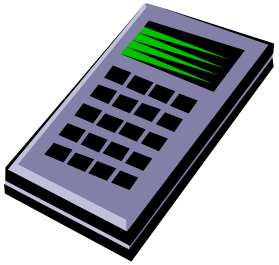
BIOSTATISTIQUE



n=2016?!

Faouzi LYAZRHI

Unité pédagogique de Biostatistique

École Nationale Vétérinaire de Toulouse

23 chemin des Capelles

B.P. 87614 31076 Toulouse Cedex 03

f.lyazrhi@envt.fr

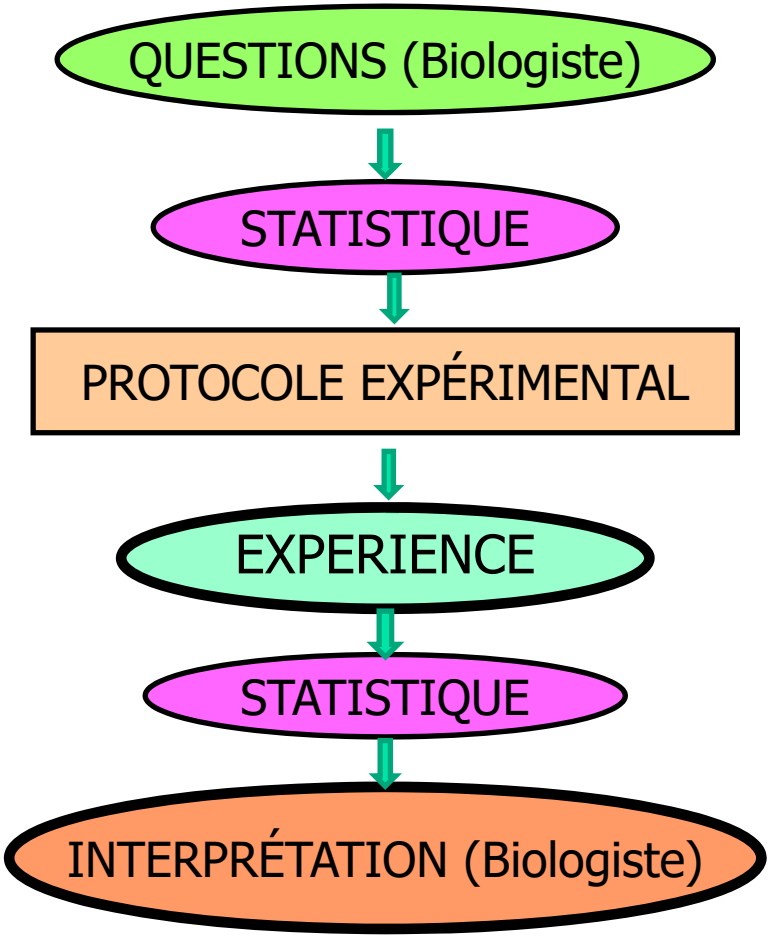
Introduction

La biostatistique est l'application de la statistique à un large éventail de sujets en biologie.
La biostatistique consiste au développement et l'application d'outils statistiques dans :

- La conception des expériences,
- La collecte et la présentation des données,
- L'analyse et l'interprétation des données recueillies en vue de tirer des conclusions ou prendre des décisions

Méthodologie

La biostatistique s'articule autour d'un dialogue entre le statisticien et le biologiste suivant le schéma ci-contre :



Avant l'expérience, on rédige un protocole expérimental qui décrit les conditions et le déroulement d'une expérience ainsi que la méthode choisie pour analyser les résultats attendus de l'expérience, d'où l'importance du dialogue entre le biologise et le statisticien.

Pourquoi de la biostatistique dans une école vétérinaire ?

Exemple 1 : On a dénombré sur 4900 naissances 2500 garçons (51%)

Ce résultat est-il compatible avec l'hypothèse d'équiprobabilité des naissances des garçons et des filles ?

Exemple 2 : Les guérisons d'une certaine maladie avec un traitement de référence et un traitement A ont été :

- traitement A : 85 guérisons sur 100 traités (85%)
- référence : 81 guérisons sur 100 traités (81%)

Est-ce que le traitement A est plus efficace que le traitement de référence ?

Pourquoi de la biostatistique dans une école vétérinaire ?

Exemple 3 :

Action d'un antibiotique sur des souches de bactéries issues d'individus atteints d'infection urinaire.

QUESTION :

L'âge, le sexe et le type de bactéries ont-ils une influence sur la CME ?

De la biostatistique dans une école vétérinaire ?

Etude comparative de l'influence d'un désinfectant (Dialox®) sur les paramètres physicochimiques d'un bain de dialyse

QUESTION :

Est-ce qu'il y a un effet du traitement sur le ph ?

De la biostatistique dans une école vétérinaire ?

Exemple 4 :

Tel chien, tel maître !!

QUESTION :

Est-ce qu'il existe un lien entre la taille d'un maître et celle de son chien ?

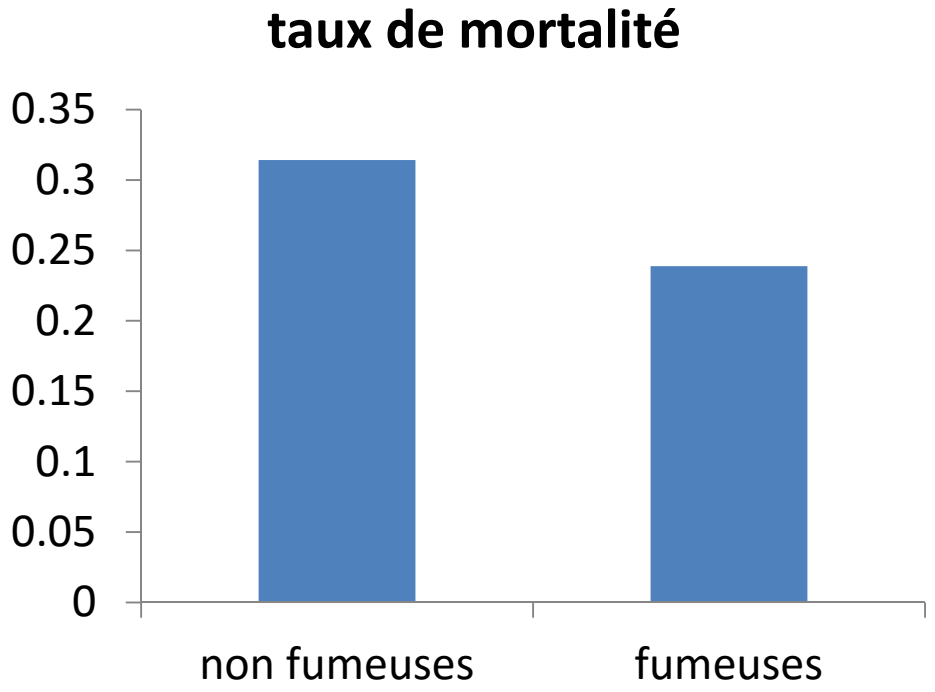
Faut-il se méfier de la statistique ?

Fumer, c'est bon pour la santé !!

(données fichier *smoking* du package *SMPractical)*

	vivantes	mortes	Total
non fumeuses	502	230	732
fumeuses	443	139	582
Total	945	369	1314

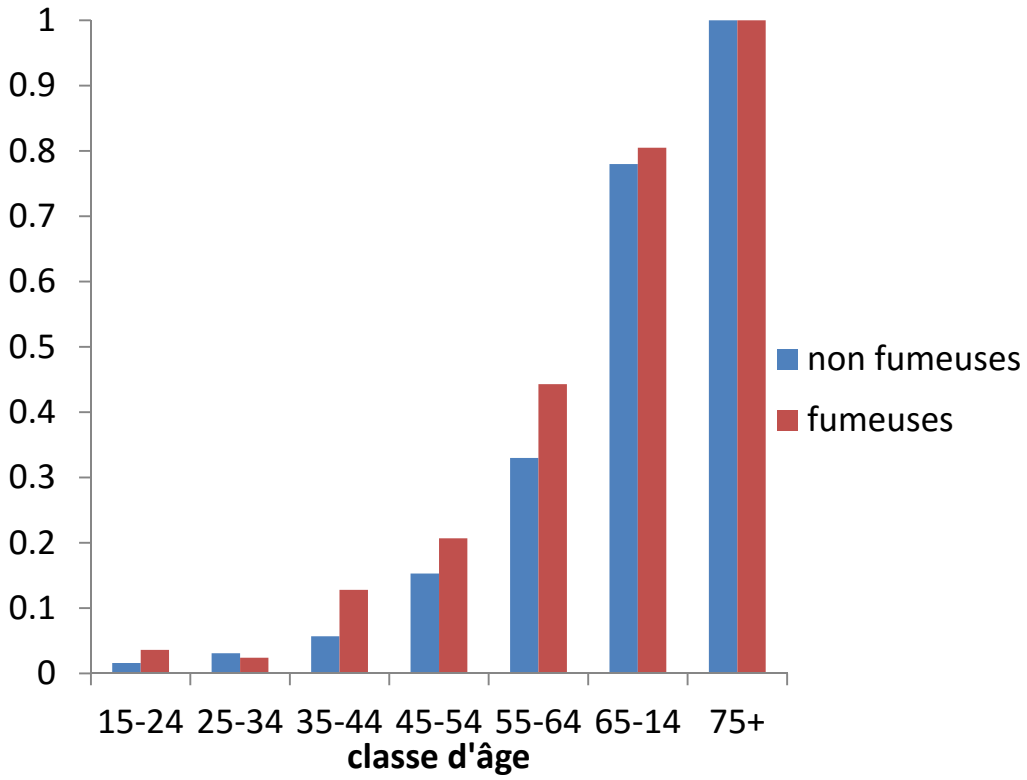
1314 femmes ont été suivies pendant 20 ans, et l'objectif était de comparer le taux de mortalité des fumeuses et des non-fumeuses.

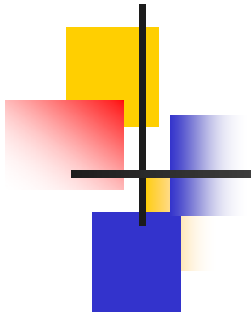


Faut-il se méfier de la statistique ?

Et si on regardait les données par classe d'âge :

âge	fumeur	vivante	morte	% mortes
18-24	fumeuse	53	2	0.0364
18-24	non fumeuse	61	1	0.0161
25-34	fumeuse	121	3	0.0242
25-34	non fumeuse	152	5	0.0318
35-44	fumeuse	95	14	0.1284
35-44	non fumeuse	114	7	0.0579
45-54	fumeuse	103	27	0.2077
45-54	non fumeuse	66	12	0.1538
55-64	fumeuse	64	51	0.4435
55-64	non fumeuse	81	40	0.3306
65-74	fumeuse	7	29	0.8056
65-74	non fumeuse	28	101	0.7829
75+	fumeuse	0	13	1
75+	non fumeuse	0	64	1





PLAN

Chap.1 Définitions de base

Chap.2 Statistique descriptive

Chap.3 Estimation d'un paramètre

Chap.4 Tests d'hypothèses

Chap1. Définitions de base

- Statistique vs. Statistiques
- Individu
- Population vs. Echantillon
- Echantillonnage aléatoire
- Statistique descriptive vs. statistique inférentielle
- Variable aléatoire
- Paramètre

STATISTIQUE vs. STATISTIQUES

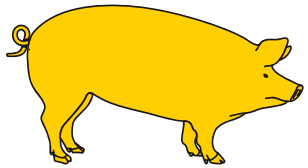
Des statistiques

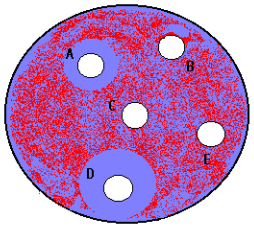
- 1,3 milliard d'habitant en chine
- 97 000 séropositifs en France en 2003
- En France, La taille moyenne d'un homme est de 1m 75 et celle d'une femme est de 1m 62
- En 205, 648.000 voitures neuves particulières ont été immatriculées en France

La statistique

L'activité qui consiste à recueillir, traiter et interpréter un ensemble de données.

INDIVIDU



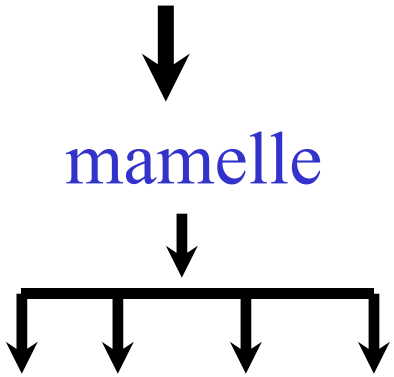


L'individu est l'unité expérimentale sur la quelle on fait la mesure.

Le terme individu doit être pris au sens large du terme, un individu peut être un patient, un animal, une plante, une souche, une mamelle, une pièce métallique fabriquée par une machine, etc.

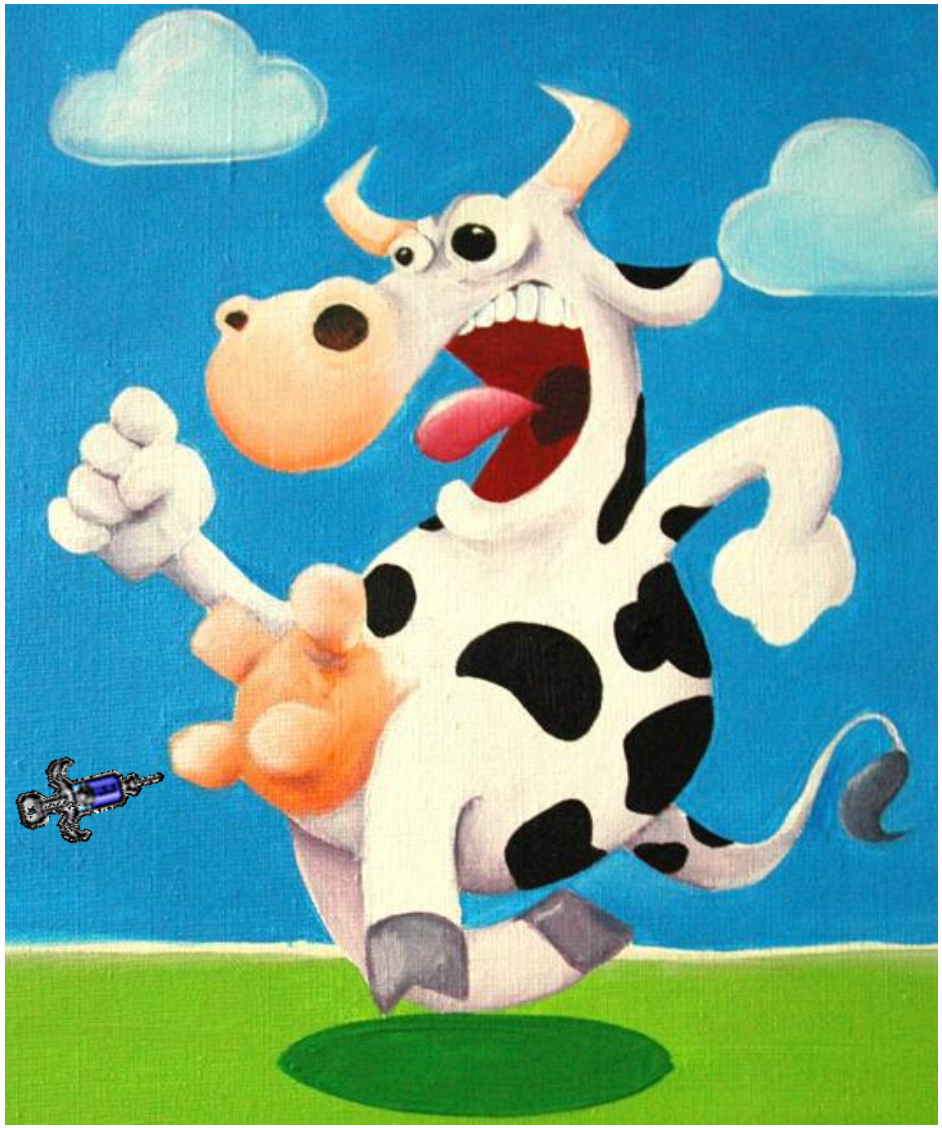
INDIVIDU

vache



mamelle

quartier



Population vs. échantillon

Population :
Ensemble d'individus de même nature auxquels on s'intéresse et sur lesquels une caractéristique pet être relevée, et auxquels les résultats d'une expérience sont extrapolés.
En général, la population est supposée de grande taille voire infinie.
On suppose qu'elle ne peut faire l'objet d'une étude exhaustive.
Les questions que l'on se pose et les hypothèses que l'on formule portent sur la population.

Échantillon :
Sous-ensemble de la population, il est constitué d'individus sur lesquels la caractéristique faisant l'objet de l'étude est effectivement mesurée. Un échantillon fournit des informations sur la population et les observations faites au niveau de l'échantillon servent à répondre aux questions que l'on s'est posé au niveau de la population.
Il est donc important que l'échantillon soit constitué de telle sorte qu'il puisse remplir ce rôle.

Échantillonnage aléatoire

Échantillon représentatif :
Un échantillon est dit représentatif si sa composition est conforme à celle de la population.
La façon la plus simple de constituer un échantillon représentatif est l'échantillonnage aléatoire.

Échantillonnage aléatoire simple :
Dan un échantillonnage aléatoire simple, la population est supposée homogène et chaque individu de la population a une chance égale de faire partie de l'échantillon.

Échantillonnage aléatoire simple stratifié :
Dans un échantillonnage aléatoire stratifié, on suppose que la population est constituée de strates (sous-populations homogènes), un échantillon est choisi de chaque strate à l'aide d'un échantillonnage aléatoire simple.

Échantillonnage aléatoire

Exemple d'échantillonnage aléatoire simple

Choix d'un échantillon de 10 chiens parmi une population de 80 chiens d'une race donnée.

- 1ère étape : on numérote les chiens de 1 à 80
- 2éme étape : on utilise un générateur de nombres aléatoires (par exemple la fonction aléa() dans Excel)
- 3éme étape : on génère 10 nombres parmi 80

Les chiens qui correspondent à ces 10 nombres feront partie de l'échantillon

Échantillonnage aléatoire

Exemple d'échantillonnage aléatoire stratifié

Taille moyenne des étudiants ayant suivi le DU de pharmacocinétique de 1987 à 1993.

Pour des raisons de parité Homme/Femme, on choisit au hasard un échantillon de 5 filles et 5 garçons.

fille	garçon
171	193
165	187
173	180
174	185
166	178

→ La moyenne calculée : 177.2 cm

La DEP nous apprend que la population des étudiants est constituée de 99 filles et 67 garçons et que la taille moyenne des 166 étudiants est égale à 174.0 cm !!

EXEMPLE

99 filles et 67 garçons → L'échantillon n'est pas représentatif

1ère solution :

Echantillonnage aléatoire simple → 187;165;180;168;165;160;174;183;168;176

Moyenne calculée : 172,6 cm

2ème solution la mieux adaptée:

Fille : 99/166=0.596 ≈ 60%

Garçon : 67/166=0.403≈ 40%

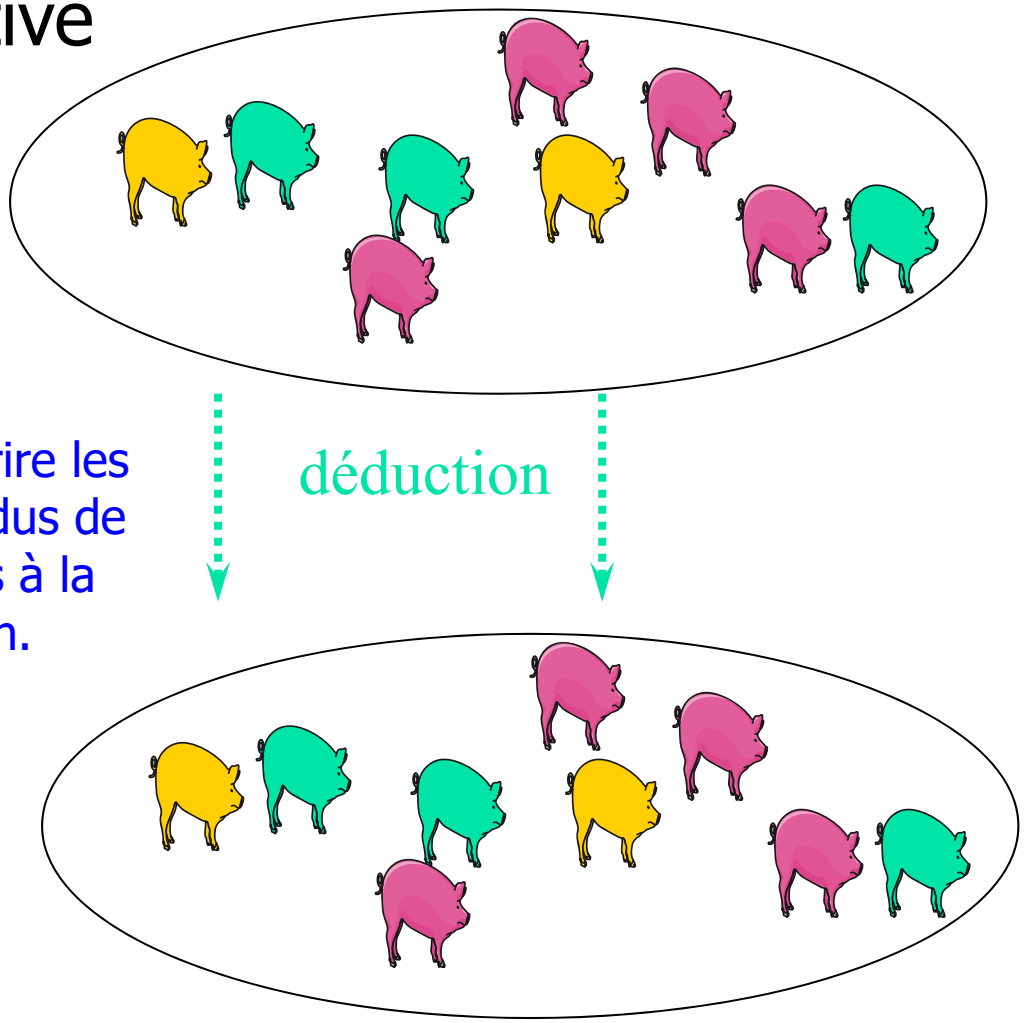
→ Echantillonnage aléatoire stratifié

→ Soit 6 filles et 4 garçons

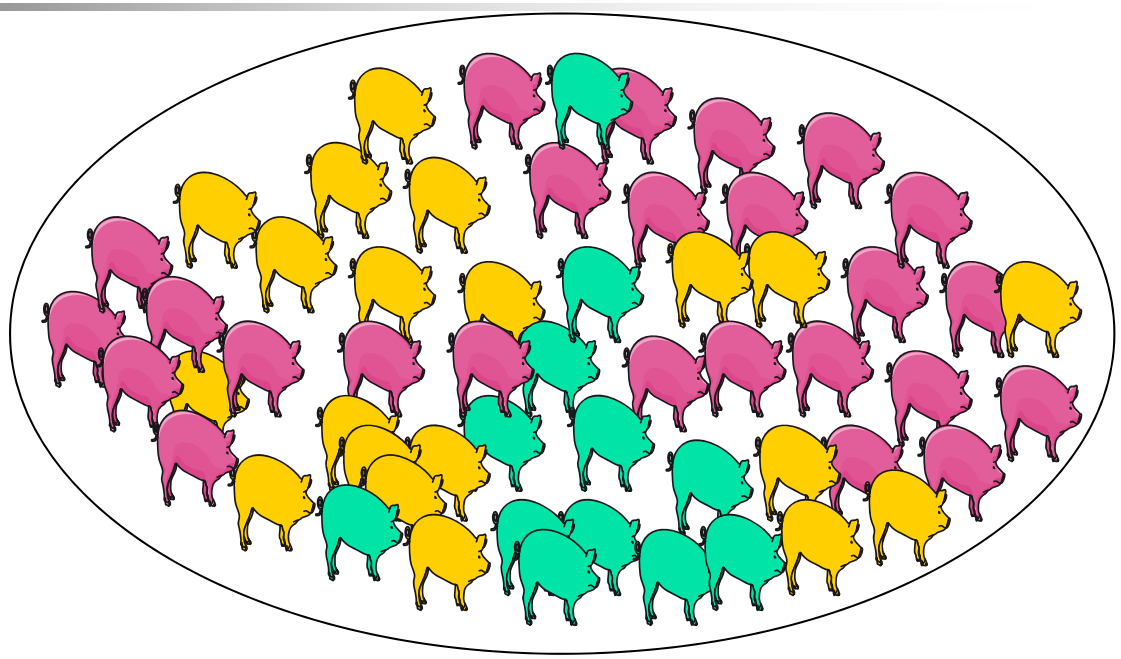
STASTIQUE DESCRIPTIVE vs. INFERENTIELLE

Statistique descriptive

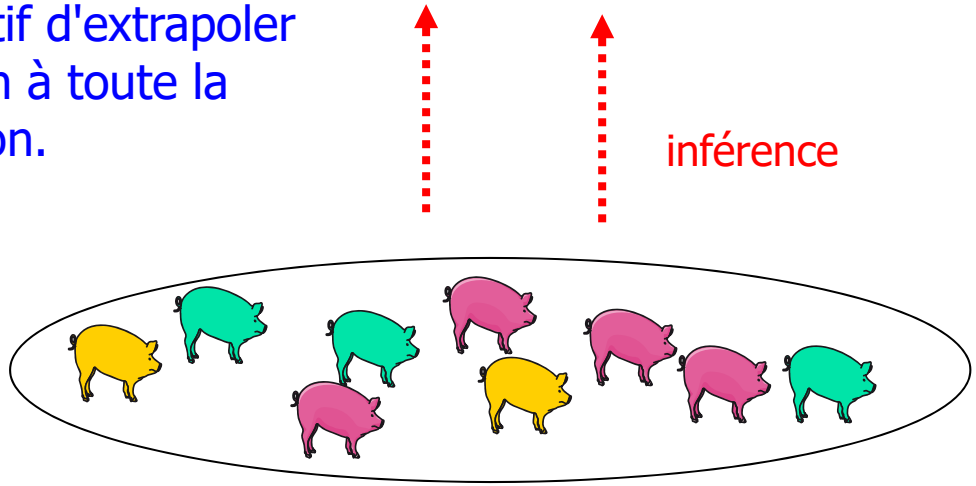
La statistique descriptive consiste à décrire les caractéristiques observées sur les individus de l'échantillon sans extrapoler les résultats à la population d'où l'on a extrait l'échantillon.



STASTIQUE DESCRIPTIVE vs. INFERENTIELLE



La statistique inférentielle a pour objectif d'extrapoler les résultats observés sur un échantillon à toute la population d'où l'on a extrait l'échantillon.



VARIABLE ALEATOIRE

✓ Une variable est une entité qui prend ses valeurs dans un ensemble fini ou infini. Cette variable est dite aléatoire si la valeur obtenue est soumise au hasard.

✓ Dans la pratique, le hasard provient du fait que la variable est mesurée sur un individu choisi au hasard dans la population.

✓ Soit Ω la population étudiée, soit ω un individu choisi au hasard dans Ω, la valeur de X mesurée sur ω est notée X(ω). En général, on note X(ω)=x et x est appelé une observation de la variable X.

On distingue deux types de variables :

- Les variables qualitatives nominales et ordinales
- Les variables quantitatives continues et discontinues

VARIABLE ALEATOIRE

Variables qualitatives :

Ce sont des variables non numériques, les valeurs qu'elles prennent sont appelées modalités.
Lorsqu'il y a un ordre dans les modalités, la variable est dite ordinale sinon elle est dite nominale

Exemples :

- Variables nominales :
l'état d'un individu :
  - M : malade,
  - G : guéri

- Type de bactéries :
1 : Staphylococcus aureus
2 : Staphylococcus spp.
3 : Streptococcus uberis

- Variables ordinales :
Score d'une douleur :
1 : très douloureux
2 : douloureux
3 : peu douloureux
4 : non douloureux

- État de la mamelle
1 : saine
2 : lésion+
3 : lésion++

VARIABLE ALEATOIRE

Variables quantitatives

Ce sont des variables numériques, la variable est dite quantitative discrète si elle prend ses valeurs dans un ensemble fini (sous ensemble de l'ensemble des entiers naturels) ou infini dénombrable (l'ensemble des entiers naturels).

Elle est dite quantitative continue si elle prend ses valeurs dans un sous-ensemle de l'ensemble des réels ou l'ensemble des réels.

Quantitatives discrètes :

- Nombre de germes dans le lait
- Nombre de porcelets dans une portée

Quantitatives continues :

- La taille
- Le poids
- Une concentration
- Une longueur

Paramètre

✓ Une caractéristique de la population supposée constante et inconnue et qui est susceptible de varier d'une population à l'autre

- Taille moyenne en France
- Pourcentage de chiens atteint d'une insuffisance rénale
- AUC moyenne

Un paramètre de population ne peut pas être calculé

Chap2. STATISTIQUE DESCRIPTIVE

- Introduction
- Graphiques
- Mesures de position
- Mesures de dispersion
- Corrélation

STATISTIQUE DESCRIPTIVE

La statistique descriptive a pour but de décrire et de synthétiser l'information contenue dans un échantillon de données à l'aide de tableaux, graphiques et indices numériques.

Si les données ne sont relatives qu'à une seule variable, on parle de statistique descriptive univariée. Dans le cas où l'on s'intéresse à deux variables simultanément, on met en œuvre la statistique descriptive bivariée.

Si l'ensemble de données provient de l'observation de plusieurs variables, on doit faire appel aux méthodes de la statistique descriptive multivariée ACP, AFC, etc.)

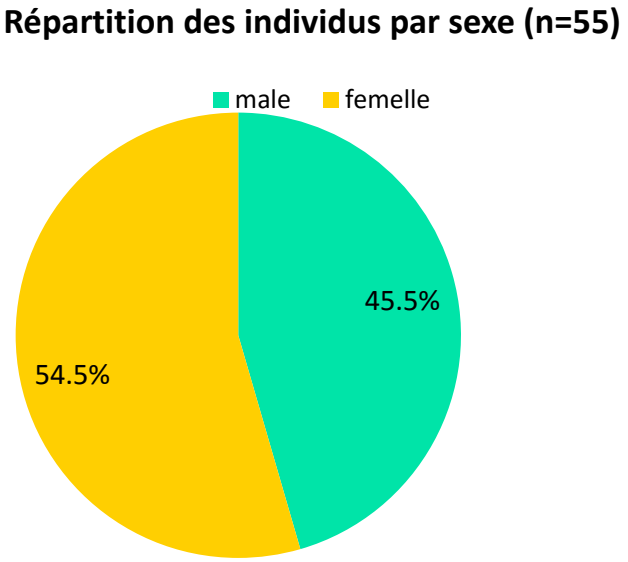
Graphiques

Variables qualitative

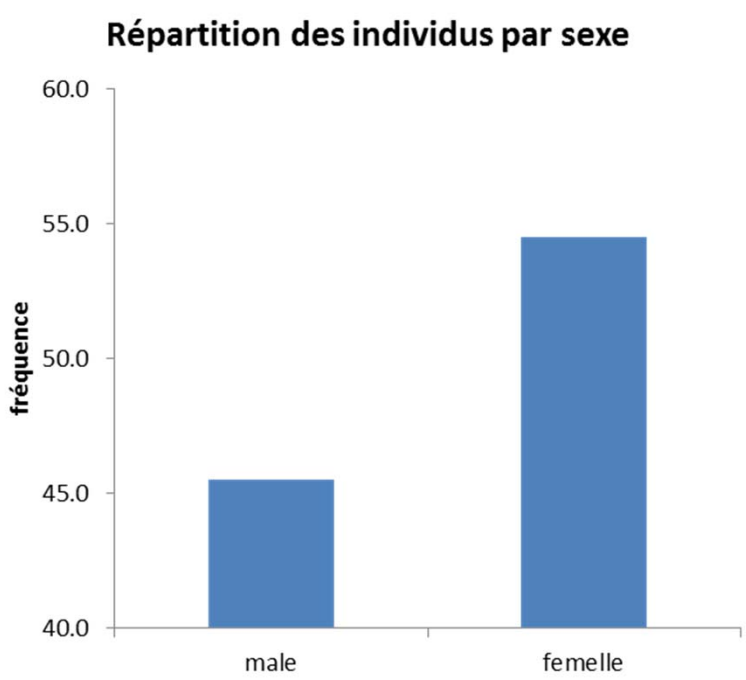
Ce type de variable peut être représentée à l'aide d'un diagramme à secteur (camembert) ou un diagramme en bâtons.

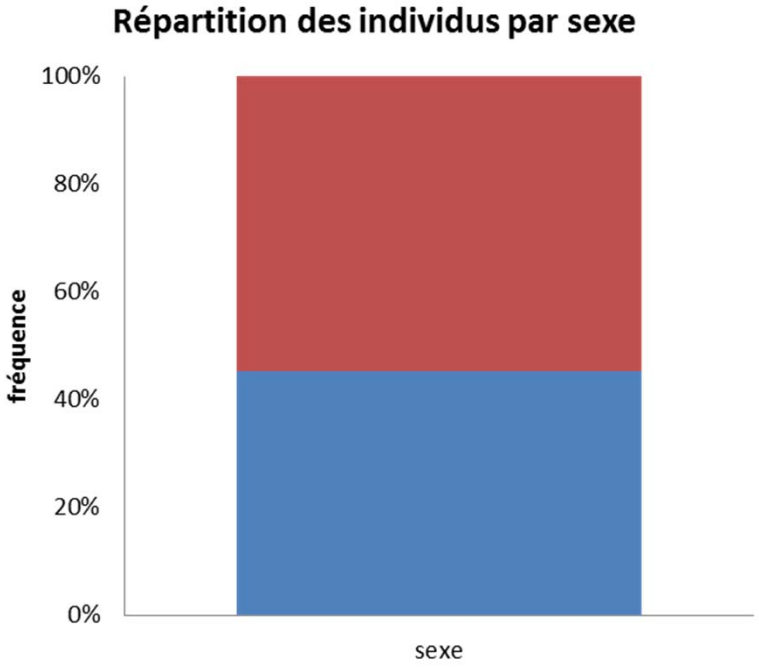
Exemple :

sexe	effectif	fréquence (%)
male	25	45.5
femelle	30	54.5
total	55	100



Graphiques


Répartition des individus par sexe
60.0
55.0
50.0
45.0
40.0
fréquence
male
femelle


Répartition des individus par sexe
100%
80%
60%
40%
20%
0%
fréquence
sexe

Graphiques

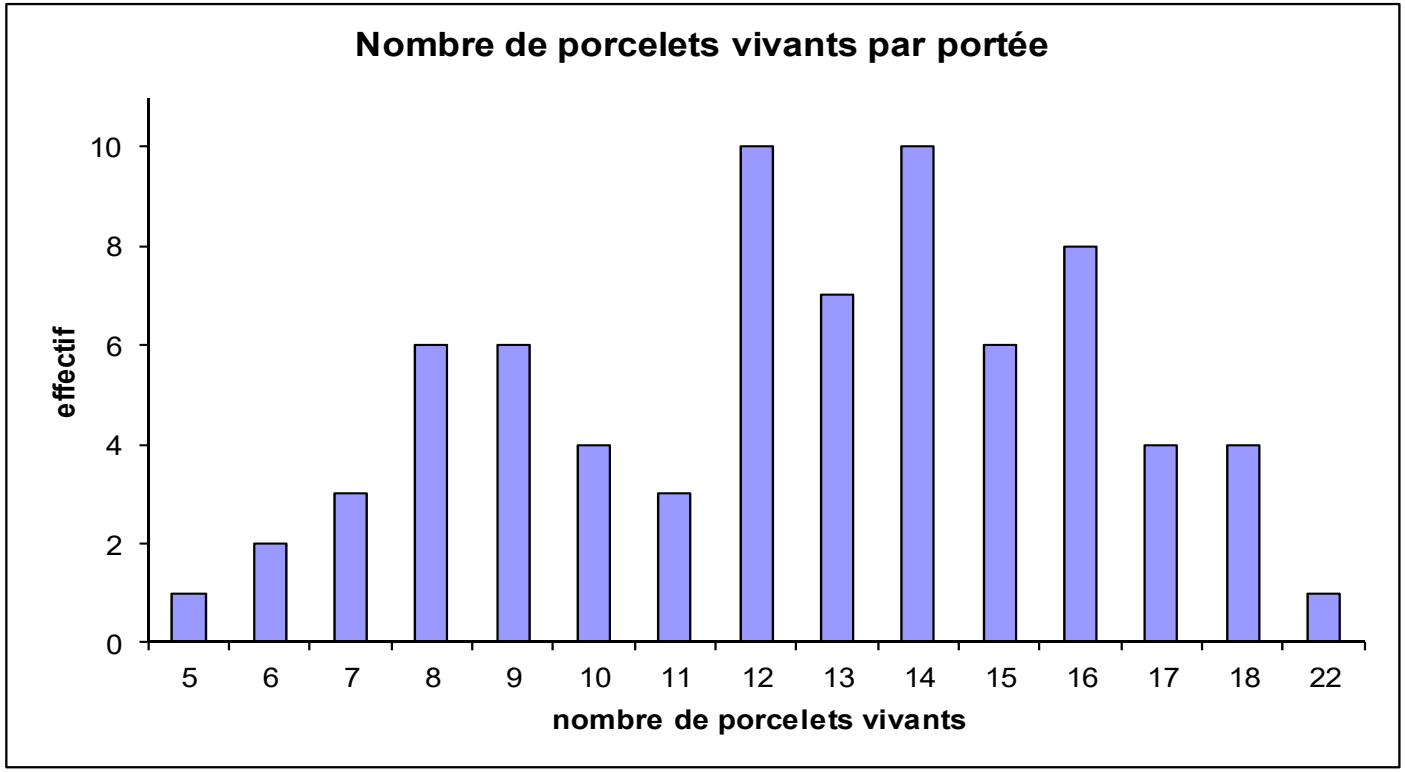
Variables quantitative discrète

Ce type de variable peut être représentée à l'aide d'un diagramme en bâtons où l'axe horizontal des abscisses porte les valeurs prises par la variable et l'axe vertical des ordonnées porte l'effectif observé.

Exemple : nombre de porcelets vivants par portée

Nombre de porcelets vivants	5	6	7	8	9	10	11	12	13	14	15	16	17	18	22
Nombre de portées	1	2	3	6	6	4	3	10	7	10	6	8	4	4	1

Graphiques



Dans un diagramme en bâtons , chaque bâton est proportionnel à l'effectif observé

Graphiques

Variables quantitative continue

Dans le cas de ces variables, il y a autant de valeurs différentes que d'individus dans l'échantillon par conséquent la représentation graphique à l'aide d'un diagramme en bâtons dans ce cas n'a aucun sens.

Exemple : poids des porcelets à la naissance (kg)

1.83	1.72	1.70	2.15	2.05	1.76	2.18	1.61	1.43	1.44	1.94
1.63	1.40	0.93	0.88	0.97	0.69	1.16	0.83	1.50	1.35	0.99
0.94	1.07	1.08	1.23	1.05	1.23	1.30	1.27	1.07	1.49	1.29
1.83	1.72	1.70	2.15	2.05	1.76	2.18	1.61	1.43	1.44	1.94
1.63	1.40	0.93	0.88	0.97	0.69	1.16	0.83	1.50	1.35	0.99
0.94	1.07	1.08	1.23	1.05	1.23	1.30	1.27	1.07	1.49	1.29
1.05	1.23	1.30	1.27	1.07	1.49	1.29	1.32	0.81	0.63	1.19
1.03	1.58	1.16	0.72	1.60	1.63	0.94	1.36	0.75	1.57	1.65
1.42	1.45	0.87	1.12	1.38	1.48	0.56	0.28	0.16	0.29	1.18
1.26	1.89	1.35	1.30	1.57	0.19	1.39	1.59	1.54	1.29	1.93
0.84	1.93	1.22	1.13	1.44	0.89	1.04	1.05	1.17	.....	.....

Graphiques

Une variable quantitative continue doit être représentée à l'aide d'un histogramme

Méthodologie :

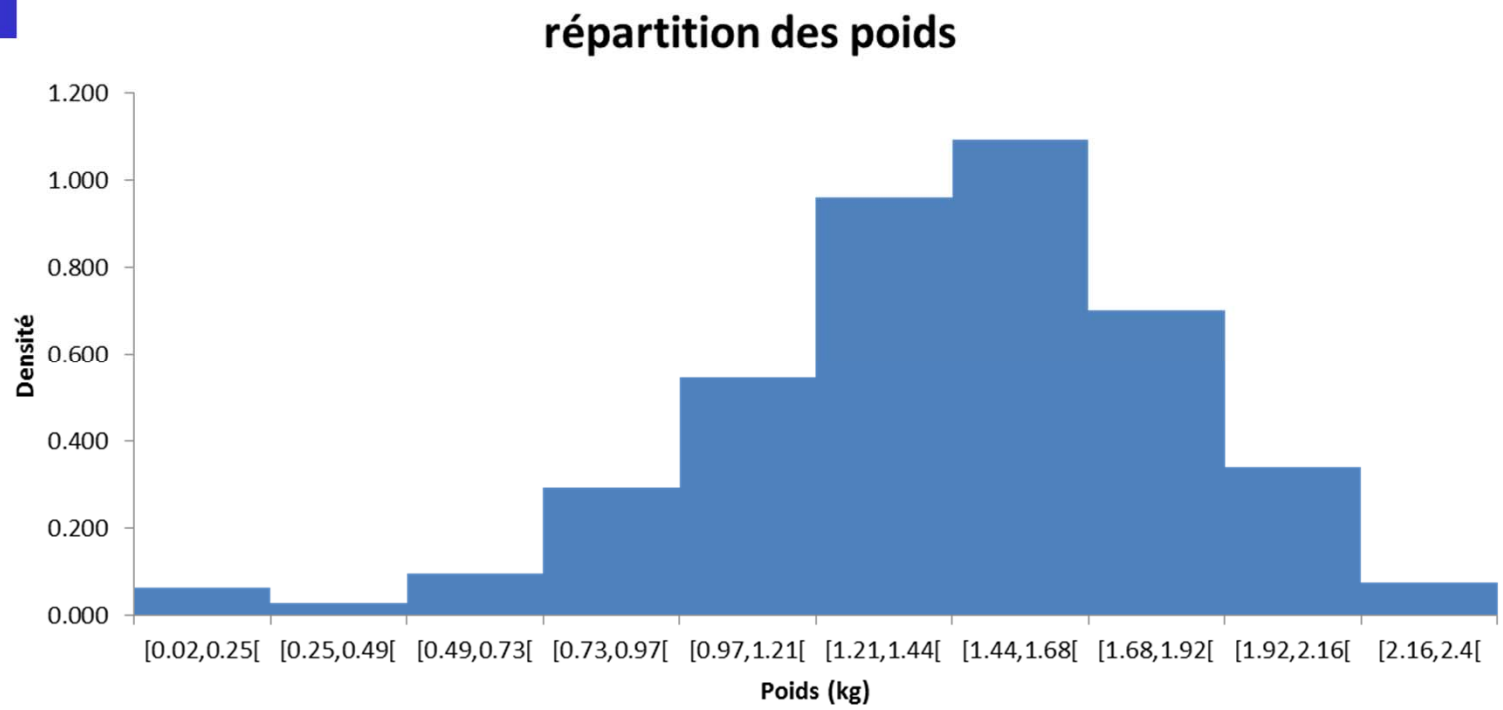
- Regrouper les données en classe
- Calculer le nombre d'observations de l'échantillon contenues dans chacune des classes appelé effectif de la classe ainsi que la fréquence (rapport entre l'effectif de la classe et le nombre total d'observations)
- Calculer l'amplitude de la classe (différence entre la borne supérieure et la borne inférieure de la classe)
- Calculer la densité de la classe (rapport entre la fréquence et l'amplitude de la classe)
- On trace l'histogramme représenté par une succession de rectangles accolés, la base du rectangle est égale à l'amplitude de la classe et la hauteur du rectangle est égale à la densité de la classe

Graphiques

Exemple : Poids de porcelets à la naissance (kg)

Borne inférieure	Borne supérieure	Amplitude	Effectif	Fréquence	Densité de la classe
0.020	0.258	0.238	14	0.015	0.0630
0.258	0.496	0.238	7	0.007	0.0294
0.496	0.734	0.238	22	0.023	0.0966
0.734	0.972	0.238	66	0.070	0.2941
0.972	1.210	0.238	123	0.130	0.5462
1.210	1.448	0.238	216	0.229	0.9622
1.448	1.686	0.238	245	0.260	1.0924
1.686	1.924	0.238	158	0.167	0.7017
1.924	2.162	0.238	76	0.081	0.3403
2.162	2.400	0.238	17	0.018	0.0756

Graphiques



- Dans un histogramme la surface d'un rectangle est proportionnelle à la fréquence de la classe
- La somme des aires des rectangles est égale à 1

Graphiques

En conclusion :

- Faire un graphique adapté à la nature de la variable mesurée
- Faire un graphique sans déformer le message et l'information contenus dans les données
- Tracer les points avec l'échelle adaptée au problème étudié et qui soit la moins suggestive possible

La représentation graphique est une étape très importante dans l'analyse des données, et constitue un des outils d'aide à la décision pour répondre à des questions qui concernent la population d'où l'on a extrait l'échantillon.

Indices numériques

Les indices numériques permettent de synthétiser l'information contenue dans l'échantillon

On distingue deux types d'indices : les indices de position (ou de la tendance centrale) et les indices de dispersion.

Les principaux indices de position sont :

- la moyenne
- les 3 quartiles.

Les principaux indices de dispersion sont :

- la variance
- l'écart-type
- le coefficient de variation

Indices de position

Indices de position (de la tendance centrale)

La moyenne arithmétique

Soit $(x_1,...,x_n)$ un échanillon de taille n.

x_i désigne l'observation faite sur l'individu i

la moyenne arithmétique notée $\bar{x}$ est définie par

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Formule simple et explicite
- Sensible aux « grandes valeurs »

Indices de position

Les trois quartiles

- Le premier quartile (noté q1) est la valeur dont au moins 25% des observations sont inférieures à q1
- Le deuxième quartile noté q2 est la valeur dont 50% des observations sont inférieures à q2
- Le troisième quartile noté q3 est la valeur dont 75% des observations sont inférieures à q3

- Le 2ème quartile est appelé aussi médiane
- La médiane est égale à la moyenne quand la distribution des données est symétrique (par ex. la distribution normale)
- La médiane est plus robuste que la moyenne (insensible aux grandes valeurs)

Indices de position

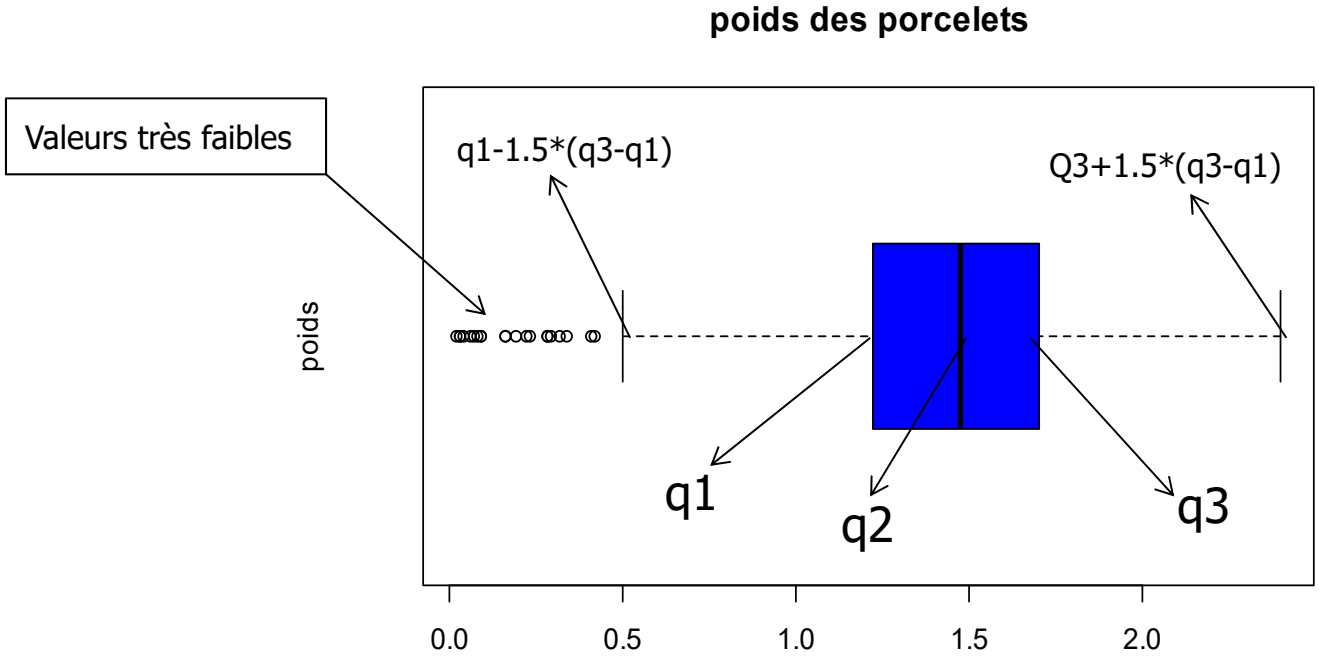
Intérêt des 3 quartiles : boîte à moustaches (boxplot)

Exemple : poids des porcelets à la naissance

q1=1.22

q2=1.47

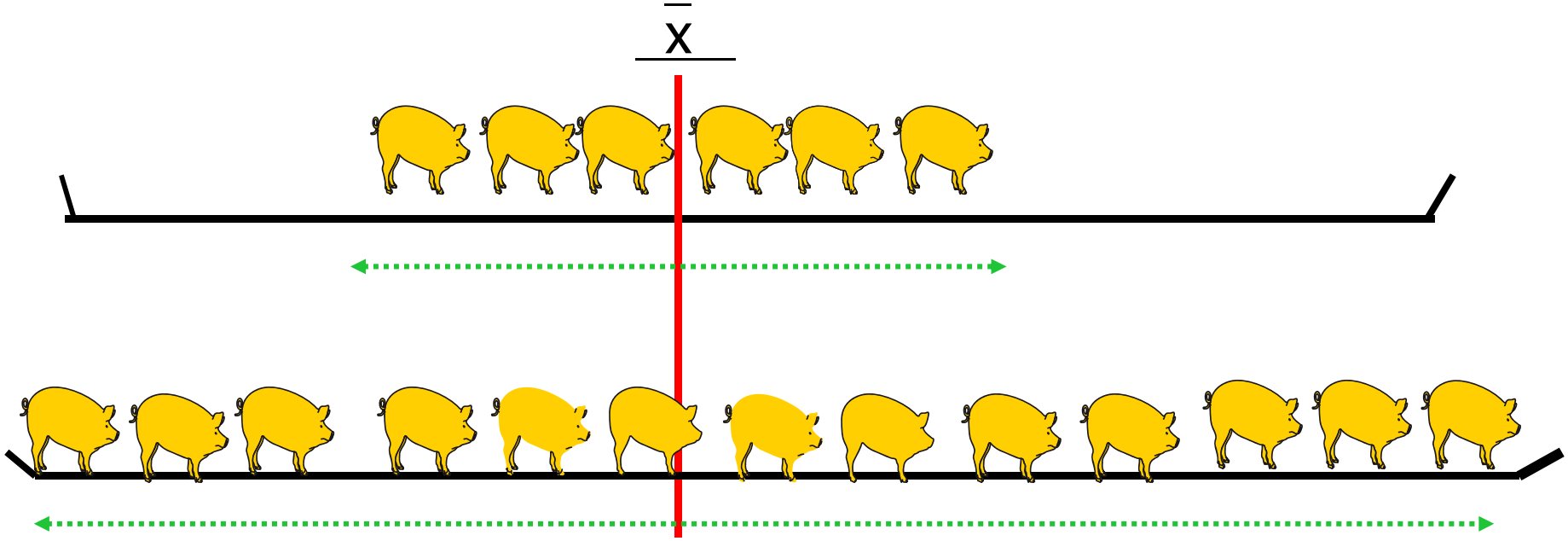
q3=1.7



- Le boxplot donne une information sur la symétrie ou l'asymétrie d'une distribution
- Le boxplot permet de détecter les valeurs très petites ou très grandes et donc qui sont susceptibles d'être considérées comme aberrantes

La tendance centrale ne suffit pas !!

Pourquoi la moyenne ne suffit pas à caractériser un échantillon ?



Une même valeur de la moyenne peut correspondre à plusieurs situations au niveau de l'homogénéité des valeurs.

D'où l'intérêt de définir un autre indice pour quantifier la dispersion des valeurs autour de leur moyenne (ou de façon équivalente la variabilité entre les individus)

INDICES DE DISPERSION

Indices de dispersion

La variance (notée s^2_{n-1}) est un indice qui permet de quantifier la dispersion des observations de l'échantillon autour de leur moyenne, elle est définie par :

$$s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

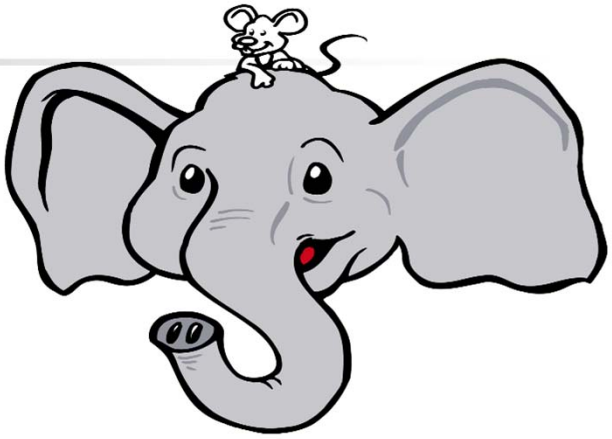
- Ecart-type :

L'écart-type (noté sd) est également un indice de dispersion défini comme étant la racine carrée de la variance, soit :

$$sd = \sqrt{s^2_{n-1}}$$

- La notation sd signifie standard déviation (notation universelle dans les logiciels et les publications)
- L'écart-type est exprimé dans la même échelle que les observations : si x_i est exprimé en kg, la variance sera en kg^2 et sd en kg.

INDICES DE DISPERSION



Coefficient de variation

Le coefficient de variation est un indice qui quantifie la dispersion des observations de l'échantillon en pourcentage de la moyenne, il est calculé à l'aide de la formule suivante :

$$CV\% = \frac{sd}{\bar{x}} \times 100$$

- Le coefficient de variation est une dispersion relative
- Le coefficient de variation est sans unité, ce qui permet la comparaison de deux échantillons dont l'un présente des valeurs très élevées et l'autre des valeurs très petites.

INDICES DE DISPERSION

Exemple :

- Danseuses : 50.6kg ± 5.2
- Sumos : 200.9kg ± 10.8

Question : Y a-t-il une variabilité plus grande chez les sumos ?

Réponse :

- Danseuses : 10.3%
- Sumos : 5.4%

Corrélation

Dans cette partie nous allons nous placer dans le cadre de la statistique descriptive bivariée en d'autres termes, nous allons nous intéresser à deux variables quantitatives simultanément.

On considère un échantillon de n individus

Sur chaque individu i (i=1,…,n) on observe un couple d'observations (xi, yi).

x_1	x_2	.....	x_{n-1}	x_n
y_1	y_2	.....	y_{n-1}	y_n

- On suppose que les caractères mesurés sont des variables **quantitatives**

Corrélation

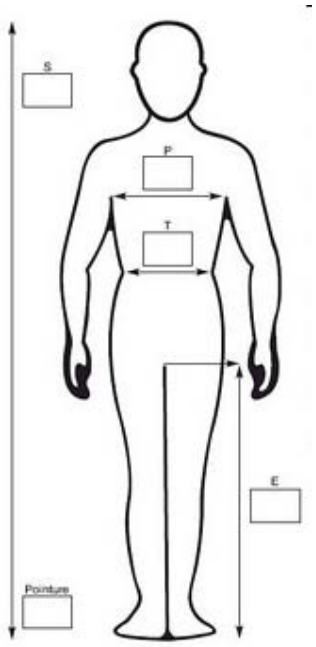
Exemple

X : taille (cm)

Y : poids (kg)

n =10

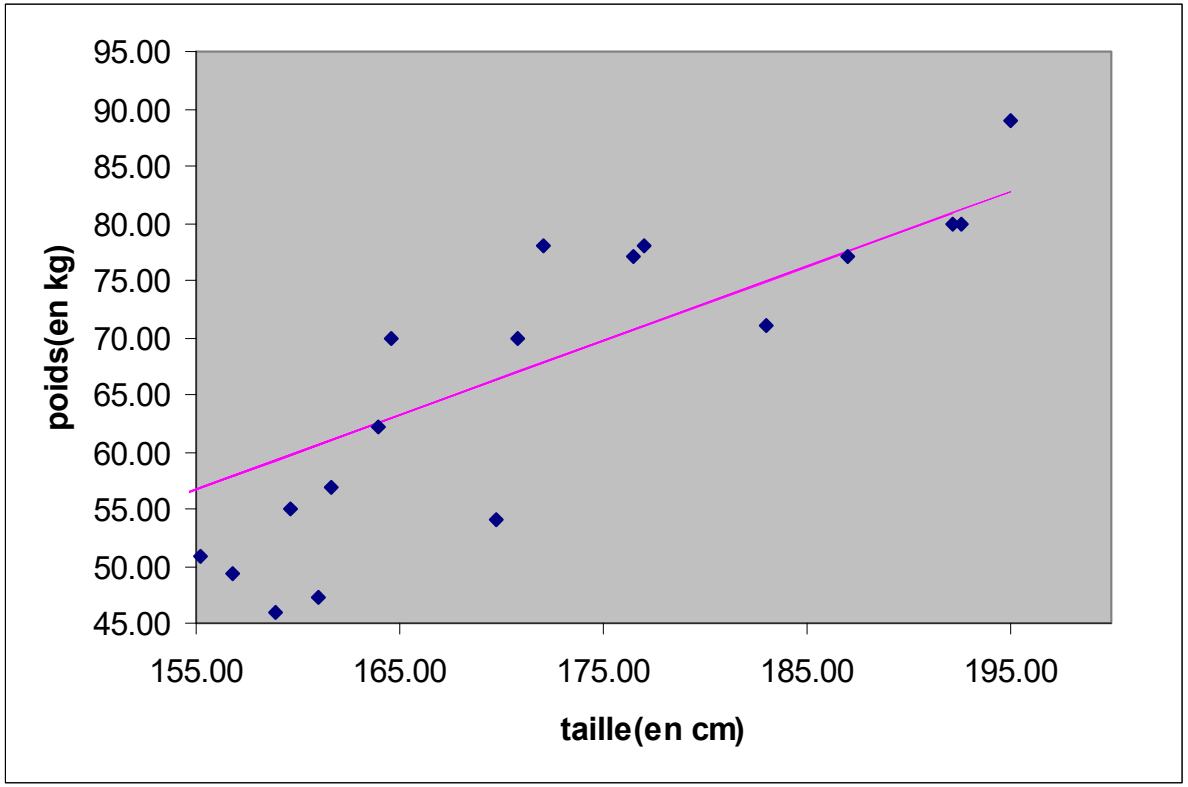
T tour de taille	Taille Française	P Tour de Poitrine	Taille Groupée	S + E Stature + Entrejambe	Taille International
69 à 72 73 à 76	36 38	78 à 85	0	156 à 162 (76 - 78)	XS (TPT)
77 à 80 81 à 84	40 42	86 à 93	1	163 à 167 (76 - 78)	S (PT)
85 à 88 89 à 92	44 46	94 à 101	2	168 à 175 (76 - 78)	M (TM)
93 à 96 97 à 100	48 50	102 à 109	3	174 à 182 (76 - 78)	L (GT)
101 à 104 105 à 108	52 54	110 à 117	4	180 à 188 (76 - 78)	XL (XG)
109 à 112 113 à 116	56 58	118 à 125	5	186 à 194 (76 - 78)	XXL (XX)
117 à 120 121 à 124	60 62	126 à 133	6	192 à 194 (76 - 78)	XXXL (XXX)



X	174	172	160	173	162	174	178	162	150	170
Y	70	57	50	70	62	67	78	57	77	63

Question : est-ce que plus l'on est grand plus l'on a tendance à peser plus?

Corrélation



- On constate qu'il y a un lien linéaire entre la taille et le poids
- On constate aussi qu'en moyenne plus on est grand plus on a tendance à peser plus !

Corrélation

Définition 1 : corrélation

On dit qu'il y a une corrélation linéaire entre deux variables quantitatives X et Y si en moyenne Y en fonction linéaire de X

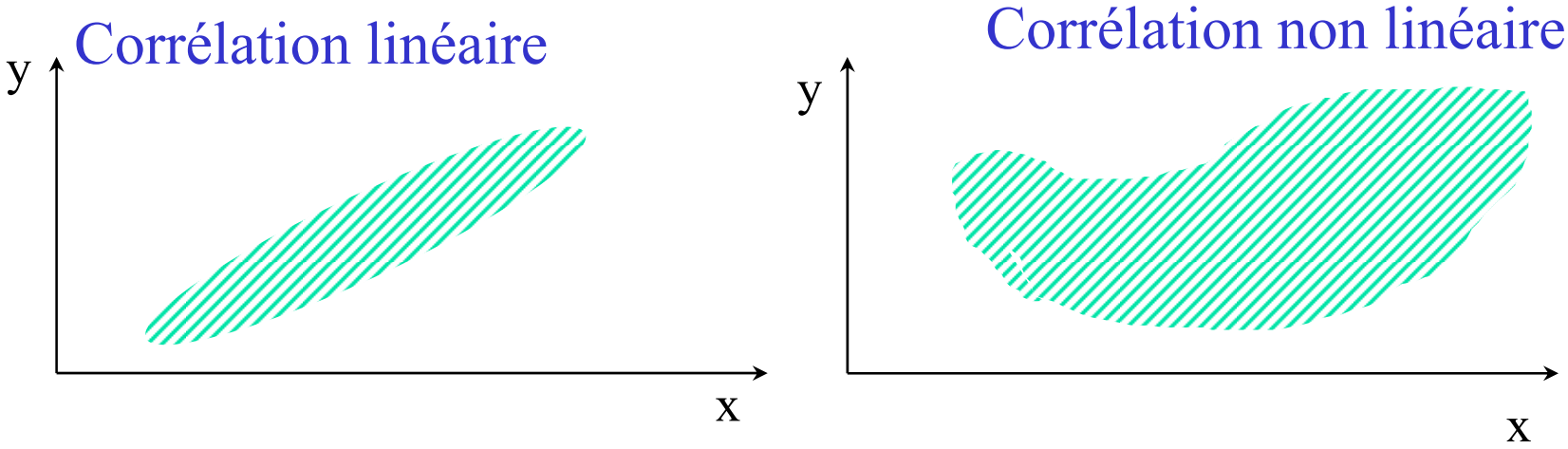
Définition 2 : Coefficient de corrélation

Le coefficient de corrélation (noté r) est un indice qui quantifie l'intensité du lien linéaire entre les deux variables quantitatives

- Le coefficient de corrélation est appelé aussi coefficient de Pearson
- Dans Excel : coefficient.corrélation(X,Y)
- Dans R : cor(X,Y)
- r est toujours compris entre -1 et 1

- Si r ≈ +1, les variables sont corrélées positivement
- Si r ≈ -1, les variables sont corrélées négativement
- Si r ≈ 0, les variables ne sont pas corrélées

Corrélation



- Le coefficient de corrélation r ne quantifie que les corrélations linéaires
- r proche de 0 signifie absence de corrélation linéaire et non pas absence de corrélation, car rien n'empêche que X et Y soient corrélées non linéairement.

- D'où l'importance de faire le graphique avant de calculer un coefficient de corrélation

Chap3. Estimation d'un paramètre

- Moyenne
- Variance, écart-type
- Pourcentage
- Propriétés des estimateurs
- Intervalles de confiance d'une moyenne et d'un pourcentage

Introduction

Exemple 1 : Pression artérielle diastolique

On a mesuré la pression artérielle sur 10 individus :

i	1	2	3	4	5	6	7	8	9	10
Pression (mmHg)	87	92	55	120	100	70	85	65	72	83

- Quelle est la pression artérielle moyenne de l'échantillon ? $\bar{x}$ = **82.90** mmHg
- Que représente cette valeur par rapport à la pression moyenne (m) de la population totale d'où l'on a prélevé ces 10 individus ?
- Peut-on dire 82.90 est une estimation (approximation) de m ? Si oui, quelle est la précision de cette estimation ?

Introduction

Exemple 2 : Pourcentage de guéris

Dans le cadre de l'étude de l'efficacité d'un traitement, on a observé 45 guéris parmi 50 traités.

1 : guéri
0 : malade

Etat	Effectif (n_{obs})	Fréquence (p_{obs})
1	45	90%
0	5	10%
total	50	100%

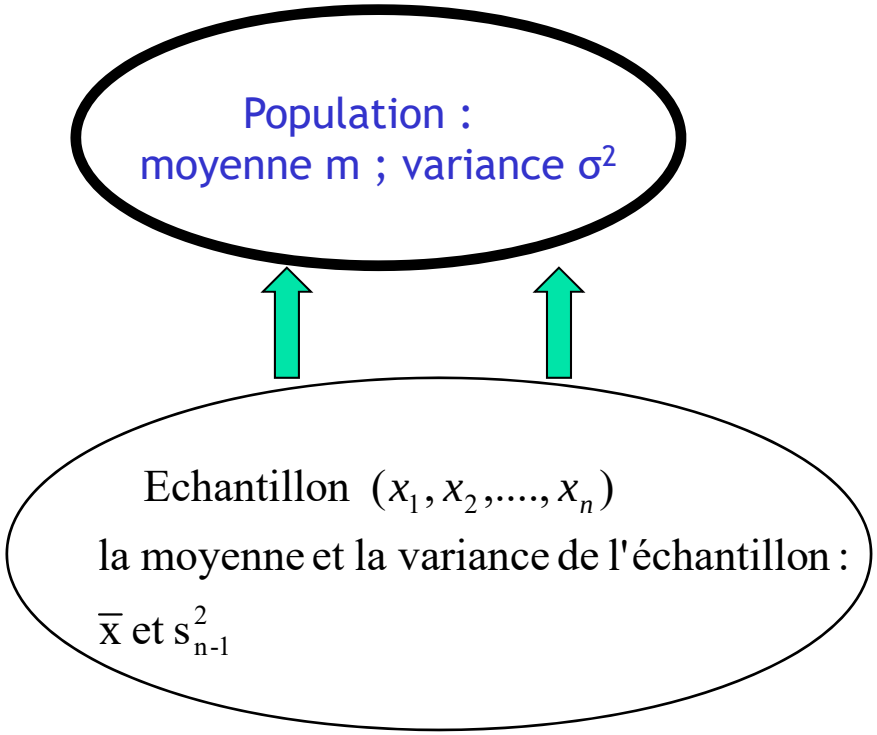
obs : observé

- Calculer le pourcentage de guéris parmi les 50 traités ? $P_{obs} = n_{obs}/n = 45/50 = 90\%$
- Que représente cette valeur par rapport au pourcentage (p) de guéris que l'on obtiendrait si l'on traitait toute la population ?
- Peut-on dire 90% est une estimation (approximation) de m ? Si oui, quelle est la précision de cette estimation ?

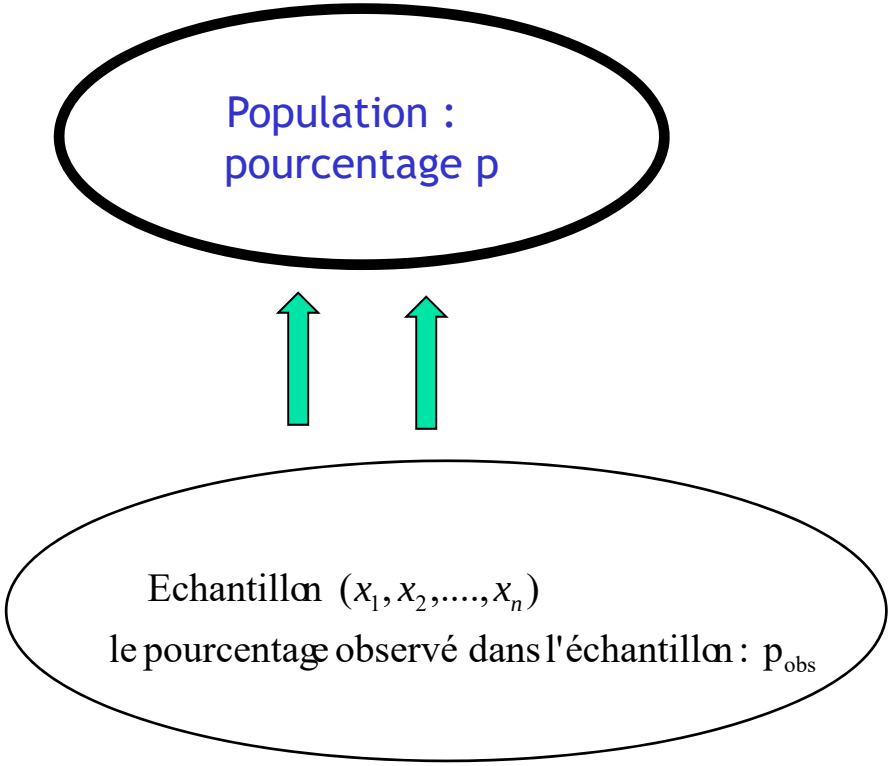
La réponse à ces questions est fournie par la statistique inférentielle

Introduction

Variable quantitative



Variable qualitative



La statistique inférentielle permet d'établir le lien entre les paramètres de la population (inconnus) et les paramètres calculés de l'échantillon (donc connus)

ESTIMATION D' UNE MOYENNE

Soit $(x_1,x_2,....,x_n)$ un échantillon de taille n où x_i représente l'observation d'une variable quantitative faite sur l'individu i (i=1,…,n)

• On appelle estimation de m, la moyenne de l'échantillon calculée comme suit :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\overline{x}$ est fonction de l'échantillon, sa valeur varie d'un échantillon à l'autre, par conséquent $\overline{x}$ peut-être considéré comme une observation d'une variable aléatoire, notée $\overline{X}$

Remarque : En général, $\bar{x}$ est noté $\widehat{m}$ pour signifier que m est estimé

Estimation de la variance

Soit $(x_1,x_2,....,x_n)$ un échantillon de taille n, x_i représente l'observation d'une variable quantitative faite sur l'individu i (i=1,...,n)

- On appelle estimation de σ^2, la variance de l'échantillon calculée comme suit :

$$s^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

De la même façon que pour la moyenne, on peut parler d'estimateur de la variance noté S^2_{n-1}

Remarque : En général, S^2_{n-1} est noté $\widehat{\sigma}^2$

- L'écart-type σ est estimé par : s_{n-1} , sd ou $\hat{\sigma}$

ESTIMATION D' UN POURCENTAGE

Soit $(x_1,x_2,....,x_n)$ un échantillon de taille n, où x_i représente l'observation d'une variable qualitative mesurée sur l'individu i (i=1,…,n) :

$$x_i = \begin{cases} 1 & \text{si le caractère est observé chez l'individu i} \\ 0 & \text{sinon} \end{cases}$$

Soit n_{obs} le nombre (l'effectif) d'individus sur lesquels le caractère a été observé.

On appelle estimation du pourcentage p, le pourcentage de l'échantillon calculé comme suit :

$$p_{obs} = \frac{n_{obs}}{n}$$

Remarque : En général, p_{obs} est noté $\hat{p}$

De la même façon que pour la moyenne, on peut parler d'estimateur de p noté $\hat{P}$

Propriétés des estimateurs

❑ Cas de la moyenne :

- Pour n assez grand, on montre que :

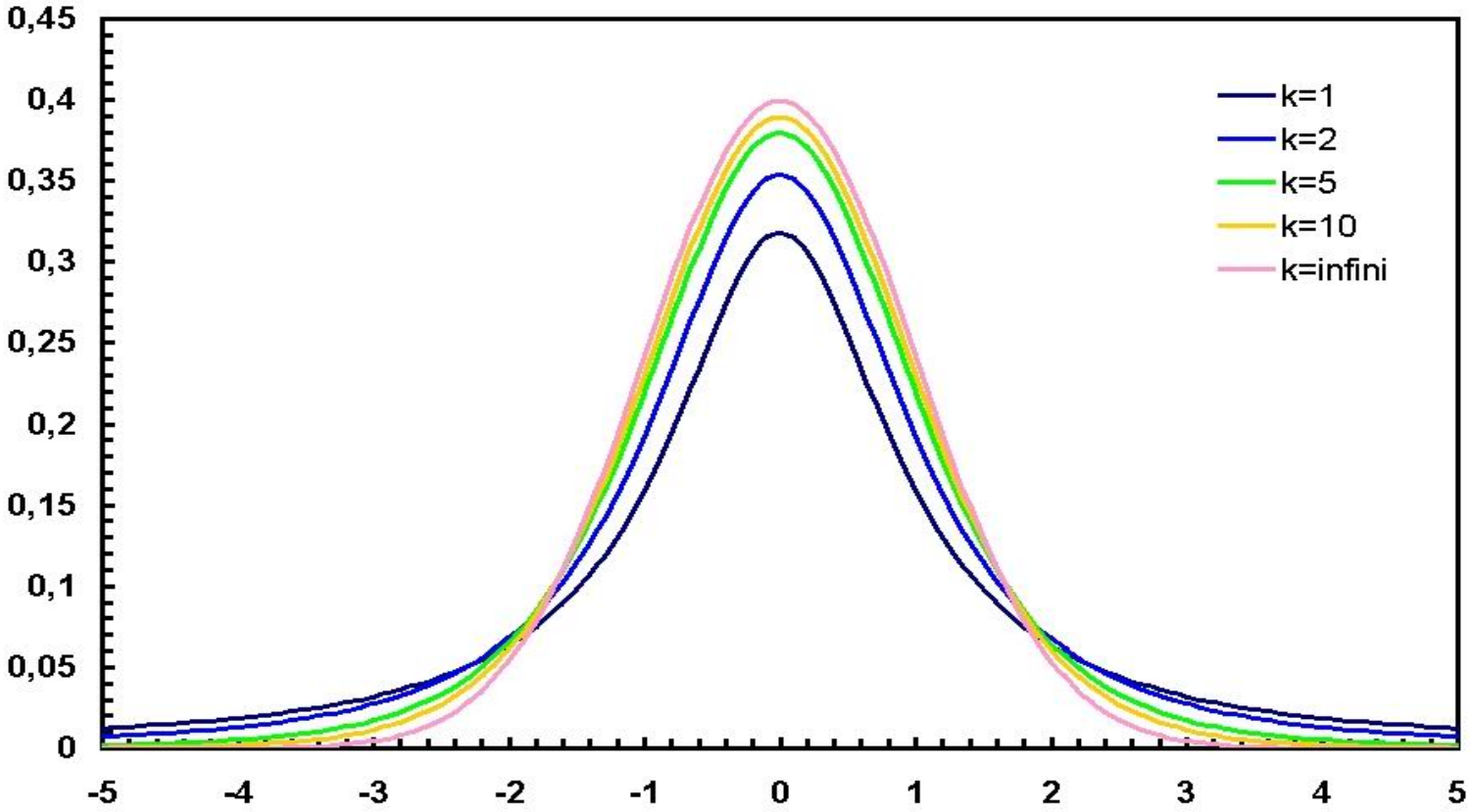
$$T_{obs} = \frac{\overline{X} - m}{\frac{s_{n-1}}{\sqrt{n}}} \text{ suit une loi noramle } N(0,1)$$

Ce résultat est connu sous le nom de théorème 'central limit'

- Pour n petit, on montre que si l'échantillon provient d'une loi normale alors T_{obs} suit une loi de Student à k degrés de liberté, dans ce cas k=n-1

LOI DE STUDENT

Allure de la loi de Student pour différents degrés de liberté



Propriétés des estimateurs

- Cas de la variance :

$$\text{On montre que } (n-1)\frac{S_{n-1}^2}{\sigma^2} \text{ suit une loi de probabilités que l'on appelle}$$

$$\text{loi du Khi - deux (notée } \chi^2\text{) à k degré de libertés, dans cas k = n - 1}$$

- Cas du pourcentage

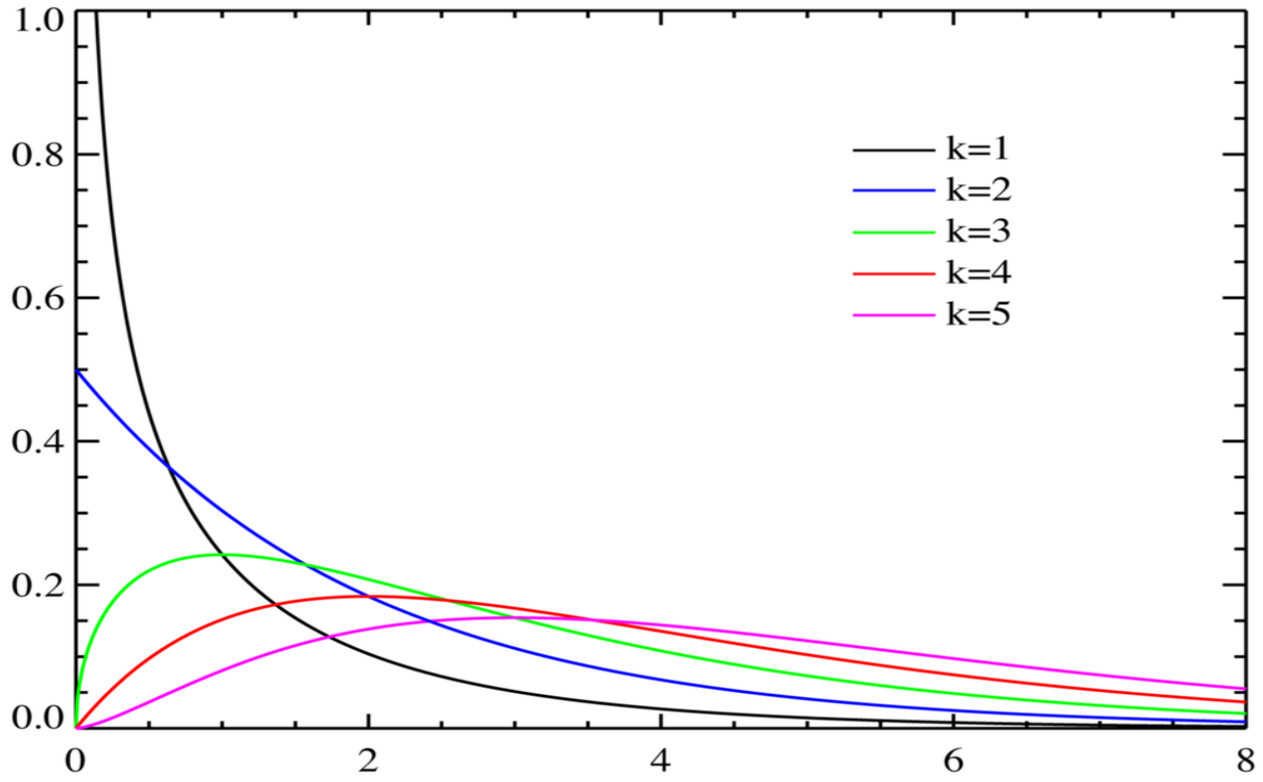
Pour n assez grand, on montre que l'estimateur de p suit une loi normale N(p, p(1-p)/n)

Remarque 1 : Ce résultat est une conséquence du théorème 'central limit'

Remarque 2 : Ce dernier résultat suppose aussi que p n'est pas très proche de 0 (autrement dit l'événement étudié n'est pas un événement rare)

LOI DU KHI-DEUX

Allure de la loi du Khi-deux pour différents degrés de liberté



INTERVALLE DE CONFIANCE

Définition :
Un intervalle de **confiance** est un intervalle construit à partir d'un échantillon et qui est susceptible de contenir la vraie valeur du paramètre estimé avec une certaine probabilité appelée **sécurité**
Dans la pratique la sécurité est notée 1-α.
En biologie la sécurité utilisée est de 95%

Autre interprétations :

- La donnée d'un intervalle de confiance de sécurité 95% d'un paramètre signifie que cet intervalle a 95% de chances de contenir la valeur du paramètre que l'on aurait obtenue si l'ensemble de la population a participé à l'expérience.

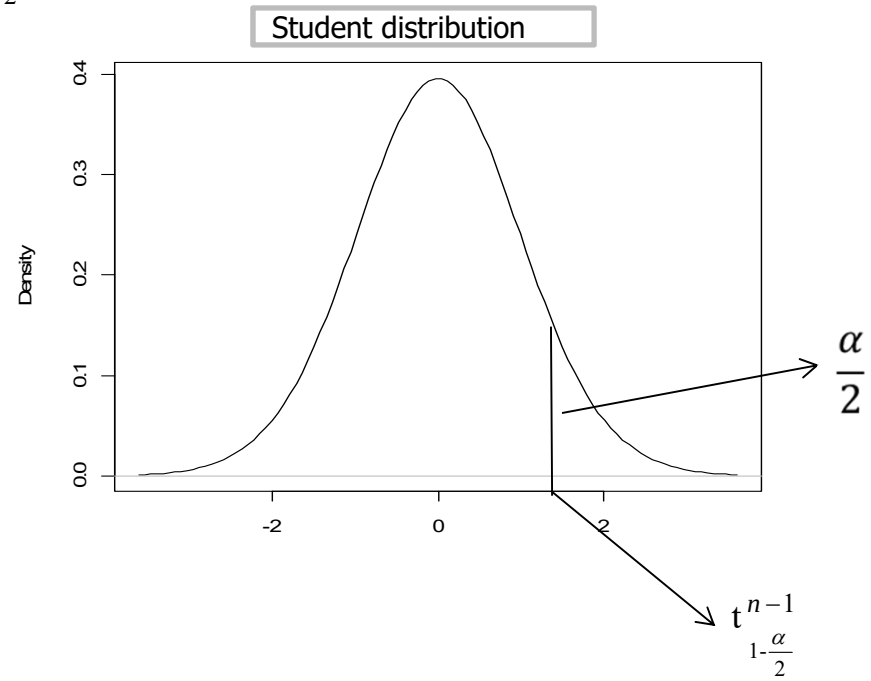
- Si on répétait la même expérience un certain nombre de fois assez élevé et si à l'issue de chaque expérience on calculait l'intervalle de confiance de sécurité 95% d'un paramètre, cela voudrait dire que 95% des intervalles de confiance calculés vont contenir la vraie valeur du paramètre estimé.

INTERVALLE DE CONFIANCE DE LA MOYENNE

Un intervalle de confiance de la moyenne m de sécurité $1-\alpha$ est calculé comme suit :

$$m \in \left[\bar{x} - t^{n-1}_{1-\frac{\alpha}{2}} \frac{sd}{\sqrt{n}} ; \bar{x} + t^{n-1}_{1-\frac{\alpha}{2}} \frac{sd}{\sqrt{n}} \right]$$

Où $t^{n-1}_{1-\frac{\alpha}{2}}$ est appelé quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à (n-1) degrés de liberté (ddl)



INTERVALLE DE CONFIANCE DE LA MOYENNE

Conséquence : Précision de l'estimation de m

La précision de l'estimation de m est donnée par le terme :

$$t_{1-\frac{\alpha}{2}}^{n-1} \frac{sd}{\sqrt{n}}$$

Et on écrit : • $m = \overline{x} \pm \underbrace{t_{1-\frac{\alpha}{2}}^{n-1} \frac{sd}{\sqrt{n}}}_{\text{précision}}$

$\frac{sd}{\sqrt{n}}$ est appelé « standard error» et on le note se

Si n est sassez grand et α =0.05, dans ce cas : $t_{1-\frac{\alpha}{2}}^{n-1} \approx 2$

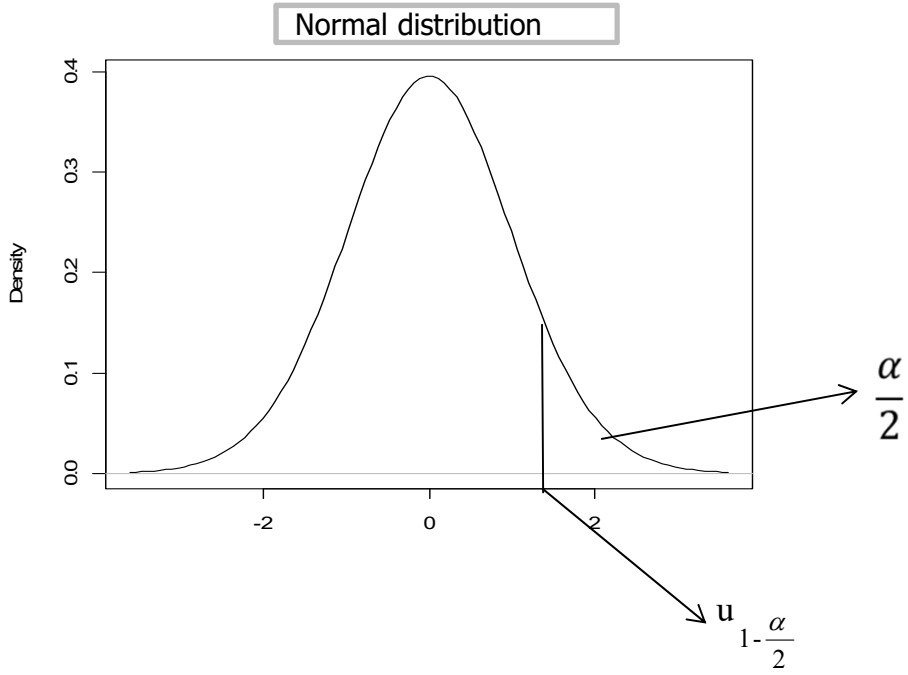
L'intervalle de confiance s'écrit : $\overline{x} \pm 2 \frac{sd}{\sqrt{n}}$

Intervalle de confiance d'un pourcentage

Un intervalle de confiance de sécurité 1-α d'un pourcentage est calculé comme suit :

$$p \in \left[\hat{p} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \; ; \hat{p} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

Où $u_{1-\frac{\alpha}{2}}$ est appelé quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale N(0,1)



Intervalle de confiance d'un pourcentage

La précision de l'estimation est donnée par le terme :

$$u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1 - \hat{p})}{n}}$$

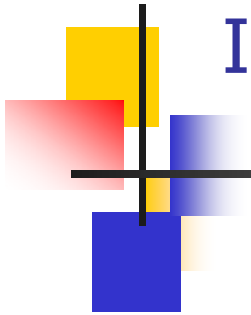
Et on écrit : • $p = \hat{p} \pm \underbrace{u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}}_{\text{précision}}$

Si α =0.05, dans ce cas : $u_{1-\frac{\alpha}{2}} \approx 2$

L'intervalle de confiance de p s'écrit : $\hat{p} \pm 2 \sqrt{\frac{\hat{p}\,(1 - \hat{p})}{n}}$

Chap4. TESTS d'HYPOTHESES

1. Introduction
2. Hypothèses
3. Risques
4. Règle de décision
5. Tests usuels



INTRODUCTION

Exemple 1 : On a dénombré sur 4900 naissances 2500 garçons (51%)

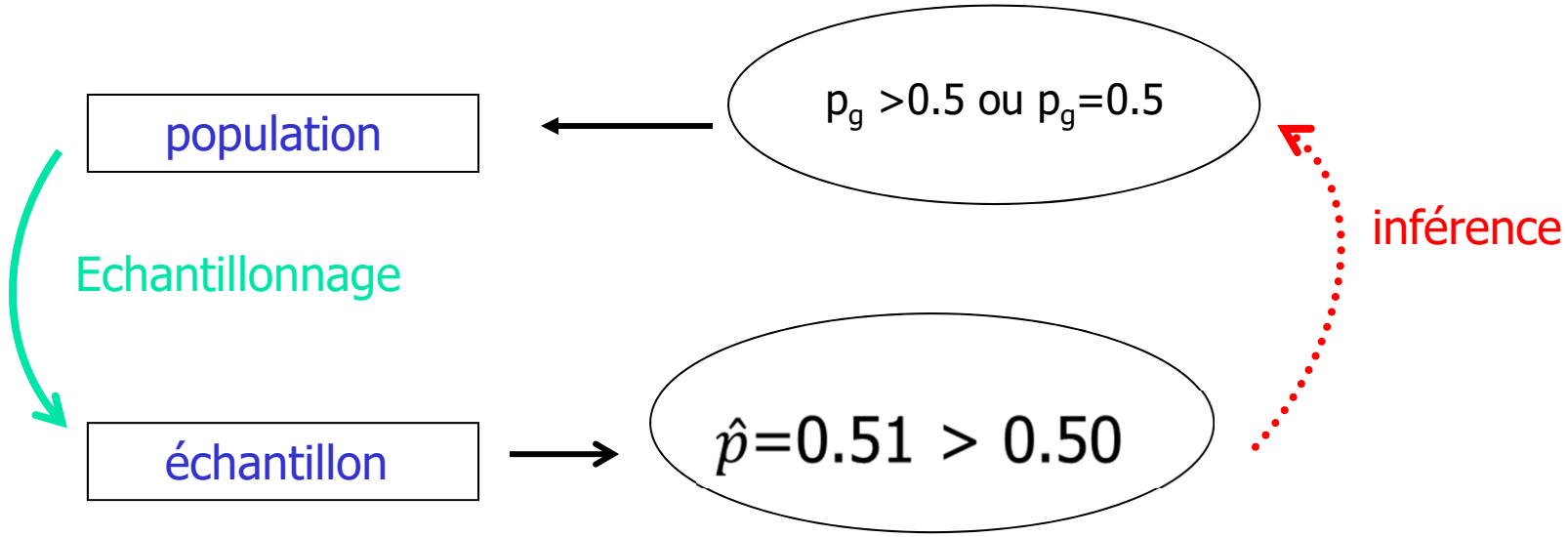
Ce résultat est-il compatible avec l'hypothèse d'équiprobabilité des naissances des garçons et des filles ?

Exemple 2 : Les guérisons d'une certaine maladie avec un traitement de référence et un traitement A ont été :

- traitement A : 85 guérisons sur 100 traités (85%)
- traitement de référence : 81 guérisons sur 100 traités (81%)

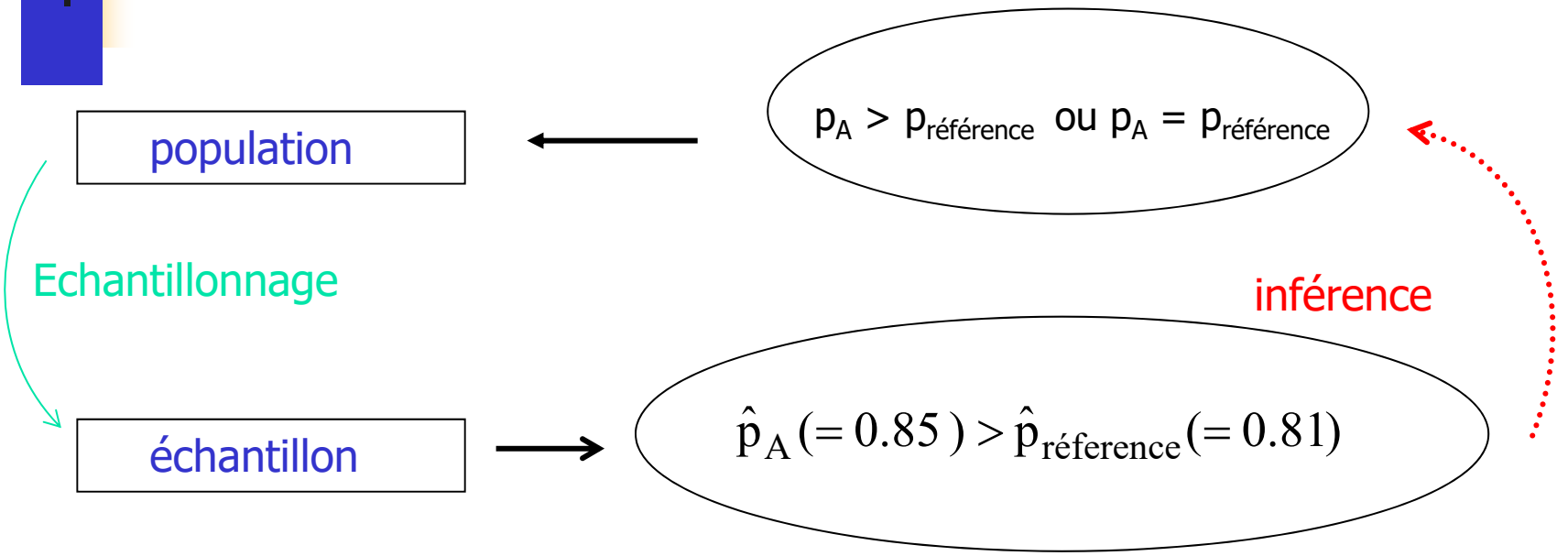
Est-ce que le traitement A est plus efficace que le traitement de référence ?

INTRODUCTION



Peut-on conclure que la probabilité d'avoir un garçon dans la population est supérieur à 0.5 ?

Introduction



Peut-on conclure que dans la population, la probabilité de guérir avec le traitement A est supérieure à la probabilité de guérir avec le référence ?

Répondre à ces deux questions revient à choisir entre deux hypothèses qui concernent des paramètres de population

HYPOTHESES

H_0 : Hypothèse nulle

H_1 : Hypothèse alternative

Exemple 1 :

$$\begin{cases} H_0 : p = \frac{1}{2} \\ H_1 : p \neq \frac{1}{2} \end{cases}$$

Test bilatéral

Exemple 2 :

$$\begin{cases} H_0 : p_A = p_{reference} \\ H_1 : p_A > p_{reference} \end{cases}$$

Test unilatéral

- L'hypothèse nulle traduit l'égalité de deux paramètres
- L'hypothèse alternative traduit la différence, la supériorité ou l'infériorité d'un des deux paramètres

RISQUES

Exemple 3 : Taux de glycémie

H_0 : ' l'individu est sain '

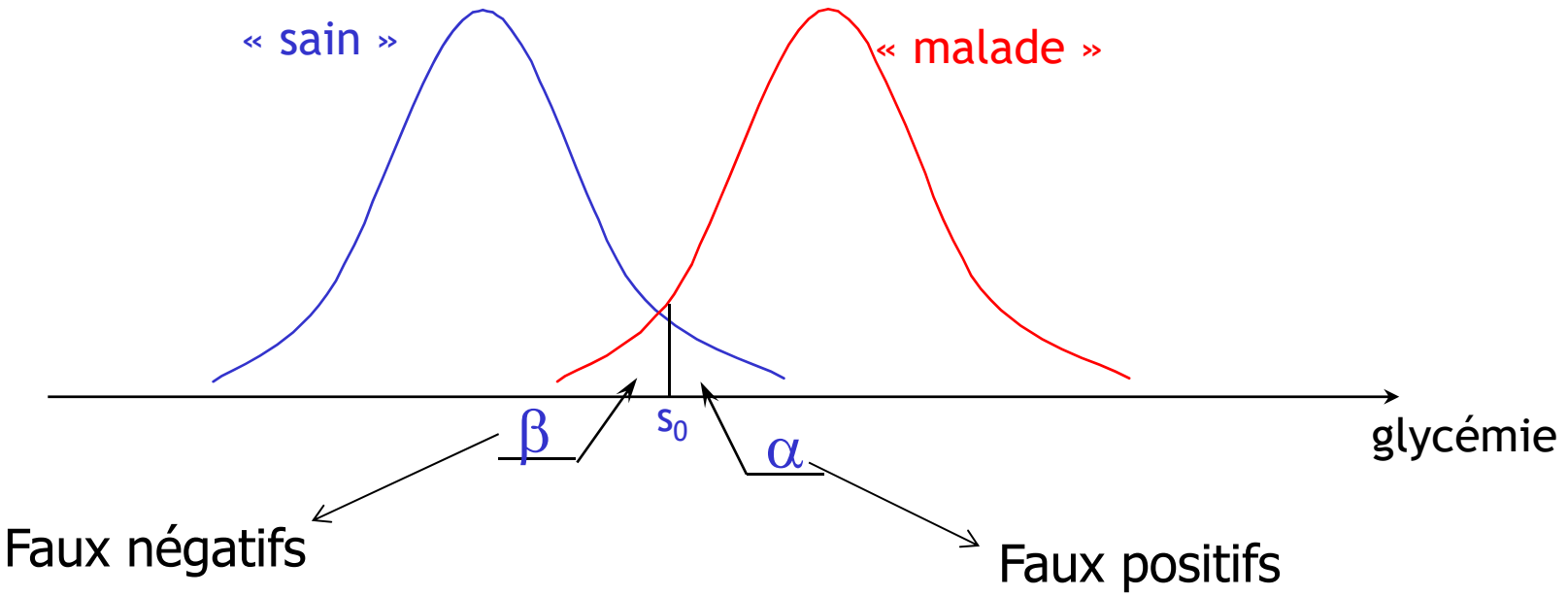
H_1 : ' l'individu est malade '

Règle de décision :

Si le taux de glycémie Tg dépasse un seuil s_0 , le médecin déclare l'individu malade

Si Tg > s_0, on rejette H_0 , sinon on l'accepte

RISQUES



Remarque : Si s_0 augmente α diminue mais β augmente

Si s_0 diminue α augmente mais β diminue

RISQUES

Définition

✓ risque de 1ère espèce

Le risque de 1ère espèce noté α est la probabilité de rejeter H_0 alors qu'elle est vraie.

✓ risque de 2ème espèce

Le risque de 2ème espèce noté β est la probabilité d'accepter H_0 alors que H_1 est vraie

Propriété :

α et β ne varient pas dans le même sens

RISQUES

décision

réalité	H_0 est acceptée	H_0 est rejetée
H_0 est vraie	$1-\alpha$	α
H_1 est vraie	β	$1-\beta$

1-β : Probabilité de rejeter H_0 alors que H_1 est vraie, on l'appelle la puissance du test

Définition : Puissance d'un test

Aptitude du test à détecter une différence entre les paramètres (décider H_1) quand cette différence existe au niveau de la population (H_1 est vraie)

RISQUES

Contrôle des risques

On est en présence de deux risques qui ne varient pas dans le même sens. D'où la difficulté (mathématique) de les contrôler simultanément.

En général, on choisit de contrôler a priori α.
La valeur communément utilisée en biologie est 0.05 (5%)

Remarque 1 : le fait de fixer a priori α suppose que le rejet de H_0 à tort est plus conséquent que le rejet de H_1 à tort.

Remarque 2 :
Le choix de la valeur de α dépend du domaine d'application, dans le domaine de l'aéronautique α est fixé à 10^{-9}, dans le domaine du dopage des chevaux il est fixé à 10^{-4}

Remarque 3 : Le calcul de β ne sera pas abordé dans ce cours

Règle de décision

Exemple 1 :

Une règle de décision simple :

$$\text{on rejette } H_0 \text{ si } \left|\hat{p} - \frac{1}{2}\right| > s_0 \text{ on rejette } H_0$$

$$\text{sinon, on l'accepte}$$

Comment choisir le seuil s_0 ?

Le calcul du seuil se fait à l'aide de la loi de probabilités de l'estimateur (ici une loi normale) et en utilisant la condition que le risque alpha est fixé a priori

Règle de décision

Résultat pour l'exemple 1 :

$$\text{on rejette } H_0 \text{ si } \left|\hat{p} - \frac{1}{2}\right| > u_{1-\frac{\alpha}{2}} \sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n}}$$

sinon, on l'accepte

Ou de façon équivalente :

$$\text{on rejette } H_0 \text{ si } \frac{\left|\hat{p} - \frac{1}{2}\right|}{\sqrt{\frac{\frac{1}{2}(1-\frac{1}{2})}{n}}} > u_{1-\frac{\alpha}{2}}$$

sinon, on l'accepte

Règle de décision

En résumé :

La construction d'un test statistique se fait en 4 étapes:

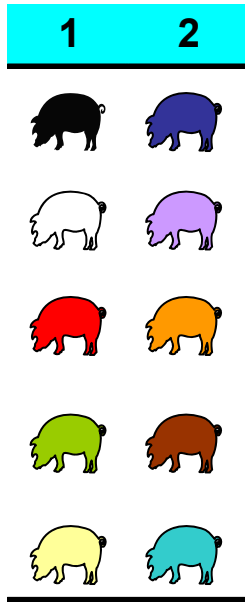
1. Traduire la question sous d'hypothèses
2. Identifier les risques que l'on veut contrôler
3. Choisir une règle de décision qui permet de choisir entre les deux hypothèses (en général, on la choisit simple pour que le calcul des risques puisse se faire !!)
4. Déterminer le seuil qui permet de choisir entre les 2 hypothèses tout en respectant les risques admis

Tests usuels

1. Tests de Student pour échantillons indépendants
2. Test d'Aspin-Welch pour échantillons indépendants
3. Test de Student apparié
4. Test d'indépendance (ou test du Khi-deux)
5. La p-valeur

Test de Student (échantillons indépendants)

Exemple :
comparaison de deux régimes d'alimentation chez les porcelets



- On dispose de 10 porcelets différents répartis dans 2 groupes (5 porcelets par groupe).
- Chaque groupe a reçu un régime d'alimentation pendant une période donnée.
- Au terme de cette période, le poids de chaque porcelet a été mesuré. Les résultats sont les suivant :

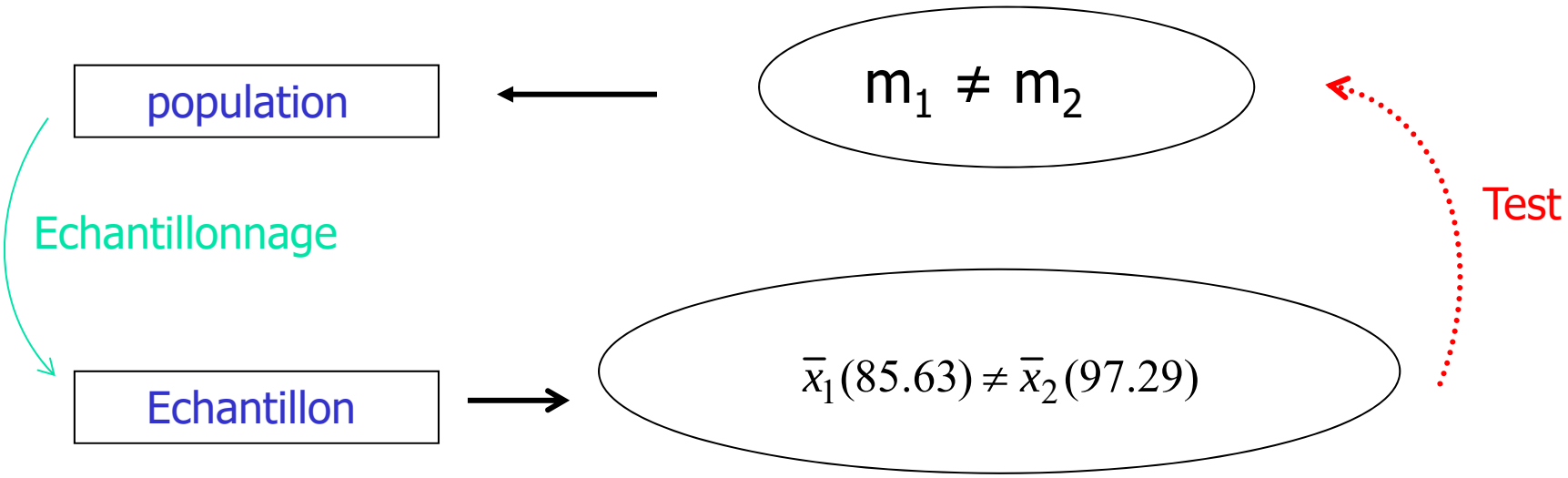
$$\text{groupe } 1 : \bar{x}_1 = 85.63 \text{ kg}, \text{sd}_1 = 6.65 \text{ kg}$$

$$\text{groupe } 2 : \bar{x}_2 = 97.29 \text{ kg}, \text{sd}_2 = 6.45 \text{ kg}$$

régime1	régime 2
92.06	87.92
83.00	98.23
92.88	95.79
82.83	105.88
77.35	98.64

Y a -t-il une différence entre les deux régimes au niveau des populations d'où l'on a extrait ces deux échantillons ?

Test de Student (échantillons indépendants)



Y a -t-il une différence ente les 2 régimes au niveau des populations d'où l'on a extrait ces 2 échantillons ?

Test de Student (échantillons indépendants)

P_1 : population susceptible de recevoir le régime 1
P_2 : population susceptible de recevoir le régime 2

Population P_1 :
moyenne m_1 , variance σ_1^2

Population P_2 :
moyenne m_2 , variance σ_2^2

Echantillonnage

Test

Echantillon 1 de taille n_1,
$\overline{x}_1 = 85.63$ kg , $sd_1 = 6.65$ kg

Echantillon 2 de taille n_2,
$\overline{x}_2 = 97.29$ kg, $sd_2 = 6.45$ kg

Hypothèses :

$$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$$

Test de Student (échantillons indépendants)

Le test qui permet de choisir entre H_0 et H_1 s'appelle le test de Student

Conditions d'utilisation du test de Student :

1. Variances homogènes (homoscedasticité) $\sigma_1^2 = \sigma_2^2$ (la plus importante)
2. Indépendance des observations entre deux individus d'un même group et deux individus de groupes différents.
3. Normalité de la variable mesurée dans l'expérience (cette dernière hypothèse n'est pas indispensable si n est assez grand)

Test de Student (échantillons indépendants)

Vérification de la condition 1

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2$$

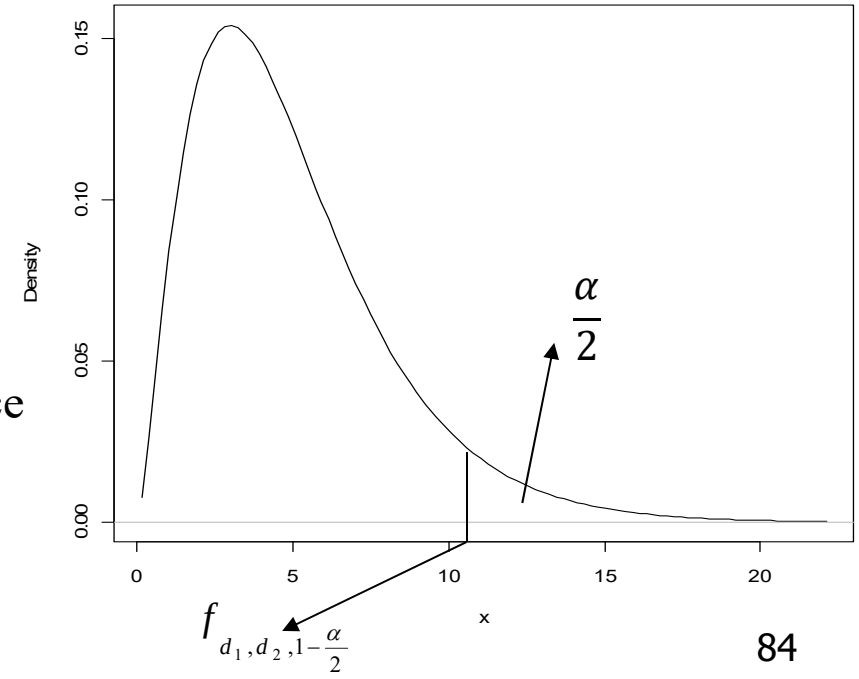
Le test qui permet de choisir entre H_0 et H_1 s'appelle le test de Fisher

on rejette H_0 si : $F = \dfrac{\max\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2\right)}{\min\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2\right)} > f_{d_1, d_2, 1-\frac{\alpha}{2}}$

sinon, on l'accepte

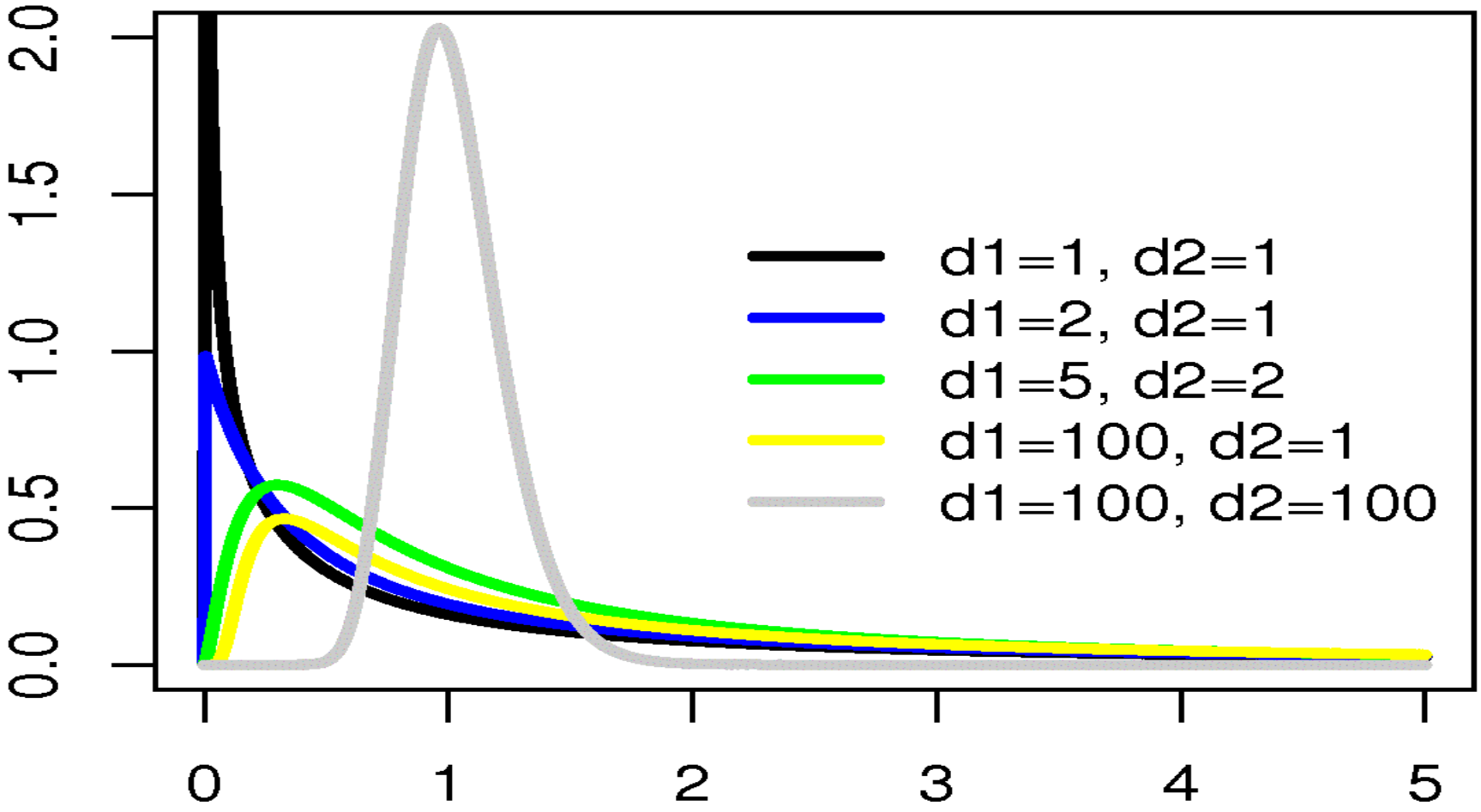
avec, d_1 le ddl correpondant à la plus grande variance

et d_2 à la variance la plus petite variance.



LOI DE FISHER

Allure de la loi de Fisher pour différents degrés de liberté



Test de Student (échantillons indépendants)

1. Si le test de Fisher ne rejette pas l'hypothèse d'égalité des variances, on suppose que l'hypothèse 1 est vérifiée et par conséquent on peut utiliser le test de Student pour comparer les deux moyennes comme suit :

$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$

on rejette H_0 si : $\dfrac{|\overline{x}_1 - \overline{x}_2|}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} > t_{1-\frac{\alpha}{2}}^{n_1+n_2-2}$

on pose $t_{obs} = \dfrac{\overline{x}_1 - \overline{x}_2}{\sqrt{\hat{\sigma}^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

sinon, on l' accepte

avec $\hat{\sigma}^2 = \dfrac{(n_1 - 1)\,\hat{\sigma}_1^2 + (n_2 - 1)\,\hat{\sigma}_2^2}{n_1 + n_2 - 2}$

En l'absence de différence (H_0 vraie), on établit que t_{obs} est l'observation d'une variable aléatoire qui suit une loi de student à (n_1+n_2-2) ddl

Test d'Aspin-Welch (échantillons indépendants)

2. Si le test de Fisher rejette l'hypothèse d'égalité des variances, dans ce cas on compare les deux moyennes à l'aide du test d'Aspin-Welch

$$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$$

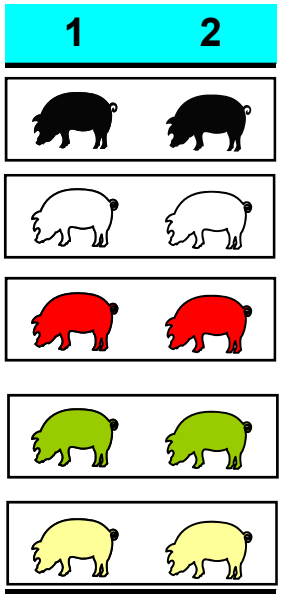
$$\text{on rejette } H_0 \text{ si : } \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} > t^{d}_{1-\frac{\alpha}{2}}$$

sinon, on l'accepte

- Le degré de liberté *d* associé à la loi de Student dans cas est fonction de de n1, n2 et des variances $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$
- Le calcul de *d* est laborieux, par conséquent il est conseillé faire ce test à l'aide d'un logiciel (Excel, R, etc.)

Test de Student apparié

Exemple :
Comparaison de deux régimes d'alimentation chez les porcelets



- On dispose de 5 porcelets.
- Chaque porcelet a reçu un premier régime pendant une période 1 , puis un deuxième régime pendant une période 2.
- Au terme de chacune des 2 périodes, le poids de chaque porcelet a été mesuré. Les résultats sont donnés dans le tableau ci-contre :

régime1	régime 2
92.06	87.92
83.00	98.23
92.88	95.79
82.83	105.88
77.35	98.64

Y a -t-il une différence entre les 2 régimes au niveau des populations d'où l'on a extrait ces deux échantillons ?

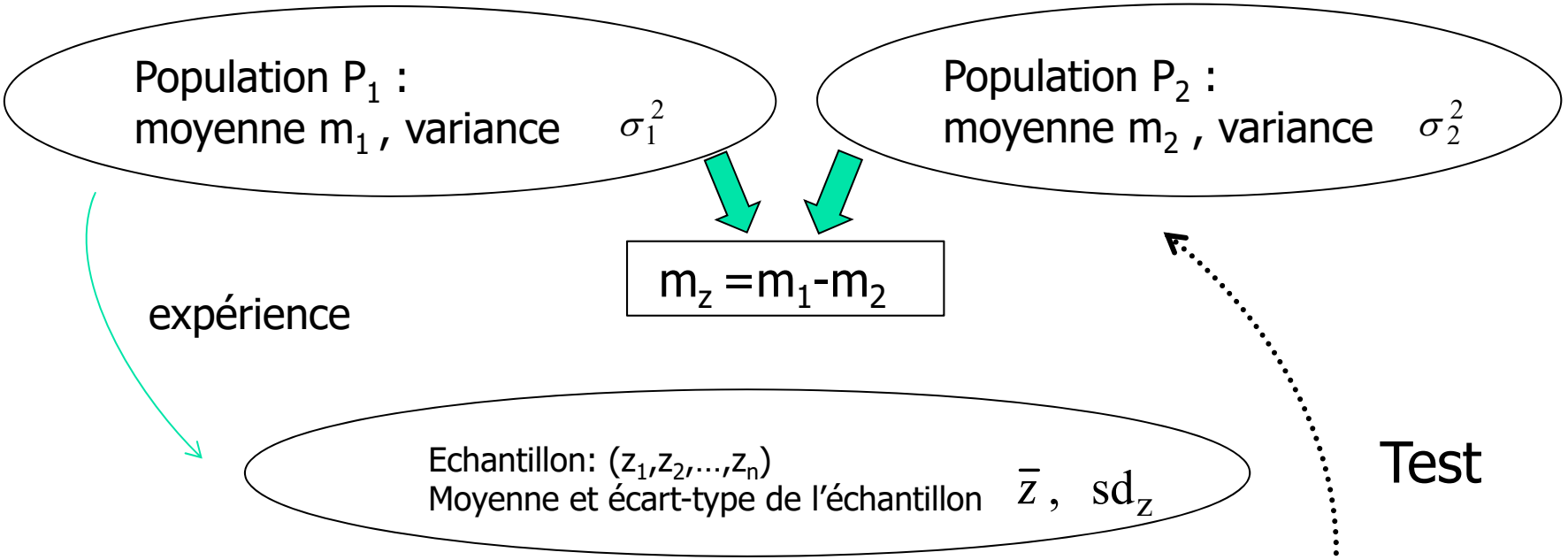
TEST DE STUDENT APPARIE

Etant donné que chaque individu a été mesuré deux fois (une mesure par période), la condition 2 du test de Student concernant l'indépendance entre individus d'un groupe à l'autre n'est plus valable.

Dans ce cas, on compare les deux régimes en se basant sur la différence (augmentation ou diminution) entre les deux mesures faites sur l'individu.
soit $z_i=x_i-y_i$ $(i=1,\ldots,n)$, où x_i désigne la mesure faite au terme de la 1ère période et y_i la mesure faite au terme de la $2^{ème}$ période.

Dans ce cas, tester l'égalité des moyennes revient à tester si la différence des moyennes est nulle. Autrement dit, tester si en moyenne l'augmentation ou la diminution est nulle.

TEST DE STUDENT APPARIE



$$H_0 : m_z = 0 \qquad H_1 : m_z \neq 0$$

$$\text{on rejette } H_0 \text{ si : } \frac{|\bar{z}|}{\frac{sd_z}{\sqrt{n}}} > t_{1-\frac{\alpha}{2}}^{n-1}$$

sinon, on l' accepte

TEST D'INDEPENDANCE

Exemple 1:
On a noté la couleur du pelage (gris/blanc)et la forme du pelage (lisse/rude) sur 133 souris, les résultats sont donnés dans la table ci-dessus.

	gris	blanc
lisse	75	20
rude	28	10

Question :
Les caractères 'couleur du pelage' et 'forme du pelage' sont-ils gouvernés par deux couples de gènes indépendants ?

TEST D'INDEPENDANCE

Exemple 2

200 individus ont été répartis dans 2 groupes (100 par groupe). Un groupe a reçu le traitement de référence et l'autre groupe un nouveau traitement.

Le nombre de guéris observé dans chaque groupe est donné dans la table ci-dessus.

	guéri	Non guéri
Nouveau traitement	85	15
Traitement de référence	80	20

Question :
L'état de l'individu (guéri ou pas) dépend-t-il du traitement reçu (référence ou nouveau traitement) ?

TEST D'INDEPENDANCE

Un test d'indépendance (ou du Khi-deux) permet de tester la dépendance ou l'indépendance de deux caractères qualitatifs.

Méthodologie :

- on calcule la table des effectifs que l'on aurait du observer s'il y avait indépendance entre les deux caractères qualitatifs, appelés effectifs théoriques.
- 2ème étape : on calcule la distance entre la table des effectifs observés et celle des effectifs théoriques
- 3ème étape: on effectue le test en comparant cette distance à un seuil choisi en fonction du risque α fixé a priori, si cette distance dépasse le seuil on rejette l'hypothèse d'indépendance sinon on l'accepte.

Hypothèses :

H_0 : 'les deux caractères sont indépendants'
H_1 : 'les deux caractères sont dépendants'

TEST D'INDEPENDANCE

1ème étape : calcul des effectifs théoriques

	gris	blanc	total
lisse	n_1=75	n_2=20	a=95
rude	n_3=28	n4=10	b=38
total	c=103	d=30	N=133

Table des effectifs observés

	gris	blanc
lisse	103*95/133	30*95/133
rude	103*38/133	30*38/133

Table des effectifs théoriques

Le calcul de l'effectif théorique se fait à l'aide de la propriété vue en calcul des probabilités : si deux événements A et B sont indépendants alors P(A∩B)=P(A)*P(B)

Exemple: A={avoir le pelage gris} et B={avoir le pelage lisse}

$$P(A) = \frac{103}{133}, P(B) = \frac{95}{133}$$

$$\text{si A et B sont indépendants alors } P(A \cap B) = \frac{103}{103} * \frac{95}{133}$$

Soit en terme d'effectif : $\frac{103}{133} * \frac{95}{133} * 133 = 103 * 95 / 133$

TEST D'INDEPENDANCE

2ème étape :
On calcule la distance entre les deux tables comme suit :

$$\chi^2_{obs} = \frac{(75-103*95/133)^2}{103*95/133} + \frac{(20-30*95/133)^2}{30*95/133} + \frac{(28-103*38/133)^2}{103*38/133} + \frac{(10-30*38/133)^2}{30*38/133}$$

Ou de façon générale :

$$\chi^2_{obs} = \frac{(n_1 - a*c/N)^2}{a*c/N} + \frac{(n_2 - a*d/N)^2}{a*d/N} + \frac{(n_3 - c*b/N)^2}{c*b/N} + \frac{(n_4 - d*b/N)^2}{d*b/N}$$

TEST D'INDEPENDANCE

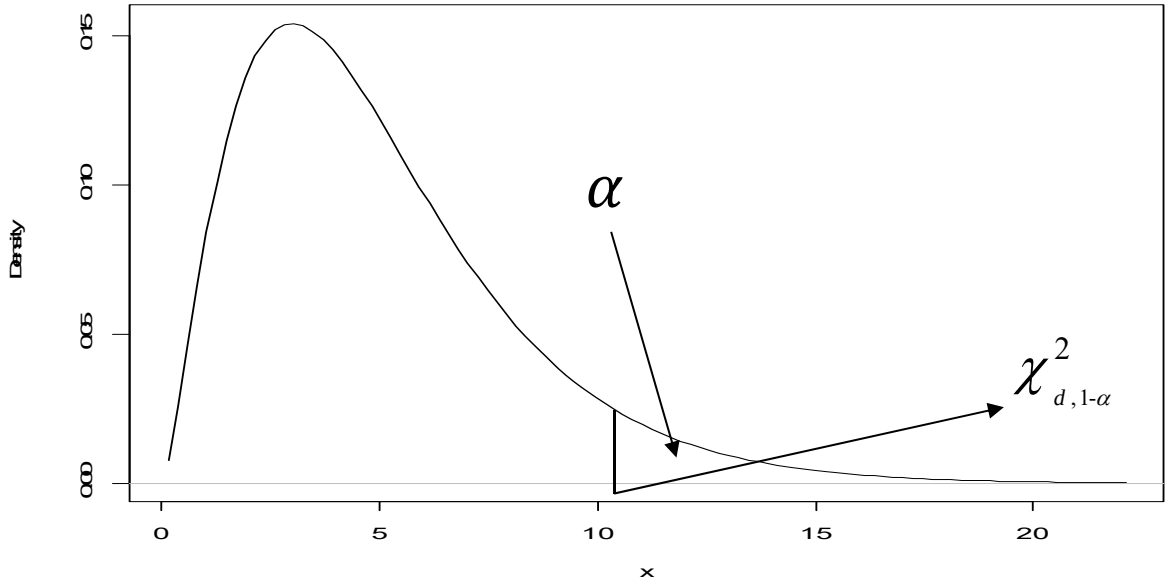
3ème étape : le test

on rejette H_0 si $\chi^2_{obs} > \chi^2_{d,1-\alpha}$ sinon, on l'accepte

où, d est le degré de liberté de la distribution du χ^2, ici d = (r - 1) (s - 1)

r : nombre de modalités du 1er carctère

s : nombre de modalités du 2ème caractère



TEST D'INDEPENDANCE

Remarques :

- Le test du Khi-deux n'est valable que si les effectifs théoriques sont tous supérieurs à 5
- Le test du khi-deux peut-être généralisé à deux caractères qualitatifs qui ont un nombre différent de modalités et supérieurs à 2
- Attention le test du Khi-deux se fait sur les effectifs et non pas sur les pourcentages
- Si un des effectifs théoriques est inférieur à 5, on peut utiliser un autre test d'indépendance par exemple le test de Fisher exact

La p-valeur

Pour introduire cette notion, nous allons nous placer dans le cadre d'une comparaison de deux moyennes (test de Student)

$$H_0 : m_1 = m_2 \qquad H_1 : m_1 \neq m_2$$

$$\text{on rejette } H_0 \text{ si : } \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > t^{n_1+n_2-2}_{1-\frac{\alpha}{2}}$$

sinon, on l'accepte

$$\text{notons } t_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

t_{obs}, est la valeur observée à l'issue de l'expérience donc elle est connue !!

Rappelons qu'en l'absence de différence (H_0 vraie), on montre que t_{obs} est l'observation d'une variable aléatoire qui sui une loi de student à (n_1+n_2-2) ddl

La p-valeur

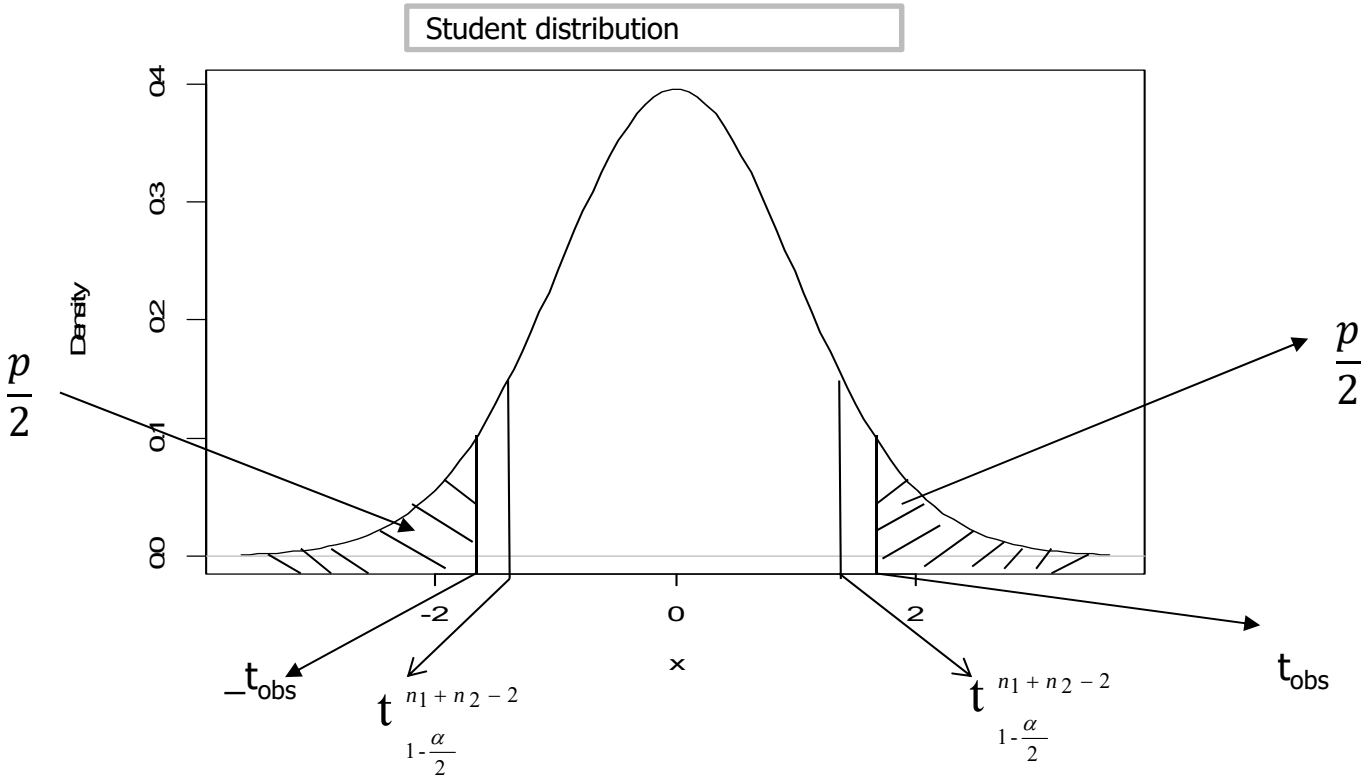
En utilisant la loi de Student on peut calculer la probabilité d'observer une valeur plus grande ou égale à t_{obs} en l'absence de différence (H_0 vraie).
Autrement dit la probabilité que la valeur du t_{obs} observée soit due au hasard
Cette probabilité est appelée p valeur et notée p.

Pour α fixé a priori,

Si $p > \alpha$, on dit que la valeur du t_{obs} n'est pas suffisamment importante par rapport aux fluctuations aléatoires pour pouvoir raisonnablement décider qu'elle est un artefact dû au hasard » et par conséquent on ne rejette pas H_0

Dans le cas contraire, on dit que la valeur du t_{obs} est suffisamment importante par rapport aux fluctuations aléatoires pour pouvoir raisonnablement exclure qu'elle soit un artefact dû au hasard » et par conséquent on rejette H_0

La p-valeur



on voit que si :

- $p < \alpha$, alors $|t_{obs}| > t^{n_1+n_2-2}_{1-\frac{\alpha}{2}}$ et donc on rejette H_0
- $p > \alpha$, alors $|t_{obs}| < t^{n_1+n_2-2}_{1-\frac{\alpha}{2}}$ et donc on ne rejette pas H_0

La p-valeur

En conclusion

- Les logiciels fournissent la p valeur (p) pour tous les tests.
- La règle de décision associée à tous les tests est la suivante :

  - Si $p \leq \alpha$, on rejette H_0
  - Si $p > \alpha$, on ne rejette pas H_0

- Lorsqu'on rejette H_0, on dit que le test est significatif
- Lorsqu'on ne rejette pas H_0, il est dit non significatif

  - Lorsqu'on rejette H_0, on le fait avec un risque α
  - Lorsqu'on ne rejette pas H_0 , on le fait avec un risque β