Tut' Rentrée Biostatistiques 2012-2013

Nice Tutorat
FACULTE DE MEDECINE

Cours 2 :

Initiation à la biostatistique.

Vos tuteurs de Biostat' !

Claire
(Cleaar)

Tom
(Tracky)

Robin
(attention83)

Plan du cours

I. Introduction

1- Quelques définitions

2- La démarche statistique

II. Statistiques descriptives

1- Les variables

2- L'échantillonage

3- L'estimation

III. Statistiques déductives

1- Les hypothèses

2- La démarche

3- Les tests

4- Les risques

I-2 Quelques définitions

Variable (ou donnée) : résultat de **l'observation** ou de la **mesure** d'un caractère
Paramètre : **information résumée** d'une variable

Série : **ensemble d'objets** de même nature mais de caractéristiques différents
Population : **tous** les individus qu'on voudrait étudier -> exhaustif
Echantillon : **extrait** de la population -> effectif limité

Statistiques descriptives ≠ Statistiques déductives

Variabilité : -> physiologique
-> hasard

I-1 La démarche statistique

Question à résoudre…. Hasard ou pas ?

Plan du cours

I. Introduction

1- Quelques définitions

2- La démarche statistique

II. Statistiques descriptives

1- Les variables

2- L'échantillonage

3- L'estimation

III. Statistiques déductives

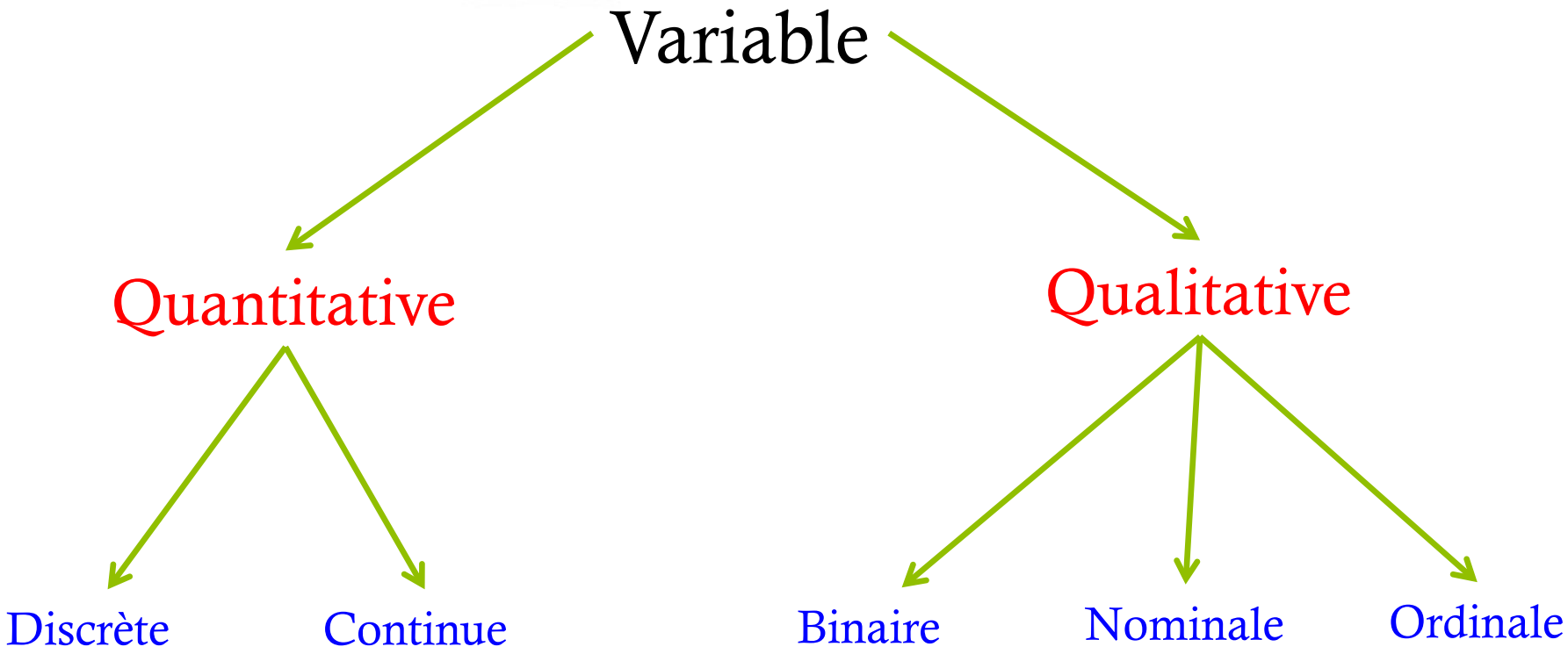
1- Les hypothèses

2- La démarche

3- Les tests

4- Les risques

II-1 Les variables



Application

Donnez la nature des variables suivantes :	Réponse :
Taille d'une personne (en cm) : 145	Quantitative discrète
Taille d'une personne (en m) : 1,45	Quantitative continue
Sexe : Féminin	Qualitative binaire
Age à l'admission : 34 ans	Quantitative discrète
Degré de douleur : faible/modéré/fort	Qualitative ordinale
Consommation de médicaments : 0-5 / 5-10 / 10-15 / + de 15	Qualitative ordinale

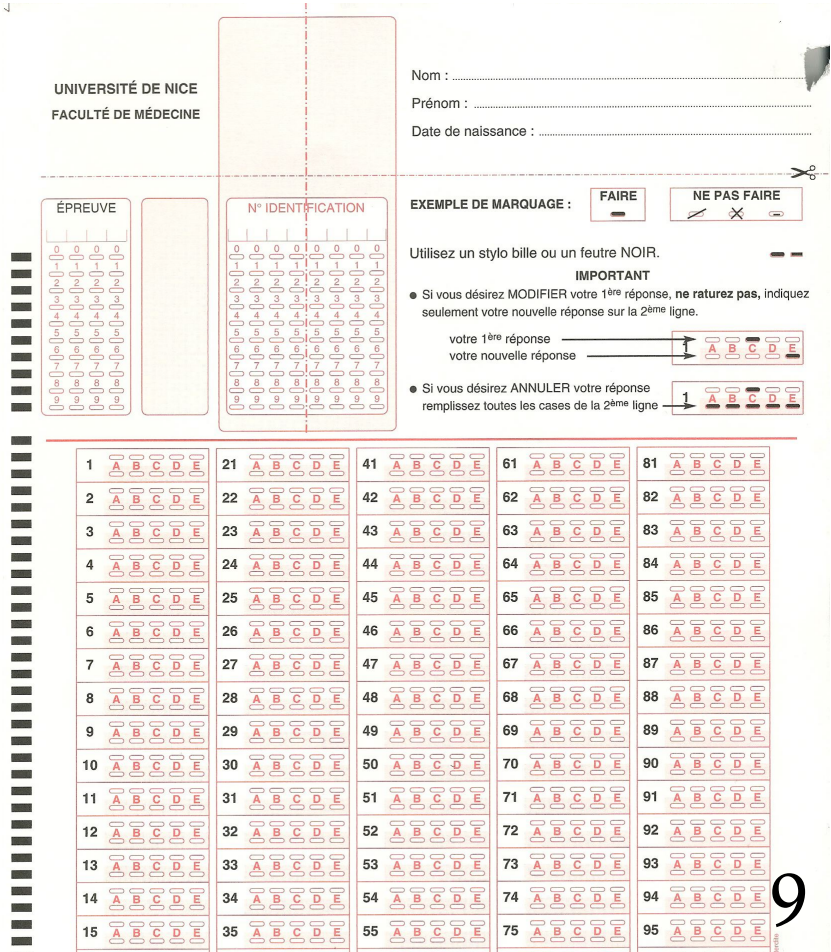
Exemple de QCM :

QCM : Donnez la ou les bonnes réponses.

A) La biostat est la meilleure des matières.
B) La physique est la meilleure des matières.
C) La chimie est la meilleure des matières.
D) La BDR est la meilleure des matières.
E) Aucune de ces réponses n'est correcte.

Une proposition en plus ou en moins c'est FAUX !

CORRECTION : A



Exemple de QCM :

QCM : Donnez les propositions justes concernant les types de variable.

A) Le nombre de tumeur pulmonaire est une variable quantitative discrète.
B) La présence de fièvre est une variable quantitative continue.
C) Le nombre de molécule de glucose dans le sang d'un patient est une variable quantitative ordinale.
D) La température du malade est une variable qualitative binaire.
E) Aucune de ces réponses n'est correcte.

CORRECTION : A

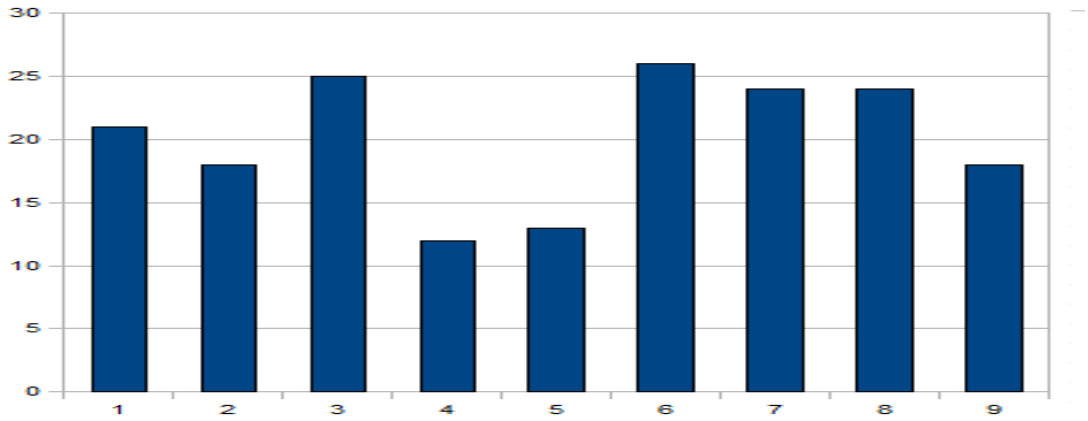
II-1 Les variables

<u>Comment les représenter ?</u>

-pour les qualitatives :
TABLEAU/HISTOGRAMME/POURCENTAGE

-pour les quantitatives : TABLEAU / HISTOGRAMME

Degré de satisfaction	Nb mères	%
Très insatisfait	6	3,9%
Plutôt insatisfait	18	11,7%
Plutôt satisfait	48	31,2%



II-1 Les variables

Comment résumer les variables quantitatives ?

→ Moyenne

→ Variance

→ Médiane

→ Quartiles

Application

On s'intéresse à la taille (en cm) de la population française.
On dispose d'un échantillon représentatif de la population de 16 personnes (n=16).

173 ; 156 ; 183 ; 134 ; 155 ; 186 ; 120 ; 149 ; 142 ; 173 ; 181 ; 134 ; 162 ; 136 ; 99 ; 201

MOYENNE :

173 +156 + 183 + 134 + 155 + 186 + 120 + 149 + 142 + 173 + 181 + 134 + 162 + 136 + 99 + 201 = 2484

2484 / n
= 2484 / 16
= 155,25

Application

On s'intéresse à la taille (en cm) de la population française.
On dispose d'un échantillon représentatif de la population de 16 personnes (n=16).

173 ; 156 ; 183 ; 134 ; 155 ; 186 ; 120 ; 149 ; 142 ; 173 ; 181 ; 134 ; 162 ; 136 ; 99 ; 201

MEDIANE :

☞Il faut classer les valeurs !
99 ; 120 ; 134 ; 134 ; 136 ; 142 ; 149 ; 155 ; 156 ; 162 ; 173 ; 173 ; 181 ; 183 ; 186 ; 201

n-> impair : (n+1)/2 ème valeur

n -> pair : moyenne de la (n)/2 et (n)/2 +1 ème valeur
→ (155 + 156)/2
→ 155,5

La médiane n'est pas obligatoirement égale à la moyenne

Application

On s'intéresse à la taille (en cm) de la population française.
On dispose d'un échantillon représentatif de la population de 16 personnes (n=16).

173 ; 156 ; 183 ; 134 ; 155 ; 186 ; 120 ; 149 ; 142 ; 173 ; 181 ; 134 ; 162 ; 136 ; 99 ; 201

QUARTILES :

☞Il faut classer les valeurs !
99 ; 120 ; 134 ; 134 ; 136 ; 142 ; 149 ; 155 ; 156 ; 162 ; 173 ; 173 ; 181 ; 183 ; 186 ; 201

1er quartile : 25% de n → 0,25 x 16 = 4
2ème quartile : 50% de n → 0,50 x 16 = 8
3ème quartile : 75% de n → 0,75 x 16 = 12

2ème quartile = médiane !

☞Si la valeur n'est pas un entier on fait la moyenne des rangs encadrés !

Application

On s'intéresse à la taille (en cm) de la population française.
On dispose d'un échantillon représentatif de la population de 16 personnes (n=16).

173 ; 156 ; 183 ; 134 ; 155 ; 186 ; 120 ; 149 ; 142 ; 173 ; 181 ; 134 ; 162 ; 136 ; 99 ; 201

VARIANCE :

Pas besoin de la calculer ☺

Ecart-type = racine carré de la variance

+ la variance est grande, + les valeurs sont dispersées (= hétérogènes)

II-2 L'échantillonnage

☞ Important !

La population -> inconnue

L'échantillon -> connu

II-2 L'échantillonnage

Comment établir cet échantillon ?

REPRESENTATIVITÉ

BIAIS

LA SOLUTION :

Tirage au sort
= Randomisation

ESTIMATION

II-2 L'échantillonnage

Pour éviter la présence de biais lors de l'échantillonnage, et donc garantir la représentativité de l'échantillon il faut :

-définir précisément les **caractéristiques de la population**

-effectuer un **tirage au sort** (TAS)

Application

On veut établir la moyenne des poids de la population mondiale, hommes et femmes confondus. Cette population étant inaccessible à une étude statistique directe, il faut constituer un échantillon **REPRESENTATIF** de cette population.

Constituer un échantillon constitué d'hommes et femmes des USA vous semble une bonne méthode ?
NON -> Erreur d'échantillonnage.

Que faudrait-il faire alors ?
Pour un bon échantillonnage : **TAS (+++)**

II-3 L'estimation

L'estimation permet de déterminer une **grandeur** (ex: le poids) concernant une **population** à travers l'étude d'un **échantillon représentatif** de la population étudiée.

II-3 L'estimation

Dans le cas général :

1) Disposer d'un **échantillon représentatif.**

2) Calculer le **paramètre sur l'échantillon.**

3) Calculer **l'intervalle de confiance.**

➔ ESTIMATION DU PARAMETRE DANS LA POPULATION

II-3 L'estimation

L'intervalle de confiance :

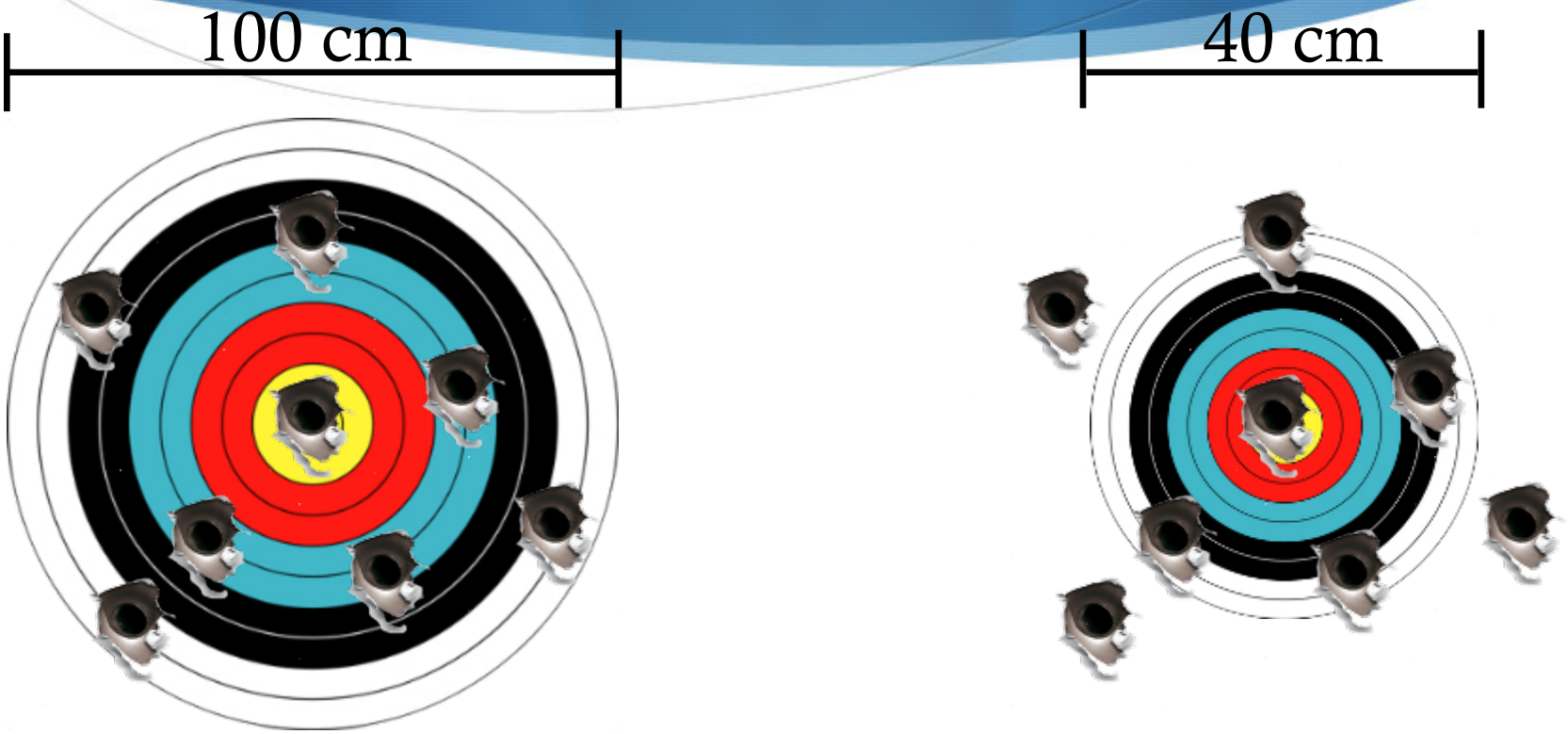
Comme l'estimation est inconnue au niveau de la population, il y a un risque d'erreur. C'est le **risque** α .

Mais c'est le paramètre ε qui intervient dans la formule.
On le lit dans la table de l'écart réduit, en fonction du risque α que l'on a choisi.

Le compromis universel est de 5% pour α .

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

II-3 L'estimation



+ α est petit, + ε est grand,
→ plus l'intervalle de confiance est grand.

II-3 L'estimation

Dans le cas des variables quantitatives

1) Disposer d'un **échantillon représentatif.**

2) Calculer le **paramètre sur l'échantillon.**

3) Calculer **l'intervalle de confiance.**

➔ ESTIMATION DU PARAMETRE DANS LA POPULATION

II-3 L'estimation

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

☞ Attention aux notations !

Dans l'échantillon :	Dans la population :
n : effectif	N : effectif
m : moyenne	μ : moyenne
s : écart-type	σ : écart-type

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

II-3 L'estimation

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

☞ Calcul de la valeur de la précision (=largeur de l'IC)

i =

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

II-3 L'estimation

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

+ n est grand,
+ i (valeur de la précision) est petit,
→ + la précision est grande

Valeur de la précision ≠ précision !!!

Application

On chercher à calculer la moyenne de taille des nains de jardin de France.

Population visée : Nains de jardins de France
Echantillon : 100 nains de jardins de France choisis par **TAS**.

Données : moyenne (m) de l'échantillon: 52cm, l'écart-type (s) est évalué à 8cm, on prend un risque dans l'estimation de μ de 5%, l'écart réduit correspondant est donc de 1,96 (dans cet exemple, on arrondit à 2).

Calcul de la précision de l'IC :
$i=(2x\ 8)/10=1,6$

Calcul de la précision de l'IC :
Ainsi μ appartient à l'IC [52 - 1,6 ; 52 +1,6], d'où [50,4 ; 53,6]

II-3 L'estimation

Dans le cas des variables qualitatives

1) Disposer d'un **échantillon représentatif.**

2) Calculer le **paramètre sur l'échantillon.** (pourcentage)

3) Calculer **l'intervalle de confiance.**

➔ ESTIMATION DU PARAMETRE DANS LA POPULATION

II-3 L'estimation

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

☞ Attention aux notations !

Dans l'échantillon :
p_{obs}: pourcentage
s : écart-type

Dans la population :
p : pourcentage

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

Application

Donnez l'estimation, sous forme d'intervalle de confiance, du pourcentage de prénom de moins de 4 lettres au niveau de la population française.

On souhaite étudier la répartition des prénoms en fonction de leur longueur dans la population française, avec un risque d'erreur de 5%. On constitue par tirage au sort un échantillon représentatif de la population française. Après étude de l'échantillon, on constate :

56% de prénoms de moins de 4 lettres, ex : Eva
44% de prénoms de plus de 4 lettres, ex : Alexis
Un écart-type de 2,5

$$P \in [\, 0{,}56 - 1{,}96 \times 2{,}5 \; ; \; 0{,}56 + 1{,}96 \times 2{,}5 \,]$$

Synthèse

Données quantitatives

Moyenne : m
Ecart-Type : s

Estimation de la moyenne inconnue dans la population :

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

Données qualitatives

% au niveau de la population : p_0
Ecart-Type : s

Estimation du pourcentage inconnu dans la population :

$$p \in [p_{obs} \pm \varepsilon s]$$

ε lu dans la table → risque d'erreur accepté

II-3 L'estimation

La **loi de Gauss** est une loi très utilisée en statistique car elle permet de décrire une population, dont l'effectif est supérieur à 30 (**n>30**).

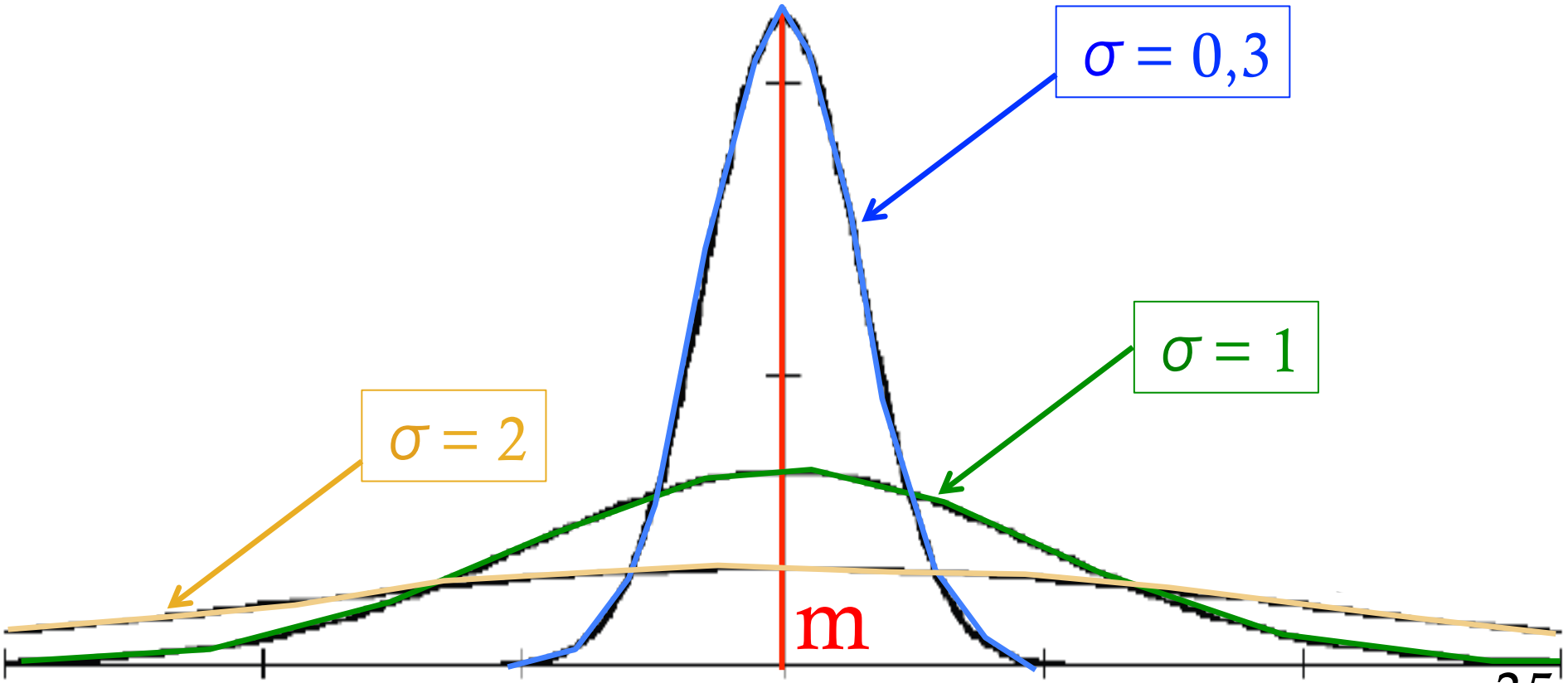
<u>Elle permet de visualiser :</u>

-moyenne

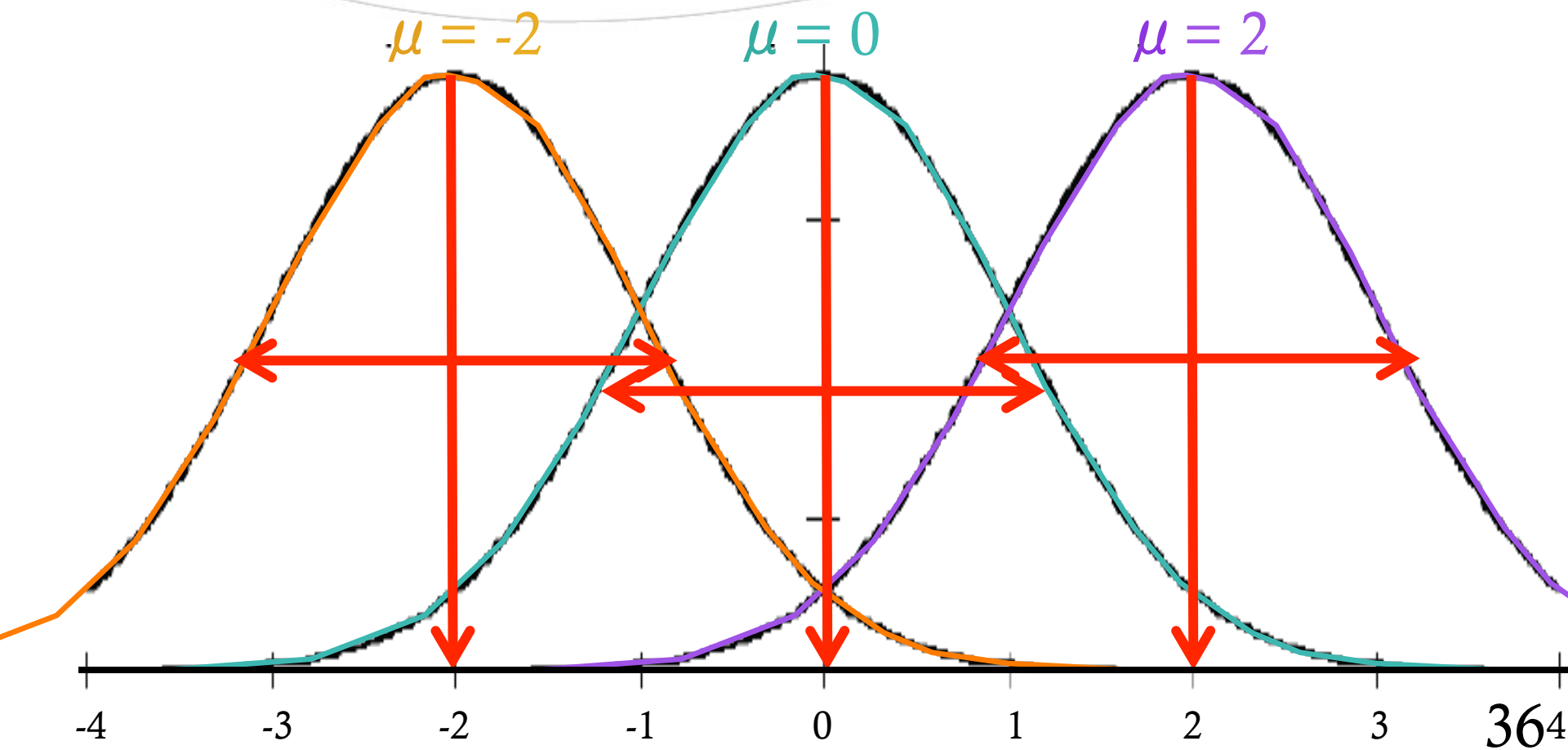
-écart-type

-intervalle de confiance

II-3 L'estimation



II-3 L'estimation



Plan du cours

I. Introduction

1- Quelques définitions

2- La démarche statistique

II. Statistiques descriptives

1- Les variables

2- L'échantillonnage

3- L'estimation

III. Statistiques déductives

1- Les hypothèses

2- La démarche

3- Les tests

4- Les risques

III-1 Les hypothèses

Quand les utilises-t-on ?

Quand on veut **comparer 2 groupes** pour un caractère donné.

2 hypothèses :

→ H0 = Hypothèse nulle ⇔ Pas de différence

→ H1 = Hypothèse alternative ⇔ Différence significative

Application

Déterminez les hypothèses H0 et H1 :

On fait une étude sur un nouveau médicament qui ralentit l'évolution de la maladie d'Alzheimer. On constitue deux groupes de malades par tirage au sort. Le premier groupe prend le médicament étudié et le deuxième groupe prend un placebo.

Réponse :

H0 : pas de différence dans l'évolution de la maladie d'Alzheimer

H1 : différence significative dans l'évolution de la maladie d'Alzheimer

III-2 La démarche

Quand une question devra être testée, procédure identique :

⇒Etape 1 : Définir les **hypothèses** H0 et H1 (symétriques).

⇒Etape 2 : Choisir le **test** en fonction du type de données.

⇒Etape 3 : Choisir le **risque** α .

⇒Etape 4 : **Recueil** des données (pas avant !).

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

III-2 Les tests

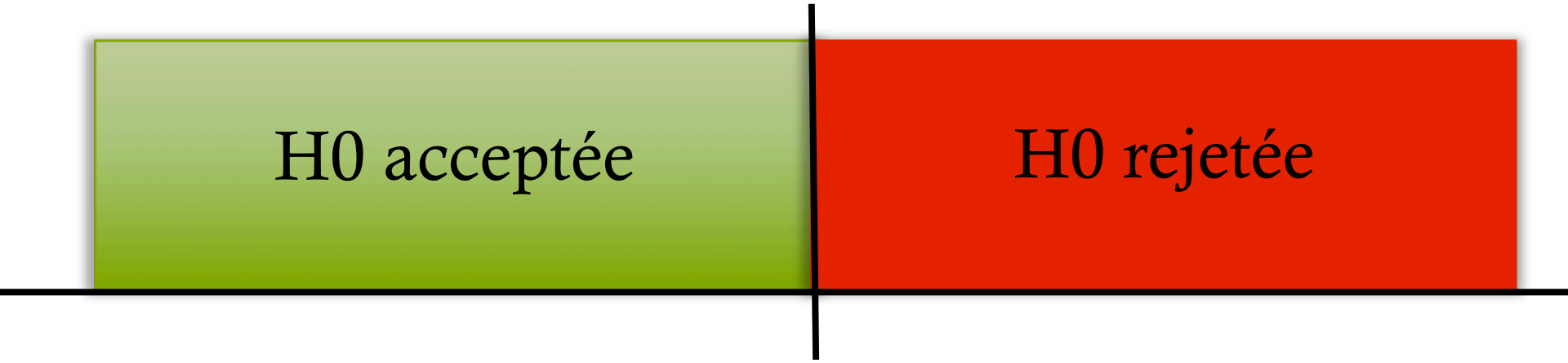
Définition :

Ce sont des **techniques qui permettent de choisir** si on doit garder ou repousser H0, en ayant fixé le risque d'erreur de notre décision.

III-2 Les tests

Acceptation de H0
⇔ Pas de différence

Rejet de H0
⇔ Acceptation de H1
⇔ Différence significative



III-3 Les risques

Risque de première espèce : α ⇔ Probabilité de rejeter H0, si H0 vraie

Risque de seconde espèce : β ⇔ Probabilité d'accepter H0, si H0 fausse

Puissance d'un test : $1 - \beta$ ⇔ Probabilité de rejeter H0, si H0 fausse

<u>Quelques règles :</u>

α et β ne varient pas dans le même sens, on privilégiera α au détriment de β !

Entre deux alternatives, on privilégiera pour H0 l'hypothèse qu'il serait le plus grave de rejeter à tort !

Synthèse

	Rejet H0	Non-Rejet H0
H0 vraie		
H0 fausse = H1 vraie		

Exemple de QCM :

QCM : Donnez les propositions justes concernant les risques statistiques.

A) Le risque de première espèce correspond à la probabilité de rejeter H0, si H0 vraie.
B) Le risque de seconde espèce est le risque α.
C) Le risque de seconde espèce correspond à la probabilité d'accepter H1, si H1 fausse.
D) La puissance d'un test correspond à la probabilité de rejeter H0, si H0 fausse.

CORRECTION : AD

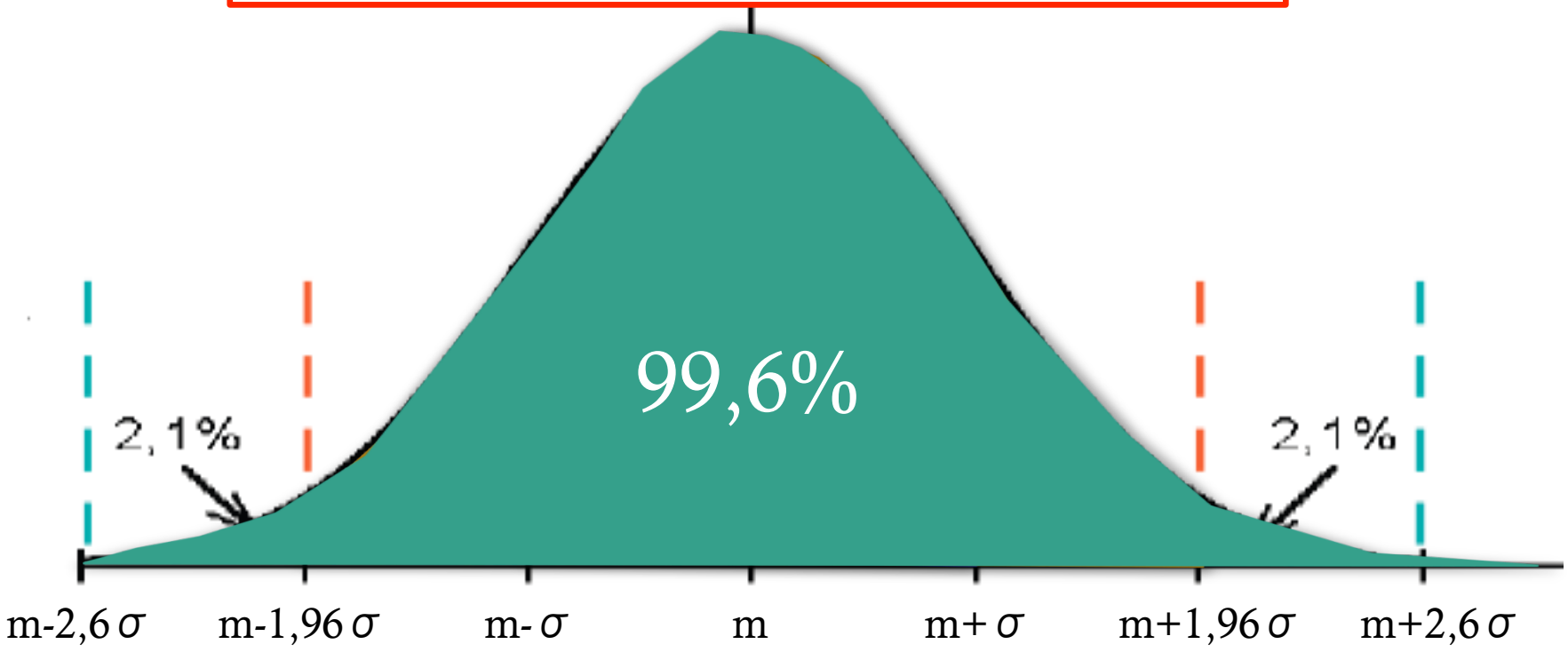
Merci pour votre attention !

Travaillez bien le cours !

Et rendez-vous samedi
pour un Concours Blanc de folie…

Pour aller plus loin…
L'estimation

$[\mu - \sigma\ ;\ \mu + \sigma]$ contient 68,2% de la population
$[\mu - 1{,}96\,\sigma\ ;\ \mu + 1{,}96\,\sigma]$ contient 95,4% de la population
$[\mu - 2{,}6\,\sigma\ ;\ \mu + 2{,}6\,\sigma]$ contient 99,6% de la population



Pour aller plus loin…
L'estimation