Master
Biologie cellulaire et physiopathologie
: initiation à la communication scientifique

Quelques notions de base de statistiques appliquées à la biologie

Étienne Roux

Adaptation cardiovasculaire à l'ischémie INSERM U 1034
UFR des Sciences de la Vie Université Bordeaux Segalen

contact: etienne.roux@u-bordeaux2.fr
support de cours :
e-fisio.net

Qu'est-ce que les statistiques ?

« There are three kinds of lies:
lies, damned lies, and statistics. »

attribué par Mark Twain à
Benjamin Disraeli

quelques définitions

La statistique : ensemble des instruments et de recherches mathématiques permettant de déterminer les caractéristiques d'un ensemble de données.

Une statistique : un nombre calculé à partir d'observations.

Les statistiques : produit des analyses reposant sur l'usage de la statistique.

caractéristiques des données biologiques

♦ *Caractéristiques en biologie de la mesure d'un certain nombre de facteurs*

– variabilité des réponses en biologie

exemples :
concentration cytosolique en calcium d'une cellule
niveau d'expression d'une protéine
valeur de la glycémie
taille d'une personne

– mesure à partir d'échantillons

principe du sondage : on travaille sur un échantillon que l'on suppose représentatif.

exemples :
prélèvement d'un échantillon de sang
étude d'une population de cellules

→ décrire mathématiquement la réalité observée

décrire = résumer et représenter les données

types de questions que l'on se pose

- *Types de questions que l'on se pose en recherche :*

– qu'est-ce qui produit un effet ?

– qu'est-ce qui produit l'amplitude d'un effet ?

– qu'est-ce qui produit l'effet le plus important ?

→ répondre à la question posée à partir des données observées

la démarche d'analyse statistique

trois étapes principales :

1 - la collecte des données

2 - le traitement des données collectées
→ caractériser la relation entre variables

- **Statistiques descriptives**

3 - l'interprétation des données
→ à partir des données obtenus sur l'échantillon, inférer les caractéristiques de la population d'origine

→ estimer si plusieurs échantillons proviennent ou non d'une même population

- **Statistiques inférentielles ou inférences statistiques**

s'appuie sur la théorie des sondages et la statistique mathématique

la notion d'individu (statistique) : unité statistique

individu (définition statistique) = unité statistique:
élément de l'ensemble étudié

ex : un patient recevant un traitement à l'hôpital
un étudiant inscrit dans un master biologie-santé
une cellule en culture

- pour chaque individu, on dispose d'un ou plusieurs paramètres.
- la définition statistique de l'individu est différence de sa définition courante
- la définition de l'individu dépend des paramètres étudiés

exemple 1 : paramètre étudié : note d'un étudiant dans un groupe de TD
un individu = un étudiant

exemple 2 : paramètre étudié : note moyenne de chaque groupe de TD d'étudiant inscrit dans une licence.
un individu = un groupe de TD

la notion de population

population (définition statistique) = ensemble d'individus sur lequel on étudie des paramètres

ex : ensemble des patients recevant un traitement à l'hôpital
ensemble des étudiants inscrits un master biologie-santé
ensemble de cellules

♦ on peut ne pas connaitre tous les individus qui composent une population

exemple : population humaine
les globules rouges de souris

notion d'échantillon :

échantillon = partie d'une population

exemple : 1000 personnes humaines

♦ on connait tous les individus qui composent un échantillon

la notion de variable

une variable (définition statistique) = paramètre étudié sur un individu

ex : âge des patients recevant un traitement à l'hôpital
sexe des patients recevant un traitement à l'hôpital
maladie des patients reçus à l'hôpital
traitement des patients reçus à l'hôpital
réussite du traitement des patients reçus à l'hôpital

- une ou plusieurs variables peuvent être associées sur un individu

- les variables peuvent être de nature variée :
  Variables qualitatives et quantitatives
  Variables indépendantes et variables dépendantes
  Variables contrôlées et non contrôlées

la notion de variable *Variables qualitatives et quantitatives*

♦ variable qualitative = variable statistique dont les valeurs s'expriment de façon littérale (ou par un codage), sur lesquelles les opérations arithmétiques comme le calcul de la moyenne n'ont pas de sens.

exemples :

mortalité dans une population de cellules, par la coloration au bleu trypan.
sexe des patients recevant un traitement à l'hôpital.

codage: la qualité de la variable peut être exprimée par un codage.

exemple : cellule morte : M cellule vivante : V

cellule morte : 1 cellule vivante : 0

attention! un codage chiffré ne transforme pas une variable qualitative en variable quantitative.

Un chiffre n'est pas forcément un nombre

ex : sudoku

la notion de variable *Variables qualitatives et quantitatives*

♦ variable quantitative = variable statistique dont les valeurs s'expriment par des nombres, sur lequel les opérations arithmétiques comme le calcul de la moyenne ont un sens.

exemples : concentration calcique cytosolique d'une cellule
âge des patients recevant un traitement à l'hôpital
dose d'un traitement administré à des patients

variable continue : peut prendre toute valeur réelle
exemple : concentration calcique cytosolique d'une cellule

variable discrète : ne peut prendre d'un nombre fini de valeurs
exemple : nombre d'enfants par femme

attention! une variable chiffrée n'est pas forcément une variable quantitative (le chiffre peut être un codage)

On peut transformer une variable quantitative en variable qualitative, avec une perte d'information.

ex: dose d'un traitement administré à des patients
→ en fonction de la dose, classement en catégories : très faible dose, faible dose, dose normale, forte dose, très forte dose.

la notion de variable *Variables indépendantes et variables dépendantes*

♦ variable indépendante = variable statistique dont les valeurs sont indépendantes des autres variables étudiées

♦ variable indépendante = variable statistique dont les valeurs sont dépendantes des autres variables étudiées

exemples :
on étudie l'effet de deux substances potentiellement cytotoxiques sur des cellules cancéreuses en culture, et on mesure la survie des cellules en fonction de la substance administrée.

variabledépendante : survie de la cellule
variable indépendante : substances cytotoxiques appliquées à la cellule

la notion de variable *Variables contrôlées et non contrôlées*

♦ variable contrôlée = variable statistique dont les valeurs sont imposées par l'expérimentateur

expérimentation (experiment)

Dans les études d'expérimentation, les variables indépendantes sont contrôlées

exemples :

- effet de l'adrénaline sur la fréquence cardiaque.

- détermination sur la souris de la quantité minimale contaminante de cerveau de bovin atteint d'ESB.

la notion de variable *Variables contrôlées et non contrôlées*

♦ variable non contrôlée = variable statistique dont les valeurs dépendent pas de l'expérimentateur

observation (survey)

Dans les études d'observations, les variables indépendantes ne sont pas contrôlées.

exemples :

- fréquence des cancers de la thyroïde après l'accident de Tchernobyl, dans une zone géographique donnée.

- admission aux urgences pour problèmes respiratoires en fonction de l'intensité de la pollution atmosphérique

exercices

données statistiques : individu, variable, population
dans chacun des exercices suivants, déterminer :
l'individu (statistique)
la population (statistique)
la ou les variables
le caractère de chaque variable : qualitatif ou quantitatif; indépendant ou dépendant, contrôlé ou non contrôlé.

exercice 1 : réponse calcique de cellules isolées stimulées
position du problème : on analyse la réponse calcique de cellules à une stimulation par la caféine. On mesure l'amplitude du pic calcique grâce à une sonde fluorescente, dont l'intensité de fluorescence dépend du calcium. Après calibration, la concentration en calcium est calculée en nM. La mesure est effectuée sur 39 cellules.

exercices *série A*

exercice 2 : détermination par Western blot du niveau d'expression de la protéine P sur culture de cellules

position du problème : sur des cellules en culture, on analyse par Western blot le niveau d'expression de la protéine P. Les suspensions de cellules sont broyées et les protéines extraites par centrifugation. Le niveau d'expression estr évalué par l'intensité de la bande correspondante à la protéine P, normalisée à par rapport à l'actine. Les mesures sont répétées sur 6 lots de cellules.

exercice 3 : influence du Ca^{2+} extracellulaire sur la réponse contractile d'anneaux de bronches

position du problème : on analyse la réponse contractile d'anneaux de bronches à une stimulation par l'acétylcholine. Chasue anneau est relié à un transducteur de force qui mesure la force développée par l'anneau, (exprimée en % d'une réponse de référence), lorsque l'acétylcholine est introduite dans la cuve. Pour déterminer le rôle possible du Ca^{2+} extracellulaire dans la réponse, des mesures sont faites sur 7 anneaux avec du Ca^{2+} extracellulaire et sur 8 anneaux sans Ca^{2+} extracellulaire.

exercices *série A*

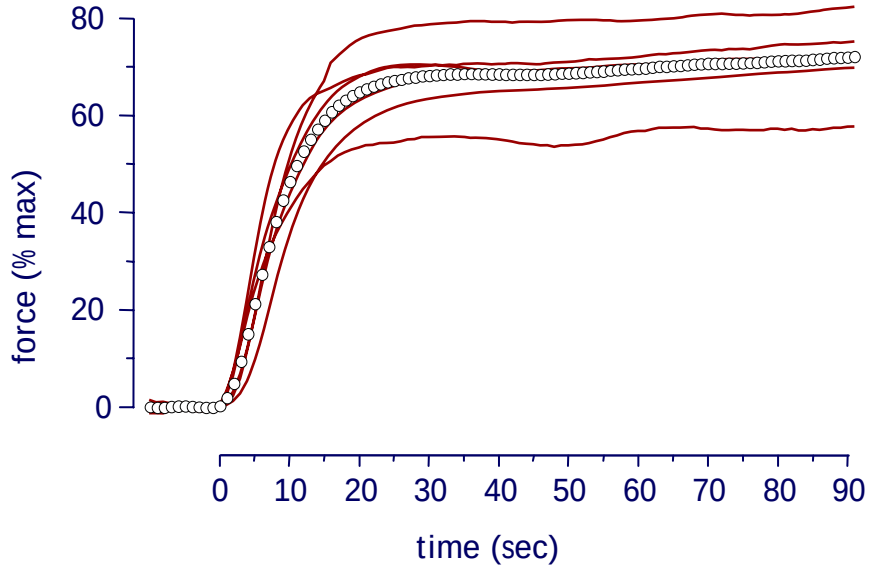
exercice 4 : relation entre la dose d'un médicament et la pression artérielle
position du problème : on analyse l'effet de 4 doses différentes d'un même traitement sur la pression artérielle d'un lot de 23 rats. La pression artérielle est mesurée au niveau de la carotide, et est exprimée en mmHg.

exercice 5 : analyse de la fréquence cardiaque au repos et à l'effort dans un groupe H/F
position du problème : on mesure la fréquence cardiaque d'un groupe de 31 étudiants, en effectuant sur chaque individu, dont on note le sexe, une mesure au repos et après effort.

statistiques descriptives variabilité des processus biologiques

caractéristiques générales

ex : tension développé par un anneau de trachée de rat en réponse à une stimulation cholinergique



- la variabilité est la règle
- la variabilité est non prévisible
- la variabilité des résultats est différente de l'erreur instrumentale
- la « marge d'imprécision » d'estimation de la tendance centrale est un intervalle de probabilité

statistiques descriptives variabilité des processus biologiques

décrire la réalité biologique

données « brutes » : ensemble des valeurs mesurées sur chaque individu

exemple : contraction d'anneaux de bronches de rat
individu : anneau de bronche de rat
variable : amplitude de la contraction

anneau	force (g)
1er	1,14596
2e	1,0461
3e	0,67606
4e	0,57967
5e	1,16159
6e	0,64212
7e	1,01782
8e	0,66019
9e	1,20027
10e	0,71591
11e	0,54514
12e	0,90245
13e	0,61038
…	
29e	1,32689

en elles-mêmes, les données brutes donnent peu d'informations utiles.

→ décrire mathématiquement la réalité observée

décrire = résumer et représenter les données

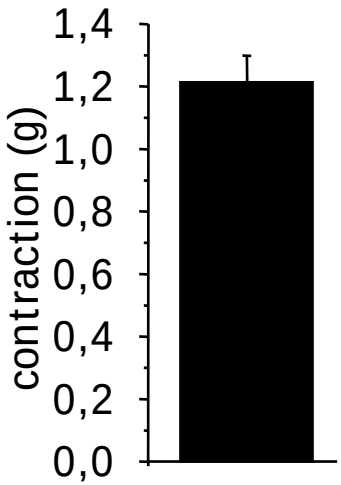
statistiques descriptives variabilité des processus biologiques

décrire la réalité biologique

statistiques descriptives : résumé mathématique de la réalitée observée

exemple : contraction d'anneaux de bronches de rat
individu : anneau de bronche de rat
variable : amplitude de la contraction

« La force mesurée était de 1,21 ± 0,08 g (n = 29) »



résumé mathématique en 3 valeurs.

NB : code ASCII pour ± : alt +0177

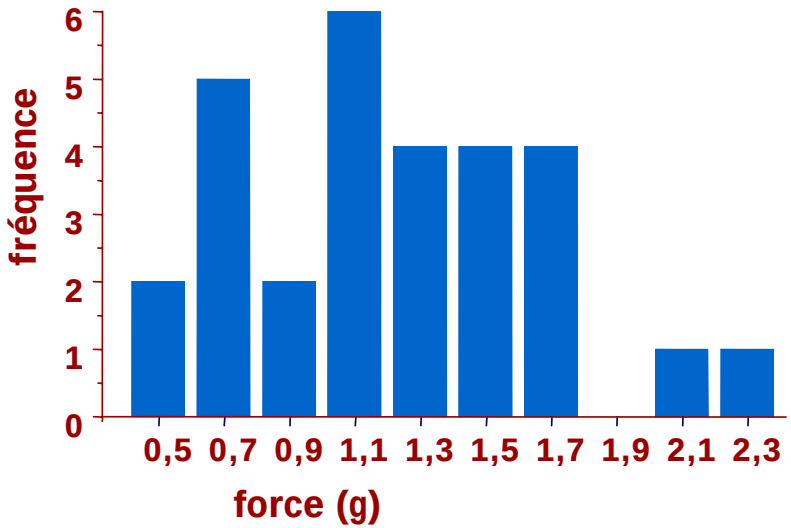
notion et types de distribution

distribution d'une variable : ensemble des valeurs, modalités ou classes d'une variable statistique, et des effectifs ou fréquences associées

exemple : contraction d'anneaux de bronches de rat

anneau	force (g)
1er	1,14596
2e	1,0461
3e	0,67606
4e	0,57967
5e	1,16159
6e	0,64212
7e	1,01782
8e	0,66019
9e	1,20027
10e	0,71591
11e	0,54514
12e	0,90245
13e	0,61038
…	
29e	1,32689

force (g)	fréquence
0,5	2
0,7	5
0,9	2
1,1	6
1,3	4
1,5	4
1,7	4
1,9	0
2,1	1
2,3	1
2,5	0



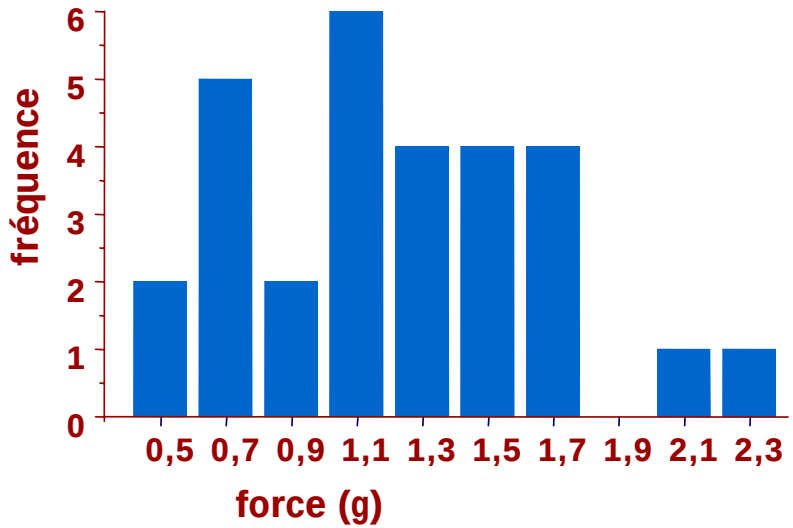
On peut décrire mathématiquement certains types de distribution

notion et types de distribution

distribution d'une variable : ensemble des valeurs, modalités ou classes d'une variable statistique, et des effectifs ou fréquences associées

exemple : contraction d'anneaux de bronches de rat

pour les variables continues, l'analyse de la distribution nécessite de regrouper les valeurs en classes

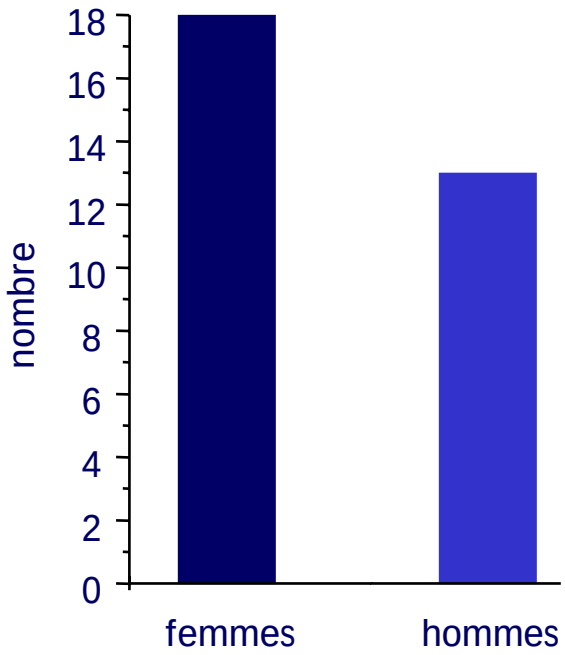


On peut décrire mathématiquement certains types de distribution

notion et types de distribution *distribution binomiale*

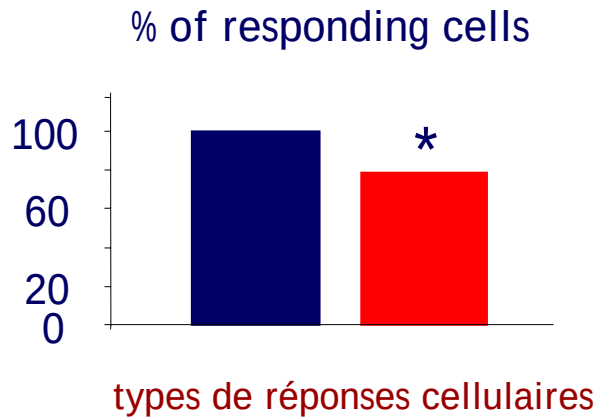
La variable peut prendre deux valeurs – pas forcément numériques.

répartition hommes/femmes dans une population



proportion de gauchers dans une population

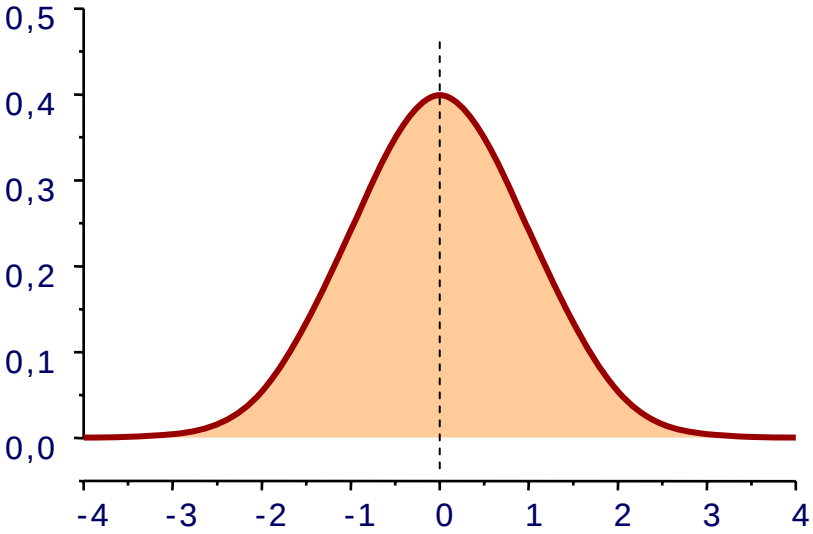
droitiers	gauchers



types de réponses cellulaires

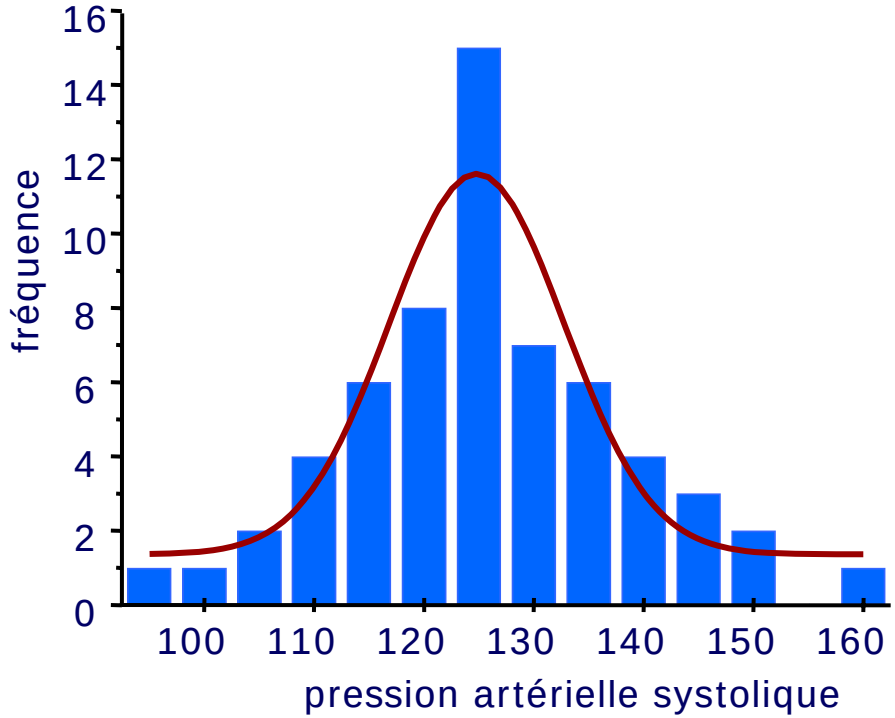
notion et types de distribution

distribution « normale » ou gaussienne



loi de distribution de probabilité, définie par une fonction de densité de probabilité de la forme :

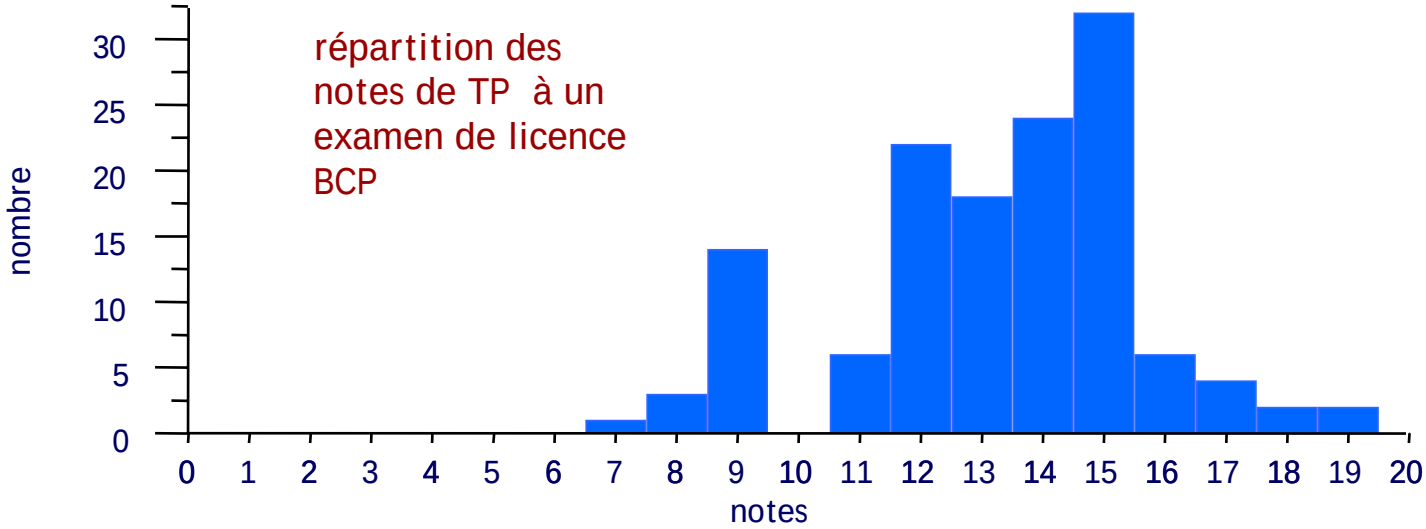
$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 / 2\sigma^2}$$

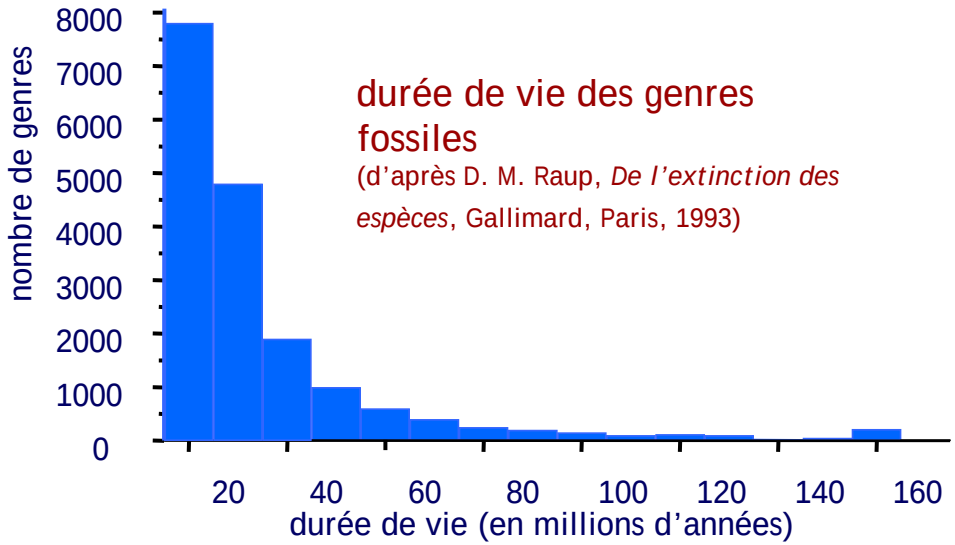


ex : valeur de la pression artérielle systémique systolique dans une population

distribution « normale » ou gaussienne : courbe « en cloche »

notion et types de distribution *autres types de distributions*



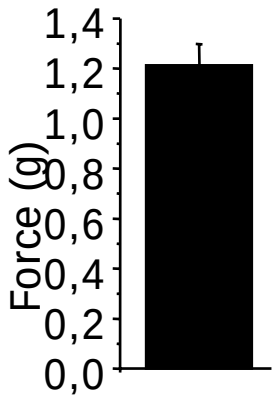


les distributions ne sont pas forcément gaussiennes

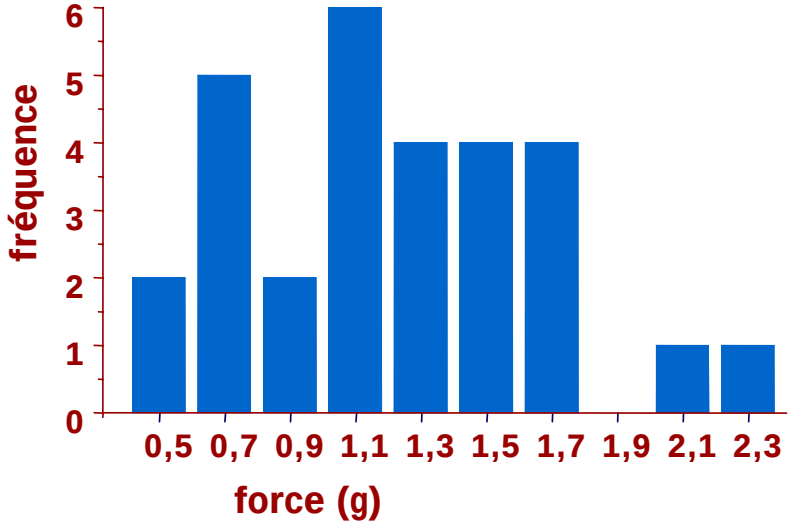
(les variables ne suivent pas forcément une « courbe en cloche »)

tendance centrale et dispersion

résumé mathématique de la réalitée observée : mesure mathématique de la tendance centrale et de la dispersion des valeurs de la variable étudiée



exemple : contraction d'anneaux de bronches de rat

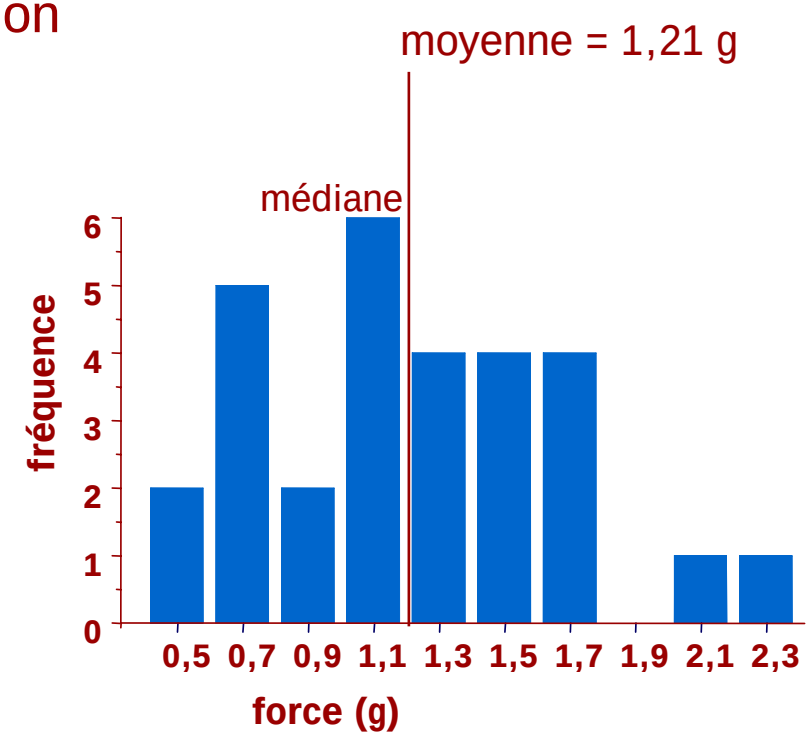


« La force mesurée était de 1,21 ± 0,08 g (n = 29) »

moyenne arithmétique (arithmetic mean)

moyenne arithmétique : somme des valeurs de la variable divisée par le nombre de valeurs
= centre de gravité de la distribution

(pour éviter les biais par simplification, faire le calcul avec une décimale supplémentaire par rapport au nombre de décimales de la valeur exprimée de la moyenne)

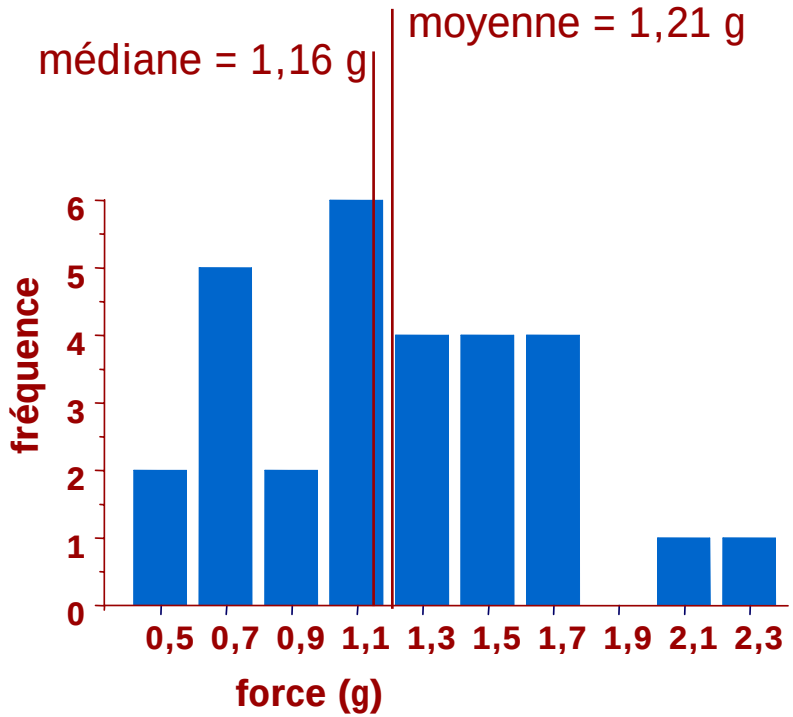


intérêts et limites :

- très utilisée en statistiques descriptive et inférentielle
- souvent, pas toujours, la mesure la plus pertinente de la tendance centrale

médiane

médiane : valeur de part et d'autre de laquelle se distribue par moitié les valeurs de la variable (50 % des valeurs sont inférieures à la médiane, et 50 % sont supérieures).

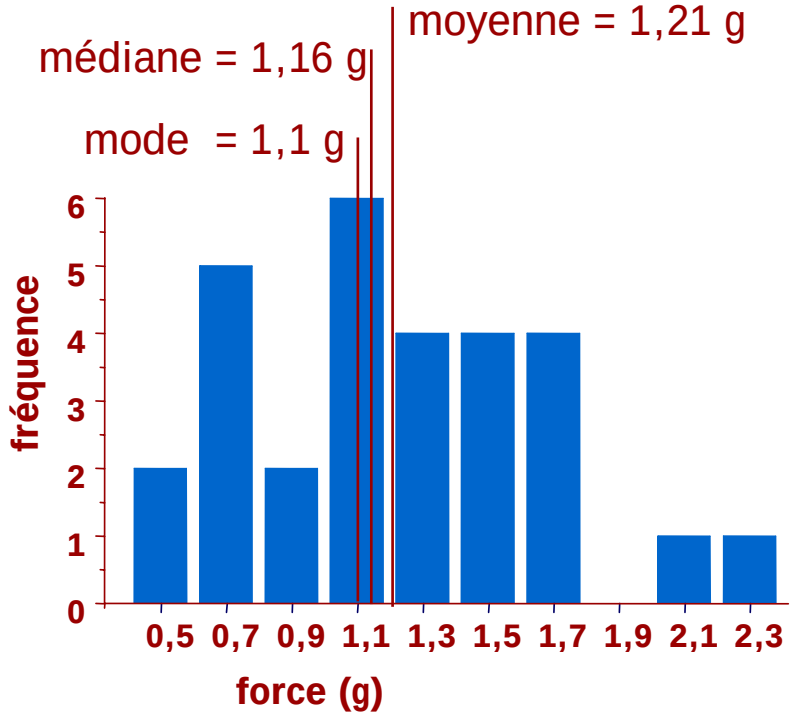


intérêts et limites :

- intérêt théorique : dans certains cas, « bonne » manière de décrire la tendance centrale
- peu utilisée pour les calculs de signification statistique

mode

mode : valeur de la variable qui survient avec la plus grande fréquence
variables discrètes (discontinues) : valeur exacte
variables continues : dépend du mode de calcul



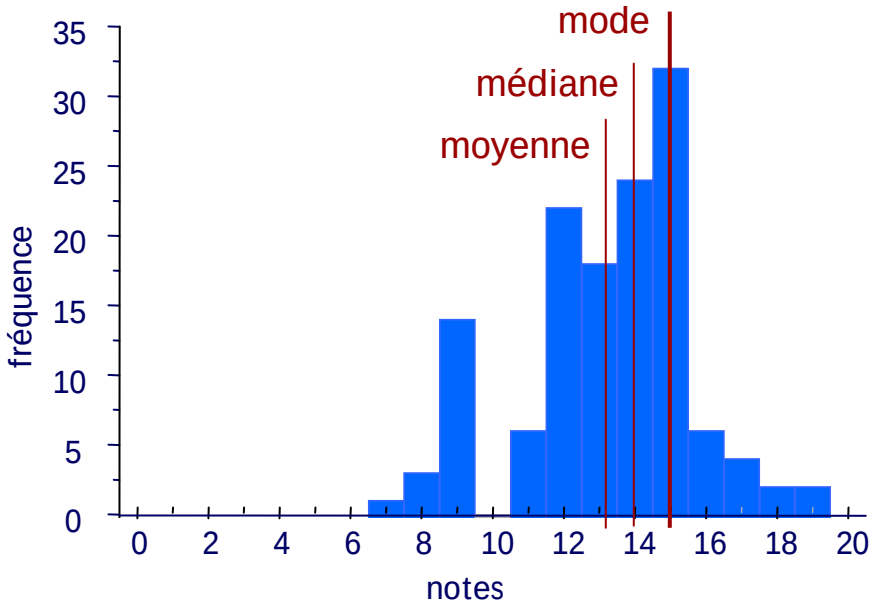
intérêts et limites :

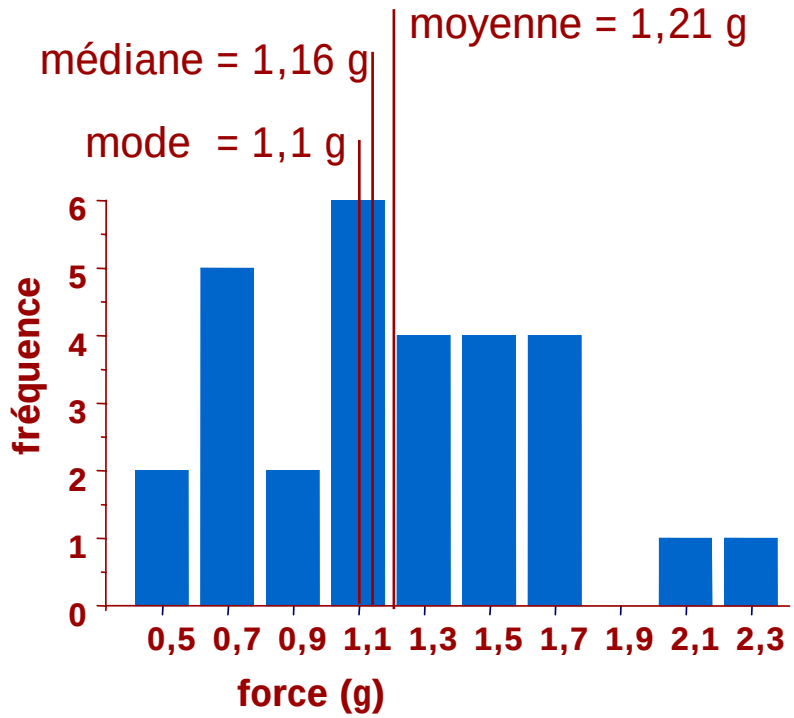
- facile à déterminer
- difficile à manipuler mathématiquement (pour tester statistiquement des hypothèses)
- intérêt théorique : dans certains cas, « bonne » manière de décrire la tendance centrale

choix de la mesure

choix de la mesure

- dépend de la loi de distribution
- dépend de la question posée

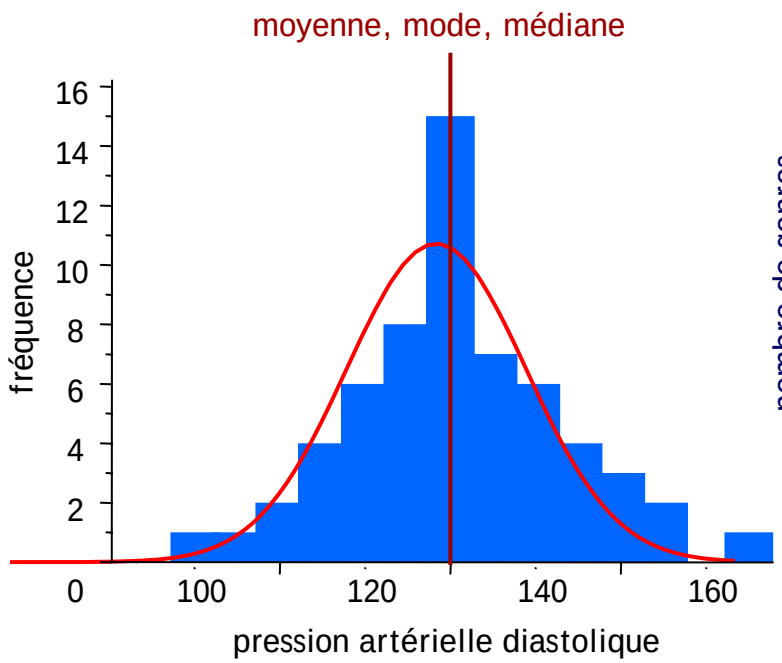


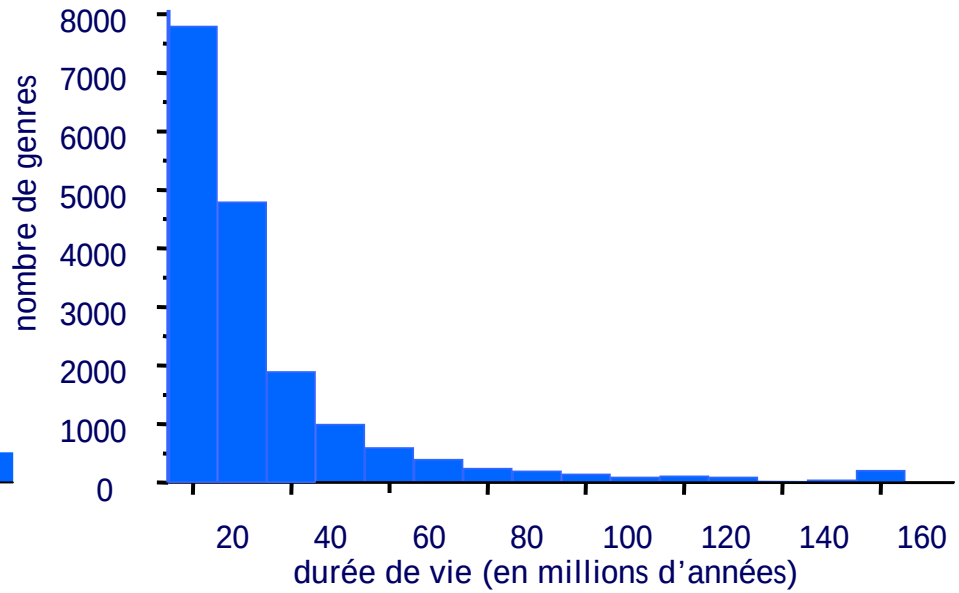


choix de la mesure

choix de la mesure

- dépend de la loi de distribution
- dépend de la question posée





si la distribution est symétrique, moyenne, médiane et mode sont similaires

- dans la plupart des cas : moyenne
- médiane et mode intéressants dans certains cas

écart (range)

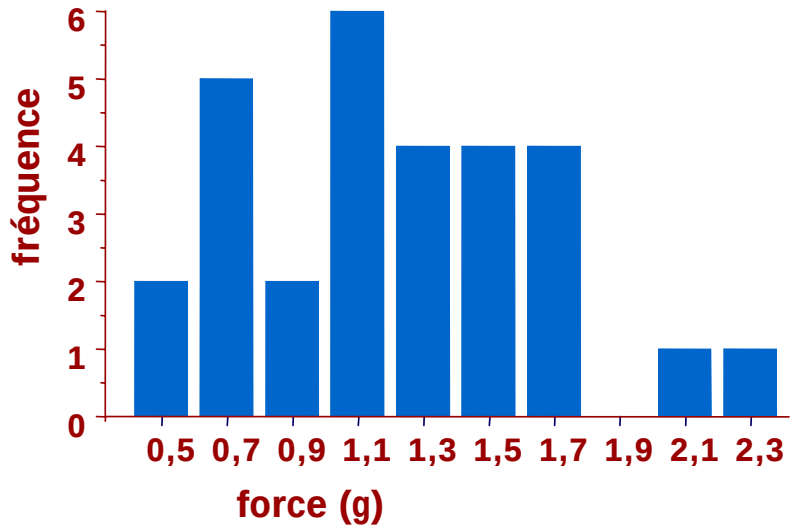
écart : différence entre la plus petite et la plus grande valeur.

intérêt et limites :

- facile à calculer
- très instable (une seule valeur extrème modifie fortement la valeur de l'écart)

mean	min	max	range	n
1.21374	0.54514	2.22593	1.68079	29

exemple : contraction d'anneaux de bronches de rat



déviation moyenne (mean deviation)

déviation moyenne : moyenne arithmétique de la différence, en valeur absolue, entre chaque valeur et la moyenne arithmétique.

principe :

1 – pour chaque point, on calcule la différence avec la moyenne

plus le point s'écarte de la moyenne, plus la différence est grande, mais elle peut être négative)

2 – pour chaque point, on prend la valeur absolue de cette différence

plus le point s'écarte de la moyenne, plus la différence est grande, et elle est toujours positive

3 – on fait la somme de la valeur absolue des différence

plus les points s'écartent de la moyenne, plus la somme des carrés est grande, mais elle dépend aussi du nombre de valeurs

4 – on divise cette somme par la taille de la population

→déviation moyenne

plus les points s'écartent de la moyenne, plus la déviation moyenne est grande, indépendamment du nombre de valeurs étudiées. Elle a la même unité que la variable étudiée

intérêt et limites :

- mesure très rarement utilisée

écart-type (standard deviation)

écart-type (standard deviation) : racine carrée de la variance

principe :

1 – pour chaque point, on calcule la différence avec la moyenne

plus le point s'écarte de la moyenne, plus la différence est grande, mais elle peut être négative)

2 – pour chaque point, on prend le carré de cette différence

plus le point s'écarte de la moyenne, plus le carré différence est grand, et il est toujours positif)

3 – on fait la somme de ces carrés

plus les points s'écartent de la moyenne, plus la somme des carrés est grande, mais elle dépend aussi du nombre de valeurs

4 – on divise la somme des carrés par la taille de la population

→variance

plus les points s'écartent de la moyenne, plus la variance est grande, indépendamment du nombre de valeurs étudiées

4 – on prend la racine carré de la variance →écart-type

plus les points s'écartent de la moyenne, plus l'écart-type est grand, indépendamment du nombre de valeurs. L'écart-type a la même unité que la variable étudiée.

écart-type (standard deviation)

écart-type (standard deviation) : racine carrée de la variance

L'écart-type est donné par la formule :

$$\sigma = \sqrt{\sum \left(x - \bar{x}\right)^2 / n}$$

exemple : contraction d'anneaux de bronches de rat

	Force (g)	F – mF (g)	$(F\text{-}mF)^2$ (g^2)
	1,14596	-0,07	0,00459
	1,0461	-0,17	0,0281
	.....	.....	.....
mF	1,21374	$\Sigma(F\text{-}mF)^2$	5,849

5,849/29 = 0,202 (g^2) → variance

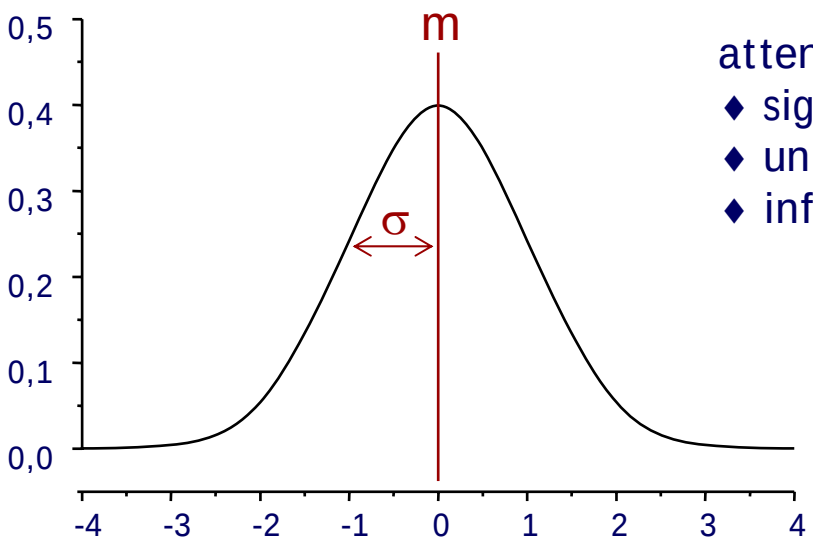
√(5,849/29) =0,449 (g) → écart-type

écart-type (standard deviation)

écart-type (standard deviation) : racine carrée de la variance

L'écart-type est donc donné par la formule : $\sigma = \sqrt{\sum \left(x - \bar{x}\right)^2 / n}$

cas particulier : loi normale



attention :

- ♦ signification de l'écart-type
- ♦ unité de l'écart-type
- ♦ influence de changement de variable

intérêts et limites :

♦ Après standardisation, permet de comparer la position de plusieurs variables entre elles, même si les unités de mesure de ces variables sont différentes.

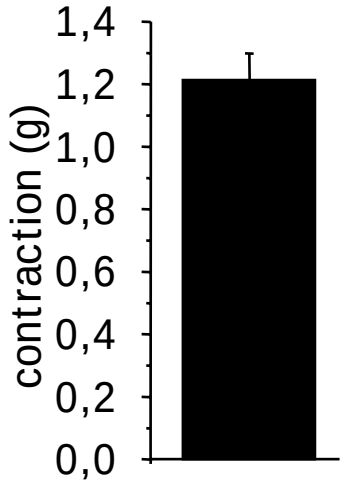
♦ quasiment la seule mesure de la dispersion utilisée

statistiques descriptives expression des données

- population étudiée
- variable(s) étudiée(s) et unité(s)
- taille de la population
- mesure de la tendance centrale (moyenne le plus souvent)
- mesure de la dispersion (écart-type le plus souvent)

exemple :
« L'étude a portée sur la contraction d'anneaux de bronches de rats. La contraction a été déterminée par la mesure de la force générée par les anneaux, en g. Les valeurs sont données sous la forme moyenne ± écart-type, avec n = nombre d'anneaux étudiés.
La force mesurée était de 1,21 ± 0,08 g (n = 29) »

figure 1 : mesure de la contraction d'anneaux de bronches de rats (en g). La colonne noire est la moyenne de 29 anneaux. La barre d'erreur représente l'écart-type.



estimation des caractéristiques d'une population à partir d'un échantillon

- fréquence de distribution
- moyenne et écart-type de la population

précision de l'estimation
intervalle de confiance

comparaison des différences entre plusieurs populations, à partir d'échantillons

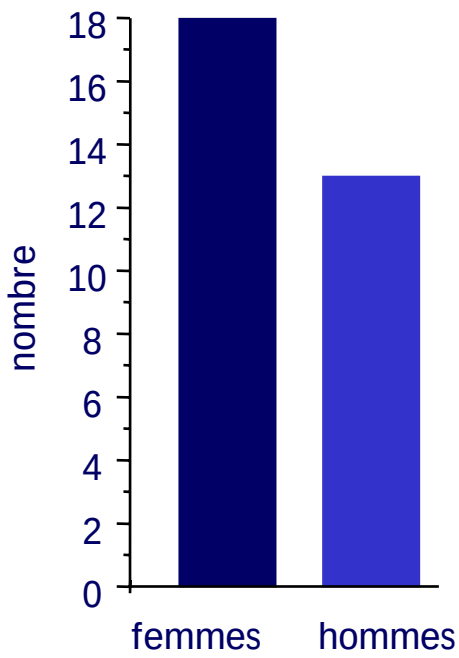
- comparaison à une population théorique
- comparaison de plusieurs (2 ou plus) échantillons entre eux

tests statistiques
estimation des erreurs
risque de première espèce (a)
risque de deuxième espèce (b)

variables qualitatives *estimation de la fréquence de distribution*

la fréquence estimée de la variable dans la population est la fréquence observée dans l'échantillon

exemple : répartition hommes/femmes dans un échantillon d'une population



échantillon (mesure) :
n = 31
femmes = 18
homme = 13
fréquence (observée)
femmes = 58,06 %
homme = 41,94 %

population (estimation) :
femmes = 58,06 %
homme = 41,94 %

variables quantitatives *estimation de la moyenne*

la moyenne estimée de la variable dans la population est la moyenne observée dans l'échantillon

exemple : mesure de la fréquence cardiaque sur un échantillon de 31 étudiants de l'UFR des Sciences de la vie

moyenne de l'échantillon (mesurée) : 86 battements/min

« la fréquence cardiaque moyenne mesurée sur un échantillon de 31 étudiants de l'UFR des Sciences de la vie était de 86 battements/min »

moyenne de la population (estimée) : 86 battements/min

« la fréquence cardiaque moyenne estimée des étudiants de l'UFR des Sciences de la vie est de 86 battements/min »

variables quantitatives

estimation de l'écart-type

l'écart-type estimé de la variable dans la population est l'écart-type observé dans l'échantillon augmenté d'un facteur de correction

→ écart-type estimé de la population ~> écart-type calculé de l'échantillon

écart-type estimé de la population : S

écart-type calculé de l'échantillon : σ

n-1 : degré de liberté

$$S = \sqrt{\frac{n}{n-1}}\sigma$$

exemple : mesure de la fréquence cardiaque sur un échantillon de 31 personnes

moyenne de l'échantillon (mesurée) : 86 battements/min

écart-type de l'échantillon (mesuré) : 13,04 battements/min

moyenne de la population (estimée) : 86 battements/min

écart-type de la population (estimé) : 13,25 battements/min

précision de l'estimation

principes généraux

variable quantitative : la précision de l'estimation de la moyenne d'une variable à partir d'un échantillon dépend de la fluctuation de la moyenne de l'échantillon

Moins, d'un échantillon à un autre, la valeur moyenne fluctue, plus grande est la précision de l'estimation de la moyenne de la population.

exemple : on mesure 2 variables A et B sur une série d'échantillons de plusieurs individus. On répète ces mesures sur 6 échantillons différents. On obtient les valeurs suivantes :

n° lot	A	B
1	12,36	18,94
2	10,10	8,93
3	7,28	6,51
4	7,90	9,36
5	10,16	17,70
6	8,99	4,16
moyenne	9,47	10,93
écart-type de la moyenne	1,67	5,50

→la fluctuation est plus faible pour la variable A. La précision de son estimation à partir d'un échantillon est plus grande que pour B

précision de l'estimation

principes généraux

variable qualitative : la précision de l'estimation de la fréquence d'une variable à partir d'un échantillon dépend de la fluctuation de la fréquence de l'échantillon

Moins, d'un échantillon à un autre, la valeur de la fréquence fluctue, plus grande est la précision de l'estimation de la fréquence de la population.

exemple : on mesure 2 variables A et B sur 2 séries d'échantillons. Pour chaque variable, on répète ces mesures sur 7 échantillons différents. On obtient les valeurs suivantes :

→la fluctuation de la fréquence est plus faible pour la variable B. La précision de son estimation à partir d'un échantillon est plus grande que pour A.

n° Lot	A	n° Lot	B
1	40 %	1	40 %
2	100 %	2	53 %
3	60 %	3	47 %
4	60 %	4	73 %
5	20 %	5	60 %
6	80 %	6	33 %
7	40 %	7	53 %
moyenne	57 %	moyenne	51 %
écart-type de la moyenne	25 %		12 %

précision de l'estimation

principes généraux

variable quantitative : la précision de l'estimation de la moyenne à partir d'un échantillon dépend de la fluctuation de la moyenne de l'échantillon

Moins, d'un échantillon à un autre, la valeur moyenne fluctue, plus grande est la précision de l'estimation de la moyenne de la population.

variable qualitative : la précision de l'estimation de la fréquence à partir d'un échantillon dépend de la fluctuation de la moyenne de l'échantillon

Moins, d'un échantillon à un autre, la valeur de la fréquence fluctue, plus grande est la précision de l'estimation de la fréquence de la population.

La précision de l'estimation de la moyenne ou de la fréquence d'une variable dépend de l'écart-type de la moyenne de la variable.

précision de l'estimation

principes généraux

la précision de l'estimation de la moyenne ou de la fréquence d'une variable à partir d'un échantillon dépend de la fluctuation de la moyenne de l'échantillon

- la fluctuation de la moyenne (ou de la fréquence) entre plusieurs échantillons dépend :

  de la fluctuation individuelle de la variable

  plus l'écart-type est petit, plus la précision est bonne

  de la taille de l'échantillon

  plus l'échantillon est gros, plus la précision est bonne

- la fluctuation de la moyenne (ou de la fréquence) est mesurée par l'écart-type de la moyenne

précision de l'estimation

principes généraux

la précision de l'estimation de la moyenne ou de la fréquence d'une variable à partir d'un échantillon dépend de la fluctuation de la moyenne de l'échantillon

♦ la fluctuation de la moyenne (ou de la fréquence) entre plusieurs échantillons dépend :

de la fluctuation individuelle de la variable

de la taille de l'échantillon

♦ la fluctuation de la moyenne (ou de la fréquence) est mesurée par l'écart-type de la moyenne

pb : comment calculer l'écart-type de la moyenne

- répéter l'expériences sur plusieurs échantillons
- estimer l'écart-type de la moyenne sur un seul échantillon

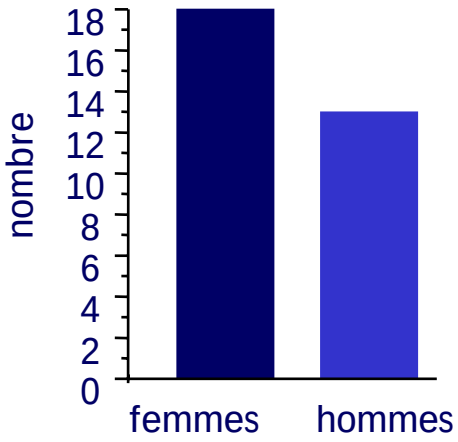
loi de probabilité de la moyenne ?

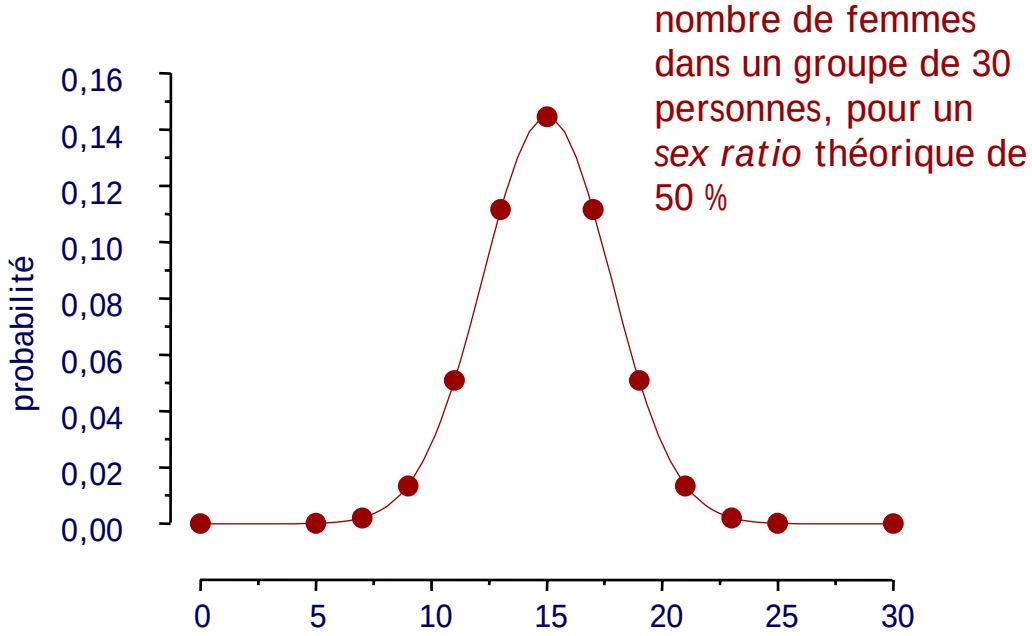
précision de l'estimation *loi de probabilité de la moyenne*

♦ si la loi de probabilité des variables des individus suit une loi normale, alors la loi de probabilité de la moyenne est également une loi normale

♦ si la loi de probabilité des variables des individus n'est pas une loi normale, la loi de probabilité de la moyenne est une loi normale, si la taille de l'échantillon est assez grande (n > 30)

ex : loi binomiale



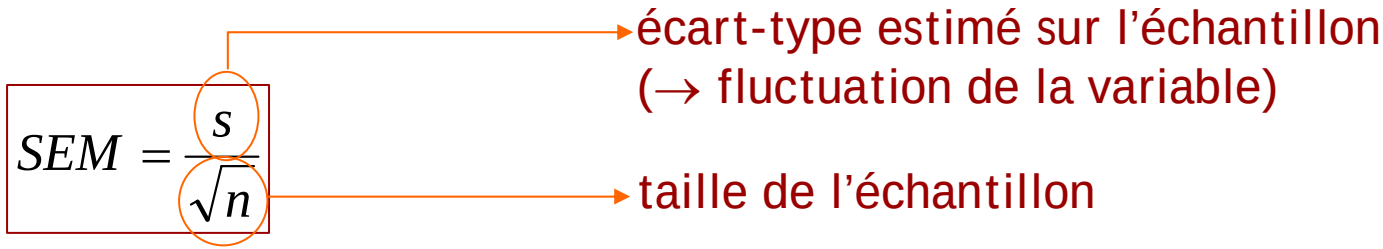


nombre de femmes dans un groupe de 30 personnes, pour un *sex ratio* théorique de 50 %

précision de l'estimation *écart-type de la moyenne (SEM)*

variable quantitative

estimation de la fluctuation de la moyenne :
écart-type de la moyenne = standard error of the mean (SEM)



exemple : fréquence cardiaque
moyenne de la population (estimée) : 86 battements/min
écart-type de la population (SD) (estimé) : 13,25 battements/min

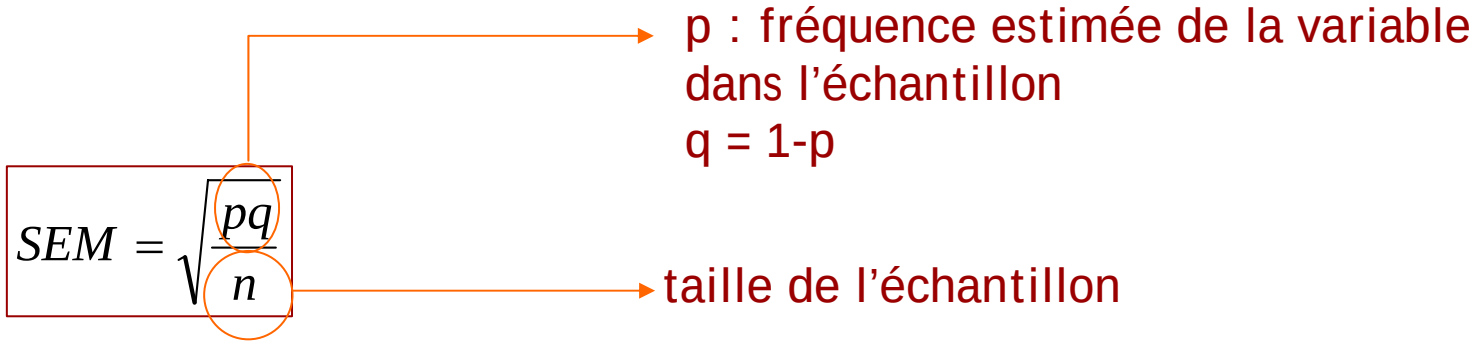
SEM = 3,38 battements/min

NB : la précision dépend de la taille de l'échantillon, pas de la taille de la population

précision de l'estimation *écart-type de la moyenne (SEM)*

variable qualitative

estimation de la fluctuation de la moyenne :
écart-type de la moyenne = standard error of the mean (SEM)



exemple : sex ratio

population (estimation) : femmes = 58,06 % homme = 41,94 %

SEM = 8,86 %

NB : la précision dépend de la taille de l'échantillon, pas de la taille de la population

précision de l'estimation *intervalle de confiance (confidence interval)*

intervalle autour de la moyenne calculée de l'échantillon dans lequel la moyenne de la population a une probabilité donnée de se trouver.

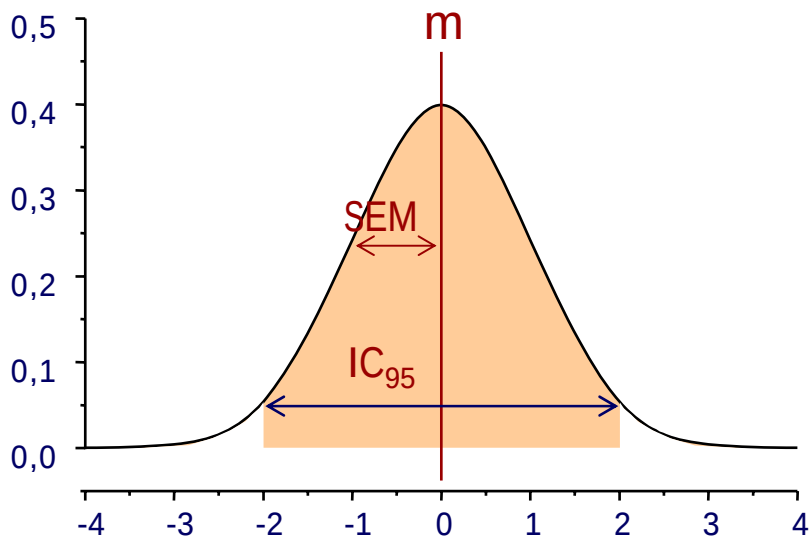
exemple : intervalle de confiance à 95 % : la valeur moyenne de la population dont est issu l'échantillon a 95 chances sur 100 de se trouver dans l'intervalle.

dépend :

- de la SEM
- du % de confiance voulu
- du degré de liberté (ddl)

ex :
ddl > 30
intervalle de confiance à 95 % =
m ± 1,96 SEM

loi normale



risque alpha : probabilité que la valeur de la population (vraie valeur) soit en dehors de l'intervalle de confiance

précision de l'estimation *intervalle de confiance (confidence interval)*

variable quantitative

exemple : fréquence cardiaque mesurée sur 31 personnes (n = 31)

moyenne de l'échantillon (mesurée) : 86 battements/min
écart-type de l'échantillon (mesuré) : 13,04 battements/min

moyenne de la population (estimée) : 86 battements/min
écart-type de la population (estimé) : 13,25 battements/min
SEM = 3,38 battements/min
n = 31 (ddl = 30)

intervalle de confiance à 95 % (IC 95 %) =
86 ± 6,7 battements/min

précision de l'estimation *intervalle de confiance (confidence interval)*

variable quantitative

calcul avec les fonctions d'Excel ou d'OpenOffice :

moyenne de l'échantillon (mesurée) : fonction « MOYENNE »
écart-type de l'échantillon (mesuré) : fonction « ECARTTYPEP »

moyenne de la population (estimée) : fonction «MOYENNE »
écart-type de la population (estimé) : fonction «ECARTTYPE»
SEM = pas de fonction disponible :
→ calculer en utilisant la formule :
= ECARTTYPE(données)/RACINE(taille)

intervalle de confiance à 95 % (IC 95 %) :
fonction « INTERVALLE.CONFIANCE »
indiquer alpha : 0,05
indiquer l'écartype estimé
indiquer la taille de l'échantillon

vrai si n>30

précision de l'estimation *intervalle de confiance (confidence interval)*

variable qualitative

exemple : sex ratio mesuré sur un échantillon de 31 personnes

sex ratio de l'échantillon (mesure) :
femmes = 18 (58,06 %) homme = 13 (41,94 %)

sex ratio de la population (estimation) :
femmes = 58,06 % homme = 41,94 %
SEM = 8,86 %
(n = 31 ; ddl = 30)

intervalle de confiance à 95 % (IC 95 %) =
femmes = 58,06 % ± 17.37 % hommes = 41,94 ± 17.37 %

vrai si n>30

précision de l'estimation *intervalle de confiance (confidence interval)*

variable qualitative

calcul avec les fonctions d'Excel ou d'OpenOffice :

sex ratio de l'échantillon (mesure) :
femmes = 18 homme = 13 n = 31 (ddl = 30)
calcul des proportions : femmes : 0,5806 (p) hommes : 0,4194 (q = 1-p)

sex ratio de la population (estimation) :
femmes = 58,06 % homme = 41,94 %
SEM = pas de fonction disponible :
→ calculer en utilisant la formule :
= RACINE((p*(1-p)/n))

intervalle de confiance à 95 % (IC 95 %) =
fonction « INTERVALLE.CONFIANCE »

NB : calcul préliminaire : écart-type estimé = RACINE((p*(1-p))
indiquer alpha : 0,05
indiquer l'écartype estimé (voir calcul préliminaire)
indiquer la taille de l'échantillon

vrai si n>30

principe des tests

♦ Les statistiques inférentielles permettent d'assigner une probabilité à l'obtention d'un résultat pour une hypothèse donnée.

exemple : intervalle de confiance à 95 %
on fait l'hypothèse que la moyenne ou la fréquence d'une variable se trouve dans l'intervalle de confiance, dont on a calculé que la probabilité était de 0,95 (95 %).
(exemple des sondages d'opinion)

♦ Si cette probabilité est trop faible, on rejette l'hypothèse.

exemple : on rejette l'hypothèse que la moyenne ou la fréquence de la variable est en dehors de l'intervalle de confiance.

→ application aux comparaisons statistiques

principe des tests — *l'hypothèse nulle*

hypothèse nulle (null hypothesis)

Le principe des tests statistiques est de postuler l'hypothèse nulle : on fait l'hypothèse que les différences observées – entre des valeurs observées ou entre une valeur observées et une valeur théorique – est due aux fluctuations d'échantillonnage.

exemple : effet de la présence de calcium extracellulaire sur la contraction d'anneau de bronche.
hypothèse nulle : le calcium extracellulaire n'a pas d'effet.
= les deux échantillons d'anneaux de bronches proviennent de la même population

le test statistique calcule la probabilité que les différences de valeur de contraction entre les deux échantillons soient dues aux fluctuations d'échantillonnage dans une même population

principe des tests — *l'hypothèse nulle*

hypothèse nulle (null hypothesis)

Le principe des tests statistiques est de postuler l'hypothèse nulle : on fait l'hypothèse que les différences observées – entre des valeurs observées ou entre une valeur observées et une valeur théorique – est due aux fluctuations d'échantillonnage.

conditions de rejet de l'hypothèse nulle

Si la probabilité de l'hypothèse nulle est trop faible, on la rejette, et on accepte l'hypothèse non nulle : les échantillons comparés proviennent de populations différentes.

exemple : le calcium extracellulaire a un effet sur la contraction
(les anneaux avec calcium proviennent d'une population différente des anneaux sans calcium)

On dit alors qu'il existe une différence statistiquement significative.

principe des tests

l'hypothèse nulle

exemples :

a) comparaison de la répartition homme/femme observée et de la valeur théorique du sex ratio de 50 %

sex ratio mesuré sur un échantillon de 31 étudiants de biologie

sex ratio de l'échantillon (mesure) :
femmes = 18 (58,06 %) homme = 13 (41,94 %)

question : y a-t-il significativement plus de femmes que d'hommes en biologie, par rapport à l'ensemble de la population?

a) formulation de l'hypothèse nulle : la population théorique dont le groupe d'étudiants est un échantillon représentatif n'est pas différente de la population « générale » dont on connaît les valeurs théoriques :
sex ratio de 50 %

principe des tests *l'hypothèse nulle*

exemples :

b) comparaison des fréquences cardiaques de groupes d'étudiants à la fréquence théorique « normale » de 70 battements/minute

exemple : fréquence cardiaque mesurée sur 31 personnes (n = 31)

moyenne : 86 battements/min
écart-type (estimé) : 13,25 battements/min
SEM = 3,38 battements/min
n = 31
(IC 95 %) = 86 ± 6,7 battements/min

question : la fréquence cardiaque des étudiants en biologie est-elle significativement différente de celle de l'ensemble de la population?

b) formulation de l'hypothèse nulle : la population théorique dont le groupe d'étudiants est un échantillon représentatif n'est pas différente de la population « générale » dont on connaît les valeurs théoriques : fréquence cardiaque de 70 batt/min.

principe des tests *l'hypothèse nulle*

exemples :

c) comparaison des fréquences cardiaques des hommes et des femmes dans un groupe d'étudiants.

exemple : fréquence cardiaque mesurée sur 31 étudiants en biologie (n = 31), 18 femmes et 13 hommes.

On calcule la fréquence cardiaque chez les hommes et chez les femmes

question : la fréquence cardiaque des étudiantes en biologie est-elle significativement différente de celle des étudiantes en biologie ?

c) formulation de l'hypothèse nulle : la population théorique dont les étudiants masculins sont un échantillon représentatif est identique à la population théorique dans les étudiants féminins sont un échantillon représentatif.

principe des tests

conditions de rejet de l'hypothèse nulle

Si la probabilité de l'hypothèse nulle est trop faible, on la rejette, et on accepte l'hypothèse non nulle : les échantillons comparés proviennent de populations différentes.

Il existe une différence statistiquement significative.

Par convention, on fixe en général le seuil de signification à 5 %
$p<0,05$: différences statistiquement significatives
$p<0,01$: différences statistiquement hautement significatives
$p<0,001$: différences statistiquement très hautement significatives

le seuil de signification est déterminé avant d'effectuer le test ; le degré de signification est déterminé par le test (= probabilité de rejeter l'hypothèse nulle si elle est vraie).

La différence est significative si le degré de signification est inférieur au seuil de signification.

principe des tests *conditions de rejet de l'hypothèse nulle*

exemple :
« On a mesuré l'effet de la présence de calcium extracellulaire sur la contraction d'anneaux de bronche. Les valeurs, exprimées en % d'une valeur de référence, sont données sous la forme : moyenne ± SEM (n = taille de l'échantillon). Les différences sont considérées comme significatives si $P<0,05$.

résultats:
En présence et en absence de calcium extracellulaire, la contraction était de 13,66 ± 1,53 (n = 8) et de 7,95 ± 1,71 (n = 7), respectivement. Le degré de signification (*P*) était de 0,029. »

question : la contraction d'anneaux de bronches dépend-elle du calcium extracellulaire ?

principe des tests *risques d'erreur*

- *risque α (risque de 1re espèce) (type 1 error)*

risque de rejeter l'hypothèse nulle si est est vraie.

Il est connu : seuil (à priori) ou degré (à postériori) de signification du test

- *risque β (risque de 2e espèce) (type 2 error)*

risque d'accepter l'hypothèse nulle alors qu'elle est fausse.

Le risque de 2e espèce correspond au défaut de puissance d'un test

Il est en général indéterminé (on ne connaît pas les caractéristiques des populations théoriques).

principe des tests *risques d'erreur*

risque a (risque de 1re espèce) (type 1 error)
risque de rejeter l'hypothèse nulle si est est vraie.

risque b (risque de 2e espèce) (type 2 error)
risque d'accepter l'hypothèse nulle alors qu'elle est fausse.

Les deux types de risques sont antagonistes.

Si on diminue le risque de 1re espèce, on augmente le risque de 2e espèce.

Étant donné que le risque de 2e espèce n'est pas connu – à la différence du risque de 1re espèce – en absence de différence significative, on ne peut pas conclure à l'absence de différence, car on ne contrôle pas le risque d'erreur attaché à cette conclusion.

Il y a une différence souvent oubliée entre ne pas conclure qu'il existe une différence, et conclure qu'il n'existe pas de différence.

méthodologie

→ poser une question

→ émettre une hypothèse

→ élaborer une procédure expérimentale de test de l'hypothèse

(NB : test ≠ confirmation)
« un protocole expérimental n'est pas une manière de prouver qu'une explication donnée est correcte, mais plutôt un système par lequel les explications alternatives sont éliminées ».

Lumley & Benjamin. *Research: some grounds rules*

= critère de réfutation
(Karl Popper. *La logique de la découverte scientifique, Conjectures et réfutations*)

procédure expérimentale

→ **constitution d'un ou de plusieurs échantillons**

« les techniques statistiques dépendent de la sélection au hasard de sujets (échantillon) dans une population définie. » Lumley & Benjamin

!attention au biais dans la constitution des échantillons!

→ **choix des procédures expérimentales**

→ **choix des procédures statistiques**

définition de l'hypothèse nulle
choix du seuil de signification
choix du test

« Le choix de la procédure statistique appropriée est une partie importante de la procédure expérimentale et ne devraît jamais être fait après la récolte des données. » Lumley & Benjamin

procédure expérimentale

→ **constitution d'un ou de plusieurs échantillons**

« les techniques statistiques dépendent de la sélection au hasard de sujets (échantillon) dans une population définie. » Lumley & Benjamin

!attention au biais dans la constitution des échantillons!

→ **choix des procédures expérimentales**

→ **choix des procédures statistiques**

définition de l'hypothèse nulle
choix du seuil de signification
choix du test

« Le choix de la procédure statistique appropriée est une partie importante de la procédure expérimentale et ne devraît jamais être fait après la récolte des données. » Lumley & Benjamin

choix du test

→ **types de variables**
qualitatives / quantitatives

→ **nombre de variables**

→ **taille de l'échantillon**

→ **loi de répartition**
« normale » ou non (+égalité des variances…)

→ **mesures répétées ou non / nombre de facteurs**

choix du test *liens entre variables qualitatives et quantitatives*

comparaison de deux moyennes

comparaison de deux moyennes observées

comparaison d'un moyenne observée à une moyenne théorique

options du test :

- comparaison d'un moyenne observée et d'une moyenne théorique (one population) ou de deux moyennes observées (two populations)

- mesures appariées (paired) ou non appariées (unpaired)

- comparaison unilatérale (one-tailed) ou bilatérale (two-tailed)

choix du test *liens entre variables qualitatives et quantitatives*

comparaison de deux moyennes

♦ les effectifs sont suffisamment grands (n>30) ou la loi de répartition est normale (faire éventuellement un test de normalité)
→ test *t* de Student
one population / two population
apparié/ non apparié

→ tests non paramétriques
options :
séries non appariées :
test *W* de Wilcoxon = test *U* de Mann et Whitney
test C1 de Fisher-Yates-Terry
séries appariées :
test T de Wilcoxon

♦ Les effectifs sont faibles et la répartition n'est pas normale (faire éventuellement un test de normalité)
→ tests non paramétriques

choix du test *liens entre variables qualitatives et quantitatives*

comparaison de plusieurs (≥ 2) moyennes

♦ étape 1 : on effectue une comparaison globale, pour tester l'existence d'une différence significative entre certains échantillons

♦ étape 2 : si l'étape 1 montre l'existence d'uen différence significative, on effetue des comparaisons 2 à 2 pour déterminer entre quels échantillons se trouvent ces différences.

→ tests « post-hoc »

NB: on ne doit pas faire de comparaisons 2 à 2 sans comparaison globale initiale

choix du test *liens entre variables qualitatives et quantitatives*

comparaison de plusieurs (≥ 2) moyennes

♦ la loi de répartition de probabilité est normale pour la variable mesurée (faire éventuellement un test de normalité)

→ Analyse de la variance (ANOVA)

→ tests non paramétriques

Kruskall-Wallis : non apparié

Friedman : appariée

options :

ANOVA à plusieurs facteurs

tests « post-hoc » :

Méthode de Bonferonni (test *t*)

Méthode de Tukey (test *t*)

Méthode de Dunnet

Méthode de Scheffé (test *F*)

choix du test *liens entre variables qualitatives et quantitatives*

comparaison de plusieurs (≥ 2) moyennes

♦ la loi de répartition de probabilité est normale pour la variable mesurée (faire éventuellement un test de normalité)
→ Analyse de la variance (ANOVA)
→ tests non paramétriques

♦ la loi de répartition de probabilité n'est pas normale pour la variable mesurée
→ tests non paramétriques

choix du test

liens entre variables qualitatives

- ♦ échantillon de taille normale (effectifs calculés > 5) :
→ test du χ^2 (chi^2).

- ♦ échantillon de taille réduite (effectifs calculés > 3) :
→ χ^2 corrigé (correction de Yates)

- ♦ échantillon de taille très réduite (effectifs calculés < 3) :
→« test exact »

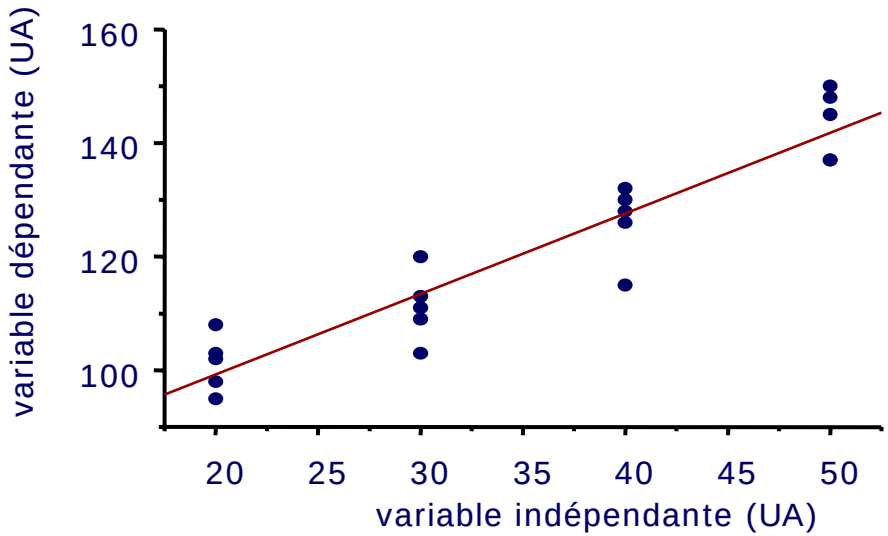
choix du test *liens entre 2 variables quantitatives*

exemple

Linear Regression
Y = A + B * X

Parameter	Value	Error
A	70,94	4,39668
B	1,416	0,11966

R	SD	N
0,94133	5,98312	20



♦ une des deux distributions liées au moins est normale avec une variance constante

→ test de corrélation (ou de régression)
détermine si la pente est statistiquement significative de 0

♦ si aucune des variables liées n'est normale et de variance constante (petits échantillons)

→ test non paramétrique de corrélation des rangs (test de Spearman)

choix du test

Que faire quand on ne sait pas quoi faire ?

demander à quelqu'un qui sait

faire appel à un statisticien
(au moment de concevoir les protocoles)

Statistiques : éléments de bibliographie

P. Lazar & D. Schwartz. Éléments de probabilités et statistiques, Flammarion, Paris, 1987.
petit livre de base, avec exercices, pour s'initier de manière pratique aux probabilités et statistiques (BU)

R. Salamon. Statistique médicale, Masson, Paris, 1988.
Petit livre de base contenant l'essentiel des notions en statistiques, et une introduction au calcul des probabilités (BU)

D. Schwartz. Méthodes statistiques à l'usage des médecins et des biologistes, 4e édition, Flammarion, Paris, 1994.
ouvrage français de référence (BU)

T. H. Wonnacot & R. J. Wonnacot. Statistique, 4e ed, Economica, Paris, 1991.
Ouvrage détaillé (900 p) sur la statistique en économie, gestion, sciences et médecine, avec exercices d'applications (BU)

J. S. P. Lumley & W. Benjamin. Resarch: some ground rules, Oxford University Press, Oxford, 1994.
guide pour savoir comment mener un travail de recherche. N'est pas consacré particulière-ment aux statistiques, mais une section est consacrée à l'analyse des résultats, avec une approche utilitaire des statistiques. (BU)

Statistiques : éléments de bibliographie

J. Fowler, L. Cohen & P. Jarvis. Practical statistics for field biology, Wiley, Chichester, 1998.
Bonne introduction aux statistiques en général, bien qu'axé plutôt sur les statistiques de biologie d'observation.

S. J. Gould, *L'éventail du vivant*, Seuil, Paris, 1997. (titre original : *Full House*)
ouvrage de vulgarisation sur l'analyse des tendances das l'évolution biologique, présente de manière claire les biais possibles et les pièdes à éviter dans l'analyse des répartitions asymétriques (en annexe, une introduction au jeu de base-ball).

D. M. Raup. *De l'extinction des espèces*, Gallimard, Paris, 1993 (titre original : *Extinction. Bad genes or bad luck?)*
Par un spécialiste de paléontologie statistique, l'analyse de la part du hasard dans les extinctions. Contient une présentation claire de quelques questions d'ordre statistique.