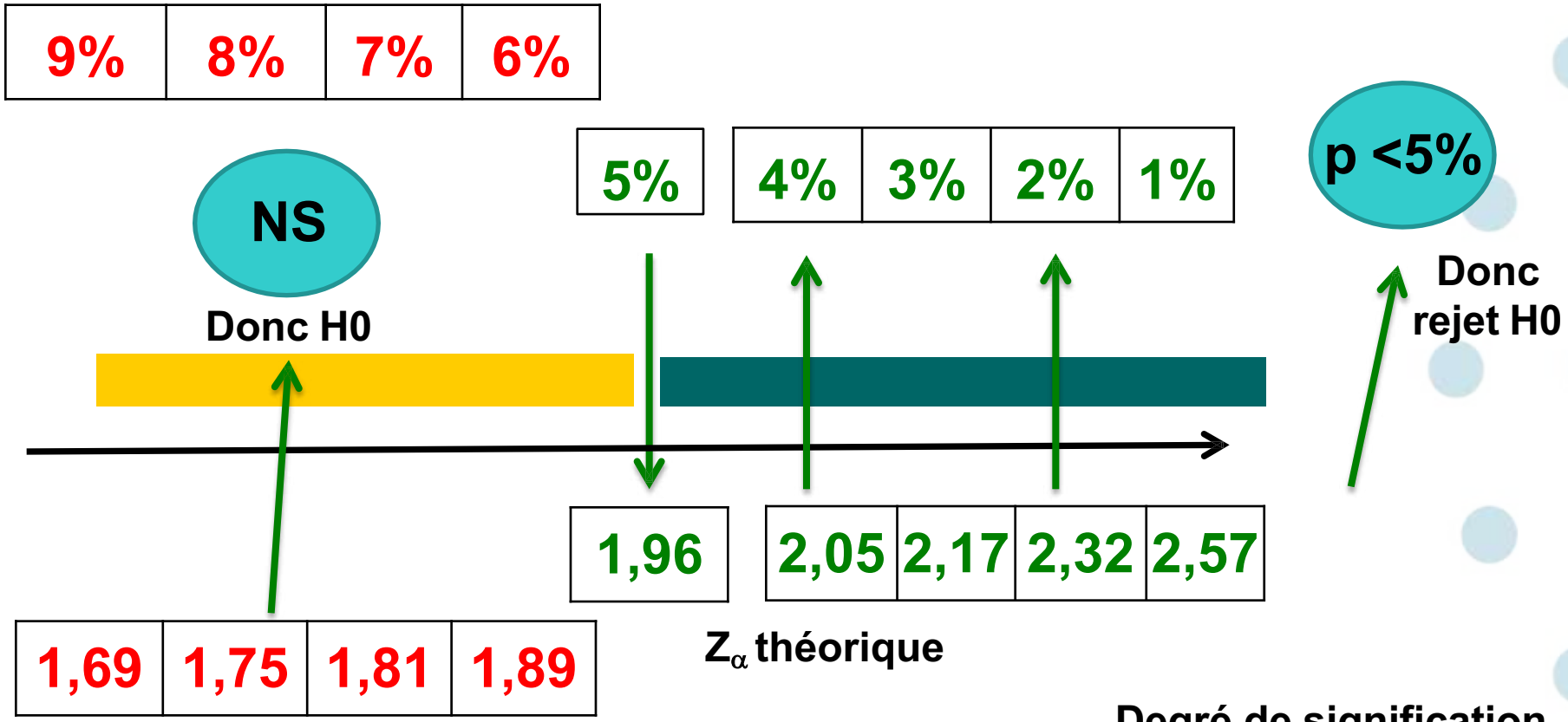
Interprétation graphique du risque α (2/3)



Degré de signification réel déterminé à posteriori

Z calculé 2,72

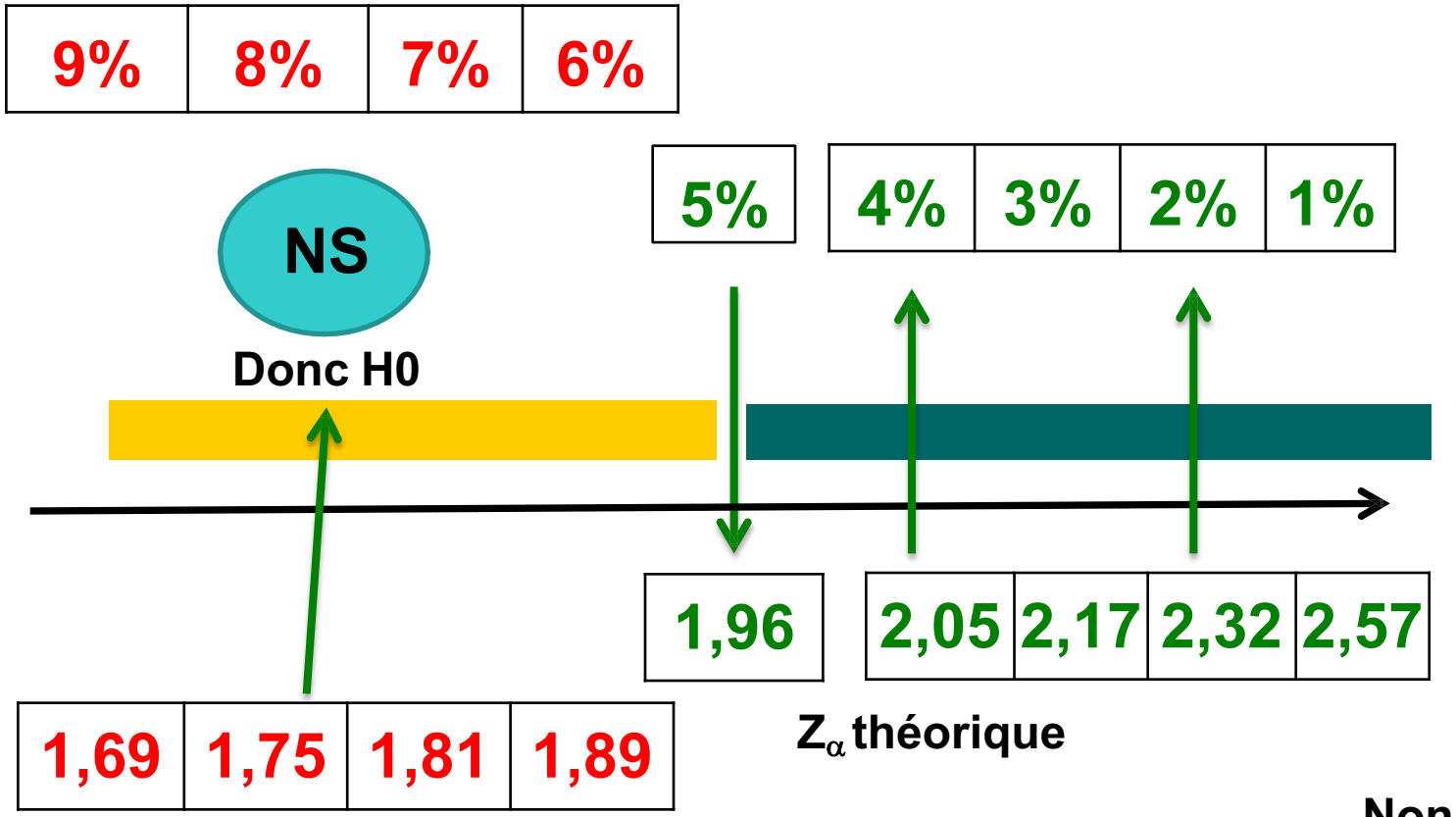
Interprétation graphique du risque α (3/3)



Degré de signification réel fixé <u>à posteriori</u>

Z calculé 2,22

Interprétation graphique du risque α (3/3)



Non significatif
<u>à posteriori</u>

Z calculé 1,3

Qu'appelle t on risque ?

Risque de première espèce : α — **Probabilité de rejeter H0 si H0 vraie.**

compromis universel : $\alpha = 5\%$

Risque de seconde espèce : β — **Probabilité d'accepter H0 si H0 fausse**

Puissance d'un test : $1 - \beta$ — **Probabilité de rejeter H0 si H0 fausse.**

Il se peut que le risque de deuxième espèce β soit assez important. L'erreur α est celle qu'on choisit de maîtriser, quitte à ignorer β. Cela induit une dissymétrie dans le traitement des deux hypothèses.

La règle de rejet du test est définie uniquement à partir de α **et H0.** Entre deux alternatives, on choisira pour H0 l'hypothèse qu'il serait le plus grave de rejeter à tort.

Les risques d'erreur

		Décision du statisticien : Rejet H0	Décision du statisticien : Non rejet H0
Réalité	H0 Vraie	Erreur 1ère espèce α	$1 - \alpha$
Réalité	H1 Vraie	Puissance $1 - \beta$	Erreur 2ème espèce β

BIG DATA (DONNEES MASSIVES)

Et si les <u>données</u> étaient le pétrole du 21ème siècle ?

Nous générons et détenons quantités d'info personnelles >>

Alimentation, achats, contributions réseaux sociaux, goûts, préférences, recherches sur Google, santé connectée,…

Données éparses mais captées par différents intervenants sur Internet.

Domaine de la santé :

Etudes épidémiologiques diverses lancées (pour le meilleur et pour le pire ?) : société privées (USA) analysent ces data et en tirent des conclusions : proposent à des femmes l'ablation des 2 seins car leur profil génétique comparé à celui de milliers d'autres femmes >> risque accru de K sein.

Les objets connectés (bracelets, balances, tee-shirts, fauteuils, iwatch,..). Suivre sa propre forme physique, la comparer à ce qu'elle devrait être (!).
Mais alimentent aussi de manière continue ces fameuses Big Data.

L'utilisation de ces masses de données remet en cause certaines théories statistiques et la notion d'échantillonnage.

Jusqu'à aujourd'hui les données recueillies dans les études cliniques sont des données démographiques (sexe, âge), cliniques (poids, taille, diag, trait, dose, durée), biologiques,…Jamais de données de type psy, émotionnel, ..
Big Data : permettent de recouper et analyser TOUS ces types de données et de remettre en cause certaines conclusions ou décisions..

De plus : échantillon traditionnel = effectif de qq dizaines, au mieux qq centaines d'individus, représentant des populations cibles souvent de plusieurs centaines de milliers d'individus. ***Schéma le plus performant ?***

Grâce aux Big Data :
effectif de l'échantillon observé et étudié est de l'ordre de la population cible
Et çà, c'est tout de même un vrai bouleversement théorique !

LES ÉTAPES DE MISE EN ŒUVRE D'UN TEST D'HYPOTHÈSE

- Etape 1 : Avant recueil des données définir H0 et H1
- Etape 2 : définir le test en fonction du type des données (qualitatives, quantitatives). Soit Z le paramètre calculé
- Etape 3 : Avant recueil des données on choisi le risque a priori (α) (dans la pratique souvent 5%)
  - Etape 4 : Recueil des données.
  - Calcul de Z.
  - Règle de décision : examiner la position de cette valeur Z, par rapport à un modèle théorique dont on connait la distribution
    Fixation du risque d'erreur réel à posteriori
  - Etape 5 : Interprétation des résultats.

Etude de la liaison entre 2 caractères qualitatifs.

Question :

Le pourcentage p_A d'un certain type d'individus dans un groupe *A* coïncide-t-il avec le pourcentage p_B du même type d'individus dans un autre groupe *B* ?

1) Comparaison de 2 pourcentages observés.

$$\varepsilon = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

ε = 1,96 avec α = 5 %

q = probabilité complémentaire de p = 1 – p

2) Test du χ^2

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

χ^2 tabulé

Nb ddl = (nb lignes-1)(nb colonnes-1)

Comparaison de pourcentages observés

On cherche à savoir si le mode de garde (crèche ou domicile) modifie le risque de rhinopharyngite des enfants. On fait une étude sur 2 groupes de 200 enfants :

Crèche	n_A=200	Nb rhino = 130
Domicile	n_B=200	Nb rhino = 96

Le mode de garde influe-t-il sur le risque d'avoir une rhinopharyngite? Quel test statistique, et conclusion ?

Comparaison de pourcentages observés

1. H0 : Pas de différence entre les 2 modes de garde vis-à-vis des rhinopharyngites
   H1 : Différence entre les 2 modes de garde

2. Caractère qualitatif 1 : Garde en crèche ou à domicile
   Caractère qualitatif 2 : Avoir une rhinopharyngite ou non
   Donc
   Test = Comparaison de pourcentages

3. α = 5% défini à priori

4. p_A = 130/200 = 65% p_B = 96/200 = 48%

$$\varepsilon = \frac{0,65 - 0,48}{\sqrt{\frac{0,65x0,35}{n_A} + \frac{0,48x0,52}{n_B}}} = 3,4$$

5. Table de l'écart réduit $\varepsilon > 3,3$ ($p < 1\,‰$)

Table de l'écart réduit

α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	**1,96**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	**3,8905**	4,41717	4,89164	5,32672	5,73073	6,10941

Le test statistique vient de démontrer **sur cet échantillon**, que le risque de rhinopharyngites est supérieur chez les enfants gardés en crèche

(p <0,001) défini à posteriori

Conclusion :

Sur cet échantillon le mode de garde est cause de cette différence

On ne pourra pas généraliser cette conclusion au niveau de tous les enfants en âge d'être gardés en France ou ailleurs, car :

Il n'y a pas eu TAS. On ne sait rien des enfants, ni des lieux de garde. Les familles n'ont peut être pas les mêmes revenus, donc l'accès aux soins n'est pas forcément le même..

Nous distinguerons toujours les 2 aspects à discuter :

a) L'aspect statistique et ses conclusions

b) L'aspect médical et ses conclusions qui peuvent être différentes.

On veut étudier l'efficacité d'un nouveau traitement (T) contre la leucémie.

On administre T à 50 souris et le traitement de référence (R) à 50 autres souris de la même espèce. On note au bout d'un mois, 33 morts dans le groupe T et 44 morts dans le groupe R.

Peut on conclure à la supériorité de ce nouveau traitement?

1. H0 : Il n'y a pas de différence significative entre les 2 traitements.
2. Traitements = 2 groupes T ou R : variable qualitative 1
3. Etat des souris dans chaque groupe (DCD ou non) : variable qualitative 2
4. Test de comparaison de pourcentages On fixe a priori α = 5%

		Traitements	
		T	**R**
Etat	**DCD**	**33**	**44**
	VIVANTES	**17**	**6**

Comparaison de pourcentages observés

groupe T : % DC = 33/50=66 % groupe R : %DC = 44/50=88%

On peut calculer ε

$$\varepsilon = \frac{0{,}88 - 0{,}66}{\sqrt{\frac{0{,}88 \times 0{,}12}{50} + \frac{0{,}66 \times 0{,}34}{50}}} = 2{,}83$$

ε calculé = 2,83 > ε théorique lu dans la table pour α=5% soit 1,96

On rejette H0. Il existe une diff significative entre les 2 groupes (H1),

α<0,01 à posteriori

On peut conclure : % DC souris traitées par T < % DC souris traitées par R

T est meilleur que R **sur cet échantillon.**

On ne sait rien des groupes, de l'état de santé des souris, et même de l'étude.

On ne peut pas généraliser ce résultat.

Table de l'écart réduit

α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	**1,96**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

α

ε

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	**3,8905**	4,41717	4,89164	5,32672	5,73073	6,10941

Etude de la liaison entre 2 caractères qualitatifs.

1 - Comparaison de 2 pourcentages observés.

$$\varepsilon = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

$\varepsilon = 1{,}96$ avec $\alpha = 5\ \%$

q = probabilité complémentaire de p = 1 – p

2 - Test du χ^2

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

χ^2 tabulé

Nb ddl = (nb lignes-1)(nb colonnes-1)

Précisions concernant les ddl

ddl = nb minimal de valeurs d'une série, nécessaire afin de pouvoir calculer les manquants si l'on dispose du total ou des totaux des valeurs de cette série

1 – Test du X^2

Exemple

CAR 1

CAR 2	A	B	Total
C	n_1	n_3	Tot C
D	n_2	n_4	Tot D
Total	Tot A	Tot B	T

ddl = 1

Test du χ^2

On cherche à savoir si l'exposition professionnelle au benzène peut entrainer une leucémie. On lance une étude dans une grande entreprise, on dénombre les salariés exposés au benzène, et ceux qui ne le sont pas. Au bout de 12 ans, on fait le bilan des leucémies apparues.

	Leucémies	Non Leucémies	Total
Expo	15	485	500
Non Expo	20	980	1000
Total	35	1465	1500

Existe-t-il une relation entre exposition au benzène et leucémies ?

Test du χ^2

1. H0 : il n'existe pas de lien entre expo benzène et leucémie
2. Variable qualitative 1 : Malades leucémie ou non malades leucémie
3. Variable qualitative 2 : Exposition au benzène ou non exposition
4. Test du $\chi 2$ permet de prendre en compte tous les cas de figure (expo-malades, expo-non malades, non expo-malades, non expo-non malades) et pas seulement deux %.
5. Si répartition au hasard : les nb de leucémies seraient à peu près identiques dans les 2 groupes Expo et Non expo.
6. Nous allons donc construire ce modèle et comparer la situation réelle à ce modèle théorique.

Test du χ^2

35/1500 = 2,33% malades, expo ou non, et 1465/1500 = 97,66% non malades.

Appliquons ces % aux salariés expo et non expo : **modèle théorique**.

2,33 % de 500 = 11,65 salariés, chiffre théorique de malades chez les expo.

2,33 % de 1000 = 23,35 salariés, chiffre théorique de malades chez les non expo.

	Leucémies	Non Leucémies	Total
Expo	15	485	500
Non Expo	20	980	1000
Total	35	1465	1500
%	2,33	97,66	

Les chiffres calculés (rouge), forment le modèle théorique. Nous allons les comparer aux chiffres observés (noir), à l'aide de la formule ci-dessous :

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

Test du χ^2

	Leucémies		Non Leucémies		Total
Expo	15	11,65	485	488,3	500
Non Expo	20	23,35	980	976,7	1000
Total	35	35	1465	1465	1500

$\chi 2 = (15-11,65)^2/11,65 + (20-23,35)^2/23,35 + (485-488,3)^2/488,3 + (980-976,7)^2/976,7$

$\chi^2 = 1,42$

Nb de degrés de liberté = (nb lignes-1)(nb colonnes – 1) = 1

La table du χ^2 indique que χ^2 calculé < χ^2 théorique (α = 5%, soit 3,84)

Nous acceptons H0 :

Il n'existe pas de relation entre expo benzène et apparition des leucémies.

Table du $\chi 2$ (extrait)

α

ddl \ α	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,016	0,455	1,074	1,642	2,706	**3,841**	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,21	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,61	4,351	6,064	7,289	9,236	11,07	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,49	7,344	9,524	11,03	13,362	15,507	18,168	20,09	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,34	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,34	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,79	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,79
...									

Etude de la liaison entre 2 caractères qualitatifs.

1 - Comparaison de 2 pourcentages observés.

$$\varepsilon = \frac{p_A - p_B}{\sqrt{\frac{p_A q_A}{n_A} + \frac{p_B q_B}{n_B}}}$$

$\varepsilon = 1{,}96$ avec $\alpha = 5\ \%$

q = probabilité complémentaire de p = 1 – p

2 - Test du χ^2

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

χ^2 tabulé

Nb ddl = (nb lignes-1)(nb colonnes-1)

Test du χ^2

On veut étudier l'efficacité d'un nouveau traitement T contre la leucémie.

On administre T à 50 souris et le traitement de référence R à 50 autres souris de la même espèce. On note au bout d'un mois, 33 DC dans le groupe T et 44 DC dans le groupe R.

Peut on conclure à la supériorité de ce nouveau traitement?

1. On dispose de 2 groupes indépendants traités T ou R : variable qualitative
2. Dénombrements des DC dans chaque groupe (DC ou non) : variable qualitative
3. Comparaison de % (vu précédemment), ou test du χ^2
4. H0 : Il n'y a pas de différence significative entre les 2 traitements.

Test du χ^2

	T	R	TOTAL
Morts	33 (38,5)	44 (38,5)	77
Vivants	17 (11,5)	6 (11,5)	23
TOTAL	50	50	100

77% DC

23% VIVANTS

77% DC
Gr T : 77% x 50 = 38,5
Gr R : 77% x 50 = 38,5
23% VIVANTS
Gr T et R : 23% x 50 = 11,5

$$\chi^2 = \sum \frac{(O_i - C_i)^2}{C_i}$$

Si hasard : même nb de DC et de vivants dans les 2 groupes…

$\chi 2 = 6,83$ ddl = 1

$\chi 2$ théorique = 3,84 ; $\alpha = 5\%$

$\chi 2$ calculé > $\chi 2$ théorique : **on rejette H0**

Il existe une différence significative entre les 2 traitements ($\alpha < 1\%$).

On a le droit d'interpréter le contenu du tableau :

on peut conclure que T est meilleur que R sur cet échantillon.

PLAN GÉNÉRAL DU COURS

1. **Biostatistique**
2. Statistique Descriptive
3. Statistique Déductive
   - Liaisons entre caractères qualitatifs
   - Liaisons entre caractères qualitatifs et quantitatifs
   - Liaisons entre caractères quantitatifs
   - Tests non paramétriques

Liaison entre caractères qualitatifs et quantitatifs

Question : En moyenne la taille des individus d'une population *A* coïncide-t-elle avec la taille des individus d'une population *B* ?

1 - Comparaison de moyennes

n_1 et $n_2 > 30$ "Grands échantillons"

$$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Table de l'écart réduit

$\varepsilon = 1{,}96$ avec $\alpha = 5\ \%$

2 - Test t de Student

n_1 ou $n_2 < 30$ "Petits échantillons"

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Table t de Student

Nb ddl = $(n_1-1)+(n_2-1)$

$$s = \sqrt{\frac{\sum (x_i - m_1)^2 + \sum (x_j - m_2)^2}{(n_1 - 1) + (n_2 - 1)}}$$

= écart type sur les 2 échantillons

Comparaison de moyennes

On cherche à comparer les taux de T3 libre (hormone thyroïdienne) chez des femmes prenant un contraceptif oral (c.o) et chez des femmes n'en prenant pas.

On dispose, après TAS de 2 groupes de femmes :

Femmes sans c.o $n_1=50$ $m_1=2$ nmol $s_1=0,35$ nmol

Femmes avec c.o $n_2=33$ $m_2=2,5$ nmol $s_2=0,30$ nmol

Les taux de T3 libre peuvent ils être considérés comme « identiques »

dans les 2 groupes ou bien sont ils significativement différents ?

1. H0 : m_1 et m_2 ne sont guères différentes. Ce sont 2 estimateurs de la valeur moyenne de T3 libre chez la femme, en général.

2. Relation entre caractères qualitatifs (prise ou non de c.o) et quantitatifs (dosages de T3 libre, ici valeur moyenne)

3. n_1 et $n_2 > 30$ ⟹ test de comparaison de moyennes

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{2{,}5 - 2}{\sqrt{\frac{0{,}35^2}{50} + \frac{0{,}30^2}{33}}} = 6{,}94$$

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

$\alpha = 0{,}0001$ $\varepsilon = 3{,}89$ Très significatif. On rejette H0 avec $p < 0{,}0001$

TAS : donc médicalement, résultat généralisable :

La prise de contraceptifs oraux augmente le taux de T3libre

Table de l'écart réduit

α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	**1,96**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	**3,89059**	4,41717	4,89164	5,32672	5,73073	6,10941

On teste un antiviral diminuant le nb de jours de symptômes cliniques chez des patients infectés par le virus de la grippe.

Soit 100 sujets non traités, atteints de grippe.

Nb moyen de jours avec symptômes $m_1 = 4{,}74$ jours et $s_1 = 1$.

Soit 100 autres sujets traités avec l'antiviral et atteints de grippe

Nb moyen de jours avec symptômes $m_2 = 4{,}2$ jours et $s_2 = 1{,}7$.

Parmi les propositions suivantes choisir celles qui sont exactes :

A) Le test de comparaison de moyennes permet de rejeter/accepter H0
B) Le test de comparaison de moyennes ne permet pas de rejeter/garder H0
C) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 1ère espèce 5%
D) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 2ème espèce inconnu
E) On ne pourra pas généraliser le résultat car l'étude n'a pas été bien menée

A) Le test de comparaison de moyennes permet de rejeter/accepter H0
B) Le test de comparaison de moyennes ne permet pas de rejeter/garder H0
C) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 1ère espèce 5%
D) On ne pourra pas conclure à l'efficacité du tt à cause du risque de 2ème espèce inconnu
E) On ne pourra pas généraliser le résultat car l'étude n'a pas été bien menée

Réponse A. H0 : m1 et m2 ne sont pas significativement différentes
Le test de comparaison de moyennes, comme tous les tests, répond à la question : peut on accepter ou rejeter H0 ?

Réponse E. L'étude aurait due être randomisée (TAS) : Traitement contre Placebo

Les items **B, C, D** sont faux (B), ou sans rapport (C, D)

Table de l'écart réduit

α

		0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	∞	2,576	2,326	2,17	2,054	**1,96**	1,881	1,812	1,751	1,695
0,1	1,645	1,598	1,555	1,514	1,476	1,44	1,405	1,372	1,341	1,311
0,2	1,282	1,254	1,227	1,2	1,175	1,15	1,126	1,103	1,08	1,058
0,3	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,86
0,4	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,69
0,5	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,6	0,524	0,51	0,496	0,482	0,468	0,454	0,44	0,426	0,412	0,399
0,7	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,8	0,253	0,24	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,9	0,126	0,113	0,1	0,088	0,075	0,063	0,05	0,038	0,025	0,013

Table pour les petites valeurs de la probabilité

0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
3,2905	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

Liaison entre caractères qualitatifs et quantitatifs

Question : En moyenne la taille des individus d'une population A coïncide-t-elle avec la taille des individus d'une population B ?

1 - Comparaison de moyennes

n_1 et $n_2 > 30$ "Grands échantillons"

$$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Table de l'écart réduit

$\varepsilon = 1{,}96$ avec $\alpha = 5$ %

2 - Test t de Student

n_1 ou $n_2 < 30$ "Petits échantillons"

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Table t de Student

Nb ddl = $(n_1-1)+(n_2-1)$

$$s = \sqrt{\frac{\sum (x_i - m_1)^2 + \sum (x_j - m_2)^2}{(n_1 - 1) + (n_2 - 1)}}$$

= écart type sur les 2 échantillons

Précisions concernant les ddl

2 – Série numérique : t de Student

Exemple série 8 valeurs donc n = 8

	2	3	5	12	10	4	7	8	Total = 51
1 valeur manquante	2	3	5	12	10		7	8	Total = 47

Manquant = 51 – 47 = 4 avec n-1 valeurs : on peut calculer la valeur manquante à partir du total

2 valeurs manquantes	2	3		12	10	7	8	Total = 42

Somme des manquants = 51 – 42 = 9 impossible de calculer les 2 valeurs manquantes

Donc ddl = n-1 = 7

t de Student : 2 séries à comparer donc ddl = $(n_1 - 1) + (n_2 - 1)$

Soient un groupe de 15 femmes obèses, et un autre groupe de 12 femmes de poids normal. On a mesuré le taux de corticoïdes sanguins moyens à l'intérieur de ces 2 groupes.

Gr 1 :	$n_1 = 15$	$m_1 = 6,3$	$s_1 = 1,8$
Gr 2 :	$n_2 = 12$	$m_2 = 4,5$	$s_2 = 1,6$

L'obésité a-t-elle une influence sur le taux de corticoïdes ?

Test t de Student

1. H0 : m_1 et m_2 ne sont pas différentes dans ces 2 groupes.

2. Relation entre caractères qualitatifs (obèses et non obèses), et quantitatifs (valeurs de dosages sanguins, valeurs moyennes)

3. n_1 et n_2 < 30 petits échantillons ⇒ t de student

4. Calcul de l'écart type commun aux 2 groupes. La formule s'écrit :

$$s^2 = \frac{(n_1-1) \times s_1^2 + (n_2-1) \times s_2^2}{(n_1+n_2-2)} = 2,53$$

nb ddl = (15+12) – 2 = 25

$$t = \frac{6,3 - 4,5}{\sqrt{\frac{2,53}{15} + \frac{2,53}{12}}} = 2,92$$

> 2,06 à 5% lu dans la table t Student

On rejette H0 avec α = 1% défini a posteriori.
Il existe une relation entre obésité et augmentation du taux de corticoïdes au niveau de ces échantillons.

TABLE DU t DE STUDENT

α

ddl	0,9	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	1	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,92	4,303	6,965	9,925	31,598
3	0,137	0,765	1,25	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,19	1,533	2,132	2,776	3,747	4,604	8,61
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,44	1,943	2,447	3,143	3,707	5,959
7	0,13	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,13	0,706	1,108	1,397	1,86	2,306	2,896	3,355	5,041
9	0,129	0,703	1,1	1,383	1,833	2,262	2,821	3,25	4,781
10	0,129	0,7	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,35	1,771	2,16	2,65	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,14
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
17	0,128	0,689	1,069	1,333	1,74	2,11	2,567	2,898	3,965
18	0,127	0,688	1,067	1,33	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,85
....									
25	0,127	0,684	1,058	1,316	1,708	2,06	2,485	2,787	3,725

Liaison entre caractères qualitatifs et quantitatifs

Séries appariées
ou
Méthode des couples

1 - Comparaison de moyennes

$$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2 - Test t de Student

$$t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

Liaison entre caractères qualitatifs et quantitatifs.

Exemple

Comparer deux méthodes de dosage de la glycémie. On dispose de n patients, auxquels on prélève 2 tubes de sang. On dose la glycémie dans chacun de ces tubes par une méthode différente.

On souhaite comparer les valeurs moyennes de ces 2 séries de n résultats. La question posée est :

Ces 2 méthodes de dosage fournissent elles des résultats identiques ?

On calcule **si n>30** $\varepsilon = m_d / \sqrt{\frac{s^2}{n}}$ **si n<30** $t = m_d / \sqrt{\frac{s^2}{n}}$

Avec d=différence des résultats pour un même sujet, m_d=moyenne des d, n=nb de couples, s=variance des différences

Puis la méthodologie est identique aux tests déjà vus : on compare cette valeur calculée aux valeurs dans la table adaptée, et la conclusion se fait de la même manière **en fixant un risque α.**

Cas des séries appariées. Méthode des couples

On souhaite évaluer l'intérêt d'une substance S capable de désintoxiquer les fumeurs. On constitue par T.A.S. 2 groupes de 40 fumeurs, un reçoit S, l'autre reçoit un placebo P. Le traitement dure 2 mois pour les 2 groupes. La consommation de cig/jours (C) est notée avant et après traitement.

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

Cas des séries appariées. Méthode des couples

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

1) Quelle est la première précaution à prendre ?

2) Dans le groupe Placebo, la conso moyenne après tt diffère t elle de la valeur avant tt ? Interpréter le résultat.

3) Les 2 groupes diffèrent ils pour leur conso moyenne après traitement?

4) Les 2 groupes diffèrent ils pour la variation de conso avant/après tt ?

Cas des séries appariées. Méthode des couples

	S (n=40)		P (n=40)	
	m_1	s_1^2	m_2	s_2^2
C avant tt	19,5	54,2	16,5	35,6
C après tt	5,4	30,4	3,8	20,1
Variation de C	14,1	9,1	12,7	8,9

1)Quelle est la première précaution à prendre ?

Conso identique dans les 2 groupes?

Les 2 groupes doivent être comparables vis-à-vis des paramètres susceptibles d'influencer la réponse au traitement (âge, sexe, CSP, conso/jour etc..).

Si ce n'est pas le cas, il faut en tenir compte lors des conclusions.

Cas des séries appariées. Méthode des couples

Comparaison des conso moyennes avant tt dans les 2 groupes:

1. **H0 = les moyennes des conso sont équivalentes dans les 2 groupes.**
2. **Etude liaison entre variables qualitatives (S ou P) et quantitatives (nb cig/j) dans 2 échantillons indépendants**
3. **n> 30 ⇒ Test de comparaison de moyennes**
4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s^2}{n_2}}} = \frac{19,5 - 16,5}{\sqrt{\frac{54,2}{40} + \frac{35,6}{40}}} = 2,00 > 1,96 \quad (\varepsilon \text{ pour } \alpha = 5\%)$$

On rejette H0, avec un risque α = 5%. Il existe donc une différence significative entre les conso moyennes des 2 groupes : on fume plus dans le groupe S : il faudra en tenir compte lors de l'étude de la variation de cette conso avant/après traitement.

Cas des séries appariées. Méthode des couples

2) Dans le groupe Placebo, la conso moyenne après tt diffère t elle de la valeur avant tt ? Interpréter le résultat.

P (n=40)	
m_2	$s_2{}_2$
16,5	35,6
3,8	20,1
12,7	8,9

1.Liaison entre variable qualitative (avant / après tt) et quantitative (nb cig/j)

2. Echantillons non indépendants (méthode des couples)

3. n >30 ⇒ **Test de comparaison de moyennes**

$\varepsilon = m_d / \sqrt{\frac{s^2}{n}} = 12,7 / \sqrt{\frac{8,9}{40}} = 26,9$ > 1,96 au risque $\alpha = 5\%$

On rejette H0. Il existe une différence très significative (p < 0,001) entre les consommations avant / après tt, dans le groupe P.

Effet psychologique : envie de profiter de l'étude pour arrêter de fumer ?

Cas des séries appariées. Méthode des couples

3) Les 2 groupes diffèrent ils pour leur conso moyenne après traitement?

S (n=40)		P (n=40)	
m_1	$s_1{}^2$	m_2	$s_2{}^2$
19,5	54,2	16,5	35,6
5,4	30,4	3,8	20,1
14,1	9,1	12,7	8,9

1. H0 = les moyennes des conso sont équivalentes dans les 2 groupes.

2. Liaison entre variables qualitatives (S ou P) et quantitatives(nb cig/j) dans 2 échantillons indépendants

3. n> 30 ⟹ Test de comparaison de moyennes

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5,4 - 3,8}{\sqrt{\frac{30,4}{40} + \frac{20,1}{40}}} = 1,42 < 1,96$$

On accepte H0 : il n'existe pas de différence significative entre les 2 groupes pour la consommation après tt.

4) Les 2 groupes diffèrent-ils pour la variation de conso avant/après tt ?

Il faut comparer les variations avant / après tt dans les 2 groupes afin de prouver l'intérêt de la substance S

1.H0 : Il n'existe pas de différence entre les variations de consommation dans les 2 groupes
2.Etude liaison entre variables qualitatives (S ou P) et quantitatives (nb cig/j) dans 2 échantillons indépendants
3. n > 30 >>>> Test de comparaison de moyennes

4. $$\varepsilon = \frac{m_1 - m_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{14{,}1 - 12{,}7}{\sqrt{\frac{9{,}1}{40} + \frac{8{,}9}{40}}} = 2{,}09 > 1{,}96$$ **au risque 5%**

<u>On rejette H0 :</u> il existe une différence significative entre les variations de conso dans les 2 groupes (p<5%) . Conclusion : efficacité de S. Il y avait eu TAS, donc résultat généralisable.

Conclusion : Pas de différence après tt dans chaque groupe (Quest 3). Mais Gr S fumait + (Quest 1) >>>> Efficacité du traitement S.

PLAN GÉNÉRAL DU COURS

1. Biostatistique
2. Statistique Descriptive
3. **Statistique Déductive**
   - Liaisons entre caractères qualitatifs
   - Liaisons entre caractères qualitatifs et quantitatifs
   - Liaisons entre caractères quantitatifs
   - Tests non paramétriques

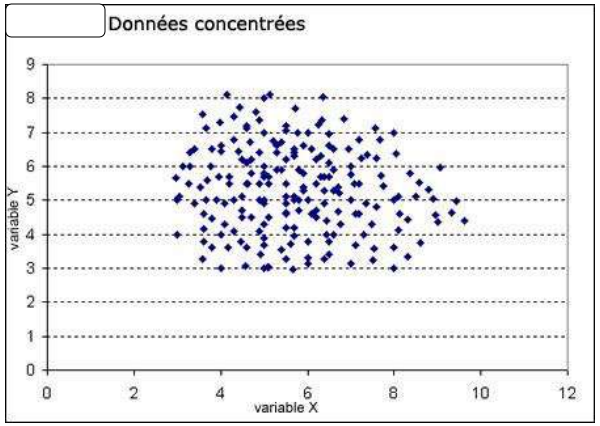
Etude de la liaison entre caractères quantitatifs

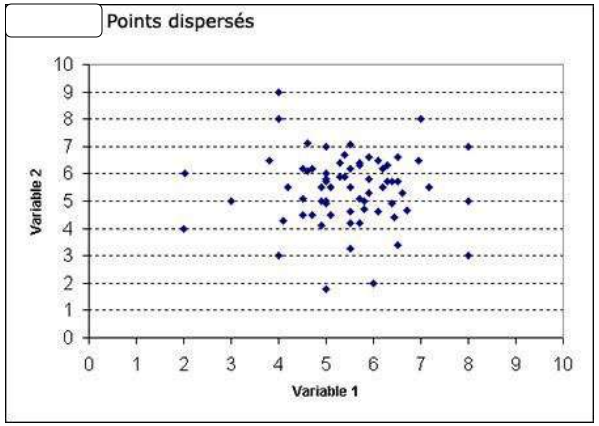
Corrélation et régression

Corrélation = Evaluation de la liaison entre 2 variables quantitatives

Régression = Méthode mathématique expliquant les relations entre variables observées.

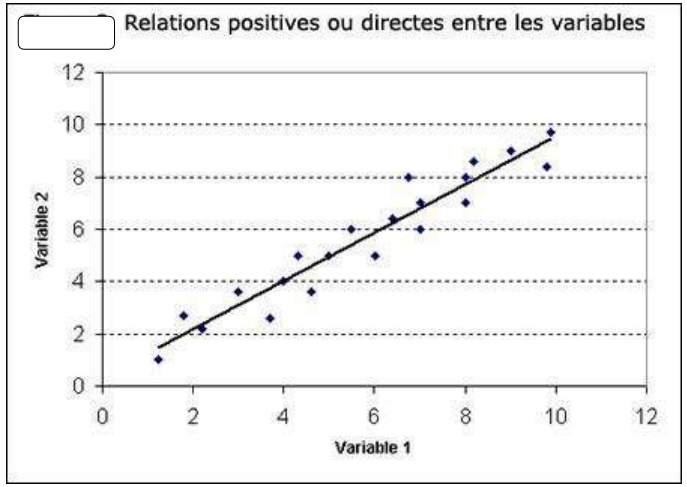
Représentation des données : **nuages de points**





Etude de la liaison entre caractères quantitatifs

Nuages de points



La droite de régression permet de visualiser si une des 2 variables est dépendante de l'autre..

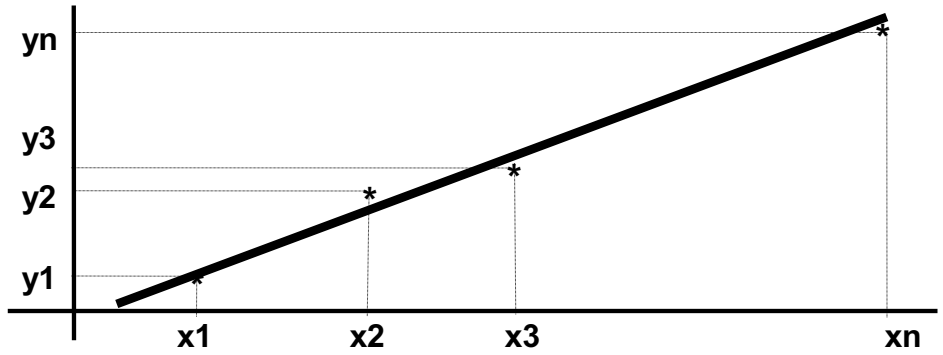
Etude de la liaison entre caractères quantitatifs

➢La capacité respiratoire est elle dépendante de la consommation de cigarettes?

➢Le poids des bébés à la naissance est il lié à l'âge de la mère?

si x et y liées alors ⇒ y = f(x) ***droite de régression de y en x***

y peut être « expliqué » en fc de x

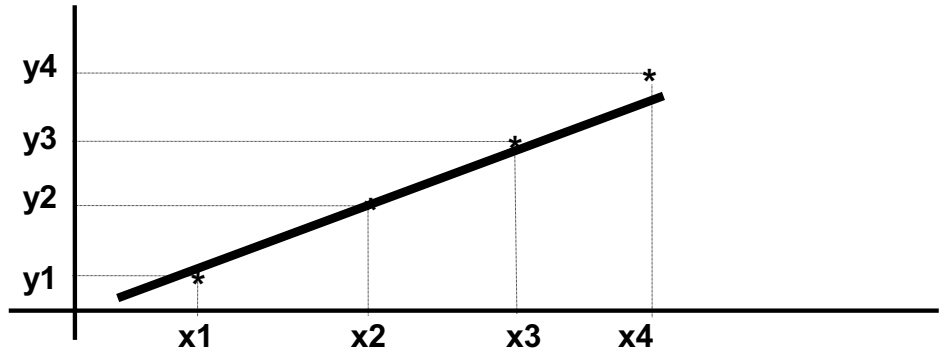


Droite de régression:

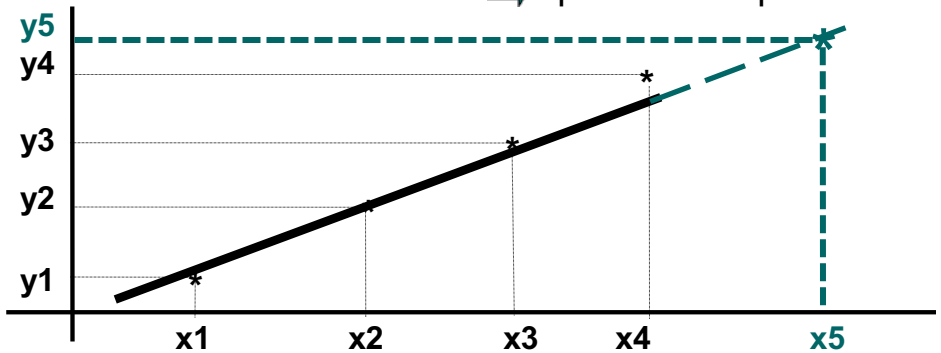
Droite des moindres carrés, passe au «plus près» de chaque point du graphe.

Dans ce cours, on ne parle que de **régression linéaire**.

Etude de la liaison entre caractères quantitatifs



La prédiction avec une droite de régression :
Pour nouvelle valeur de x ➡ quelle valeur possible de y ?



Corrélation

Sur un échantillon de 10 sujets d'âges différents, on recueille les données suivantes : âge (années) et concentration de cholestérol dans le sang (g/L)

X âges	30	60	40	20	50	30	40	20	70	60
Y chol	1,6	2,5	2,2	1,4	2,7	1,8	2,1	1,5	2,8	2,6

Le taux de cholestérol est il lié à l'âge ?

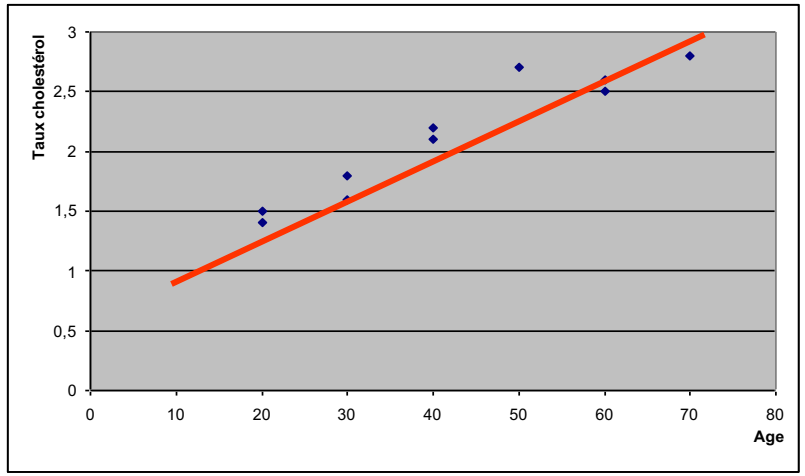
Corrélation

r calculé = 0,955

r théorique (α = 1%) avec 10-2 = 8 ddl = 0,76 Table

r calculé > r théorique **Rejet de H0**

Il existe une relation significative (α= 1%) entre l'âge et le taux de cholestérol



Plus l'âge augmente, plus le taux de cholestérol augmente
Résultat non généralisable

Corrélation n'est pas **causalité**

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119

df \ α	0.2	0.1	0.05	0.02	0.01	0.001
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512

- **Utilisation obligatoire si les effectifs sont trop faibles**
- **(< 5), avec des caractères quantitatifs. Dans ce cas, les populations ne se distribuent pas normalement.**
- **Présentent une excellente robustesse**

1) **LIAISON QUANTITATIFS / QUALITATIFS**
   - **U de Mann & Whitney**

2) **LIAISON ENTRE QUANTITATIFS**
   - **Spearman (= Corrélation)**

U de Mann & Whitney

- Le test de Wilcoxon-Mann-Whitney (ou test U de Mann-Whitney ou encore test de la somme des rangs de Wilcoxon) est un test statistique non paramétrique qui permet de tester l'hypothèse selon laquelle les moyennes de chacun de deux groupes de données sont proches.

2 échantillons indépendants E1 et E2 de taille n1 et n2

On souhaite tester l'hypothèse H0 disant que les moyennes expérimentales dans les 2 échantillons sont égales $\mu 1=\mu 2$

On trie les valeurs obtenues dans la réunion des 2 échantillons par ordre croissant

Pour chaque valeur xi issue de E1, on compte le nombre de valeurs issues de E2 situées après lui dans la liste ordonnée (celles qui sont égales à xi ne comptent que pour 1/2)

On note u1 la somme des nombres ainsi associés aux différentes valeurs issues de E1. On fait de même en échangeant les rôles des deux échantillons, ce qui donne la somme u2. Soit u la plus petite des deux sommes obtenues : $u = \min\{u1 ; u2\}$.

U de Mann & Whitney

On note U la variable aléatoire associée.

Pour n1 et n2 quelconques, on lit dans les tables du test de Mann et Whitney le nombre mα tel que, sous (H0), $P(U \leq m\alpha) = \alpha$.

On rejette (H0) au risque d'erreur α si $u \leq m\alpha$. Autrement on accepte (H0).

Si n1 et n2 sont assez grands (≥20 en général), sous (H0), U suit approximativement la loi normale N(μ,σ)

r' Spearman

$$r' = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

On a recensé pour 6 étudiants les notes obtenues au concours de PACES en Biostatistique, et le classement final à ce même concours.

On cherche à établir si il existe une relation entre cette note et le classement final.

X Biostat	Y Classement
12,4	210
4,9	555
18,1	6
5,4	445
19,4	5
16	14

r' Spearman

H0 : Il n'y a pas de lien entre ces 2 séries de valeurs numériques. Il s'agit de 2 séries indépendantes.

Données quantitatives : Coeff de corrélation, faibles effectifs ➡ r' Spearman

X Biostat	Rg X	Y Classement	Rg Y	d_i rang	$(d_i)^2$
12,4	3	210	4	-1	1
4,9	1	555	6	-5	25
18,1	5	6	2	3	9
5,4	2	445	5	-3	9
19,4	6	5	1	5	25
16	4	14	3	1	1

$$r' = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

r' Spearman

$$r' = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r' = 1 - \frac{6x70}{6x35} = -1$$

Dans la table théorique, avec n=6 : r'= 0,89 avec α= 5%

r'= 1 avec α= 1%

Le r' calculé à -1 ≤ r' théorique lu dans la table.

On repousse donc H0 (p<1%)

On met en évidence un lien très significatif entre ces 2 séries.

Il s'agit de 2 séries corrélées. Plus la note de Biostat est élevée, plus petit est le rang de classement (d'où le signe – pour r').

Table r' de Spearman

	α	
n	0.05	0.01
5	1.00	-
6	0.89	1.00
7	0.79	0.93
8	0.74	0.88
9	0.68	0.83
10	0.65	0.79

MÉTHODOLOGIE D'UTILISATION DES TESTS

Effectif	Données Quantitatives	Données Qualitatives	Données Qualitatives - Quantitatives
>4 & <12	r' de Spearman	Comp % ou χ^2	U Mann & Withney
>=12 & < 30	Coeff de corrélation r	Comp % ou χ^2	t Student
>= 30	Coeff de corrélation r	Comp % ou χ^2	Comp moyennes

Exercice de réflexion

On veut savoir si une nouvelle molécule présente un effet anti dépresseur. Pour cela, on organise un essai portant sur 20 malades dépressifs, répartis en 2 groupes.

Les 20 malades sont répartis par TAS en 2 groupes de 10 sujets, l'un recevant la nouvelle molécule, l'autre recevant un placebo.

On évalue les patients à l'aide d'une échelle numérique de 0 (non déprimé) à 50 (très déprimé). Le groupe témoin reçoit le placebo.

Les patients des 2 groupes sont évalués avant traitement, puis après traitement au bout de 28 jours.

Exercice de réflexion

Témoins		Traités	
J0	**J28**	**J0**	**J28**
34	**31**	**27**	**22**
30	**28**	**32**	**25**
25	**26**	**30**	**23**
27	**25**	**28**	**26**
31	**24**	**25**	**20**
24	**25**	**33**	**27**
28	**26**	**29**	**21**
30	**27**	**31**	**26**
35	**32**	**32**	**25**
26	**25**	**29**	**23**

Y a-t-il un effet placebo ?

A : On compare les scores (J0 et J28) Témoins

B : On compare les scores (J0 et J28) Témoins / Traités

C : On compare les scores J0 Témoins / Traités

D : On compare les scores J28 Témoins / Traités

H0 = le placebo n'a aucun effet (les scores J0 ne diffèrent pas des scores J28)

Comparaison de moyennes (n=10) donc

Test t student séries appariées ou Mann et Whitney

Groupe Témoins

J0	J28	Diff
34	31	3
30	28	2
25	26	-1
27	25	2
31	24	7
24	25	-1
28	26	2
30	27	3
35	32	3
26	25	1

$m_d = 21/10 = 2,1$

$S^2_d = 5,21$

$$t = \frac{2,1}{\sqrt{\frac{5,21}{10}}} = 2,91$$

t théorique (avec 10-1 = 9 ddl et α = 5%) = **2,26**

t calculé > t théorique Rejet de H0

Le placebo a un effet significatif

Exercice de réflexion

2. On évalue les patients à l'aide d'une échelle numérique de 0 (non déprimé) à 50 (très déprimé). Le groupe témoin reçoit le placebo.

Les patients des 2 groupes sont évalués avant traitement, puis après traitement au bout de 28 jours.

Témoins		Traités	
J0	**J28**	**J0**	**J28**
34	**31**	**27**	**22**
30	**28**	**32**	**25**
25	**26**	**30**	**23**
27	**25**	**28**	**26**
31	**24**	**25**	**20**
24	**25**	**33**	**27**
28	**26**	**29**	**21**
30	**27**	**31**	**26**
35	**32**	**32**	**25**
26	**25**	**29**	**23**

Le traitement est il efficace ?

Pour chaque patient :

A : On compare les scores (J0 et J28) Témoins

B : On compare les scores (J0 et J28) Témoins/Traités

C : On compare les scores J0 Témoins / Traités

D : On compare les scores J28 Témoins / Traités

Efficacité du traitement différent de Effet traitement !

Comparer les différences ($J_{28} - J_0$) de chaque patient, entre les 2 groupes

Le traitement est il efficace ?

Comparer les différences ($J_{28} - J_0$) de chaque patient, entre les 2 groupes.

Questions préliminaires :

Quelles sont les variables? Quel est le bon test ?

Variables qualitatives / variables quantitatives (Tt ou Placebo / Scores)

2 groupes indépendants de faible effectif (n=10).

Test t de Student ou Mann et Whitney.

Mann & Whitney :

Témoins $d=J_0-J_{28}$	3	2	-1	2	7	-1	2	3	3	1
Traités $d=J_0-J_{28}$	5	7	7	2	5	6	8	5	7	6

Valeurs de différences rangées par ordre croissant : ex aequo, classés ensemble.

Témoins
Traités

Ex aequo : rangs 1 et 2 donc -> rang 3/2=1,5

Ex aequo : rangs 4,5,6,7 donc -> rang 22/4=5,5

Différences	-1	-1	1	2	2	2	2	3	3	3
Rangs	1,5	1,5	3	5,5	5,5	5,5	5,5	9	9	9

Différences	5	5	5	6	6	7	7	7	7	8
Rangs	12	12	12	14,5	14,5	17,5	17,5	17,5	17,5	20

Mann & Whitney :

-1	-1	1	2	2	2	2	3	3	3	Témoins
1,5	1,5	3	5,5	5,5	5,5	5,5	9	9	9	Traités
		5	5	5	6	6	7	7	7	7	8
		12	12	12	14,5	14,5	17,5	17,5	17,5	17,5	20

Calcul des paramètres U

Soit U_1: pour chaque témoin, cumul des traités classés AVANT

U_1= 9 **U2** = 100-9 = **91** car on calcule n_1x n_2 = 100.

On compare donc **U=9** (le plus petit des 2) avec la table théorique

(α=5%, n_1=10, n_2=10) **U théorique = 23** : **U calculé < U théorique :**

Peu d'imbrication :

Rejet de H0 : les différences sont significativement plus importantes pour le traitement que pour le placebo, avec α < 5%

Table U de Mann Whitney (α = 5%)

n_1 est le plus petit des 2 effectifs, U le plus petit des 2 U calculés

n_2-n_1	n_1 = 1	2	3	4	5	6	7	8	9	10
0	-	-	-	**0**	**2**	5	8	13	17	**23**
1	-	-	-	1	**3**	6	10	15	20	26
2	-	-	0	2	5	8	12	17	23	29
3	-	-	0	3	6	10	14	19	26	33
4	-	-	1	4	7	11	16	22	28	36
5	-	-	2	4	8	13	18	24	31	39
6	-	0	2	5	9	14	20	26	34	42
7	-	0	3	6	11	16	22	29	37	45
8	-	0	3	7	12	17	24	31	39	48
9	-	0	4	8	13	19	26	34	42	52
10	-	1	4	9	14	21	28	36	45	55
11	-	1	5	10	15	22	30	38	48	
12	-	1	5	11	17	24	32	41	50	
13	-	1	6	11	18	25	34	43		
14	-	1	6	12	19	27	36	45		
...										
18	-	2	8	16	24	33				
19	-	3	9	17	25					
20	-	3	9	17	27					