OBJECTIFS

- **Capables de traiter et analyser vos données expérimentales**
- **« Inférence » vous parle**
- **Le modèle statistique associé au modèle biologique (supposés)**
- **Dialoguer avec un(e) *bioinformaticien(ne)***
- **Capable d'effectuer Tests d'hypothèses classiques**
- **Faire une représentation graphique soignée**
- **Variables statistiques et variables aléatoires**
- **Echantillons, population et échantillonnage**
- **Avoir une idée de l'ordre de grandeur d'une information**
- **Décider, agir à partir des chiffres – risque associé à une décision (terrain, labo, entreprise)**
- **Utiliser les outils de bases sur un ordinateur**

BioStatistique

Licence BMC/BHS & Magistère de Biotechnologies

Pascal RIGOLET [30 heures] L3S5

- **Introduction : mesures, gestion des erreurs et analyse graphique**
- **Jugement sur échantillon - Tests d'hypothèse [inférence-déduction]**
- **ANOVA**
- **Maximum de vraisemblance**
- **Statistique à 2 variables : le tri croisé**
- **Plan factoriel d'expérience , principe et initiation**

Fluctuations d'échantillonnage

Pour une grande population, il y a une infinité de façons de tirer un *N-échantillon*.
Les différents échantillons possibles d'un tirage au hasard étant plus ou moins proches, on parle de **fluctuations d'échantillonnage**.

Les indicateurs (proportion **p** / moyenne **μ** / écart type **σ**) exacts des populations sont, en fait, dans la plupart des cas, des **estimations** obtenues grâce à des tests effectués sur des échantillons.

On étudie les populations à partir d'échantillons (représentatifs)
(ex : sondages)

Nous en reparlerons à propos des lois de probabilités…

Acquisition et analyse des données expérimentales

1. Fiabilité des données expérimentales et gestion des erreurs
2. La représentation graphique des données
3. La part du stochastique dans un graphe déterministe
4. Variabilité et fluctuations selon le modèle de la loi normale
5. Le squelette d’un script graphique sous R

Fiabilité

Une mesure ne peut rendre compte de façon **absolue** de la grandeur étudiée. Il faut, on le sait, procéder à une **série de mesures** en faisant varier certains **paramètres** pour obtenir le **modèle** biologique mais il faut également **répéter plusieurs fois la même expérience** pour obtenir l'**ordre de grandeur** de la variable d'étude ainsi que sa **dispersion** par rapport à cet ordre de grandeur.
Dans cette démarche, visant à approcher la vérité sans l'atteindre vraiment, on doit obtenir les **données les plus fiables possibles** et gérer les sources d'**erreurs**.

Erreurs

Les **erreurs** étant **incontournables**, il faut en limiter la portée dans nos expériences de biologie. Il faut en tenir compte dans la **représentation graphique.**

Source des erreurs

- Identifiez les sources d'erreurs dans cette expérience
- Indiquez les variables aléatoires associées
- Proposez une loi de probabilité pour chacune de ces variables

Questionnaire anonyme

- Prendre un ¼ de feuille
- Répondre dans l'ordre aux questions posées

Source des erreurs

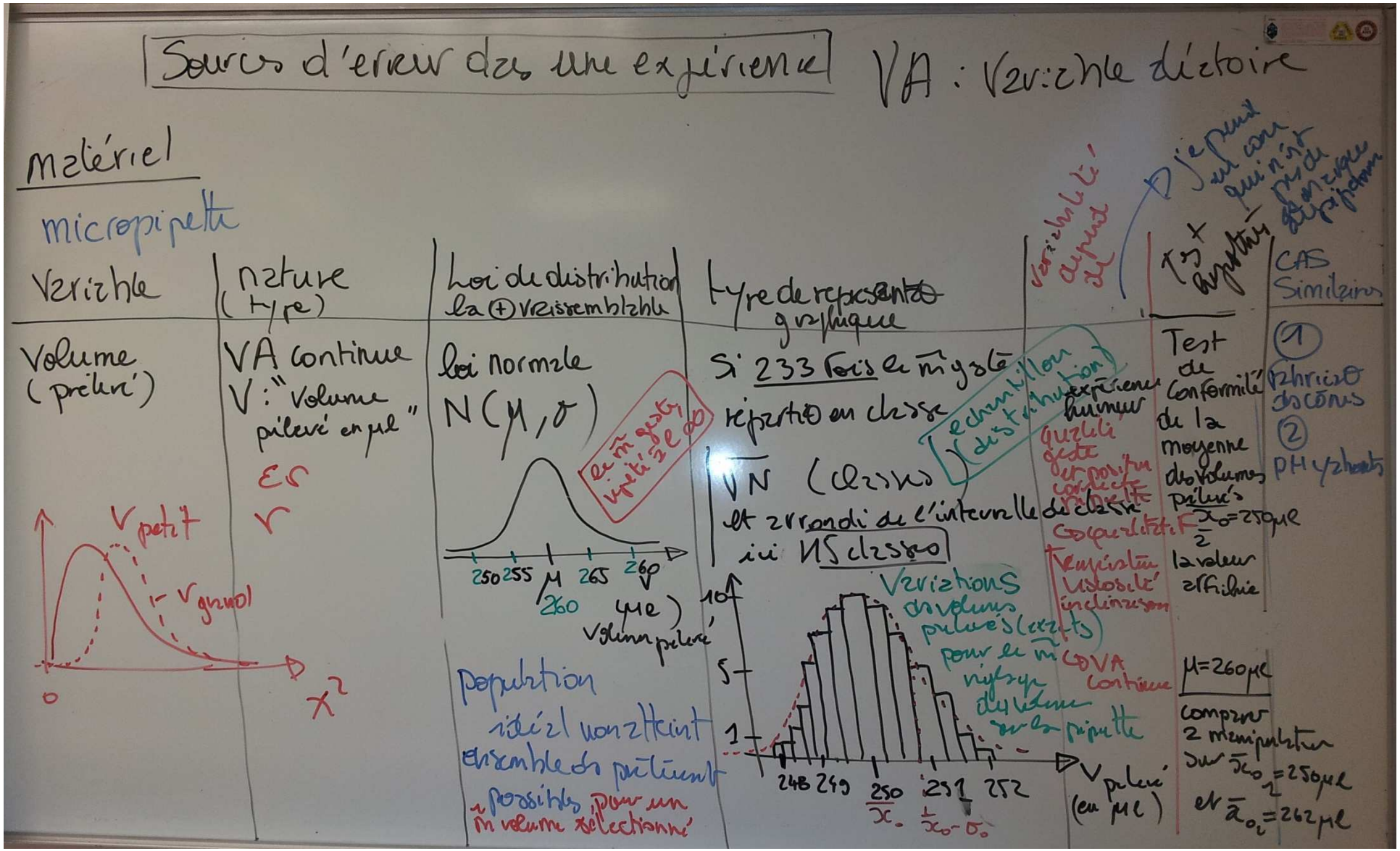
- Les sources sont multiples
- On peut les regrouper en trois catégories :

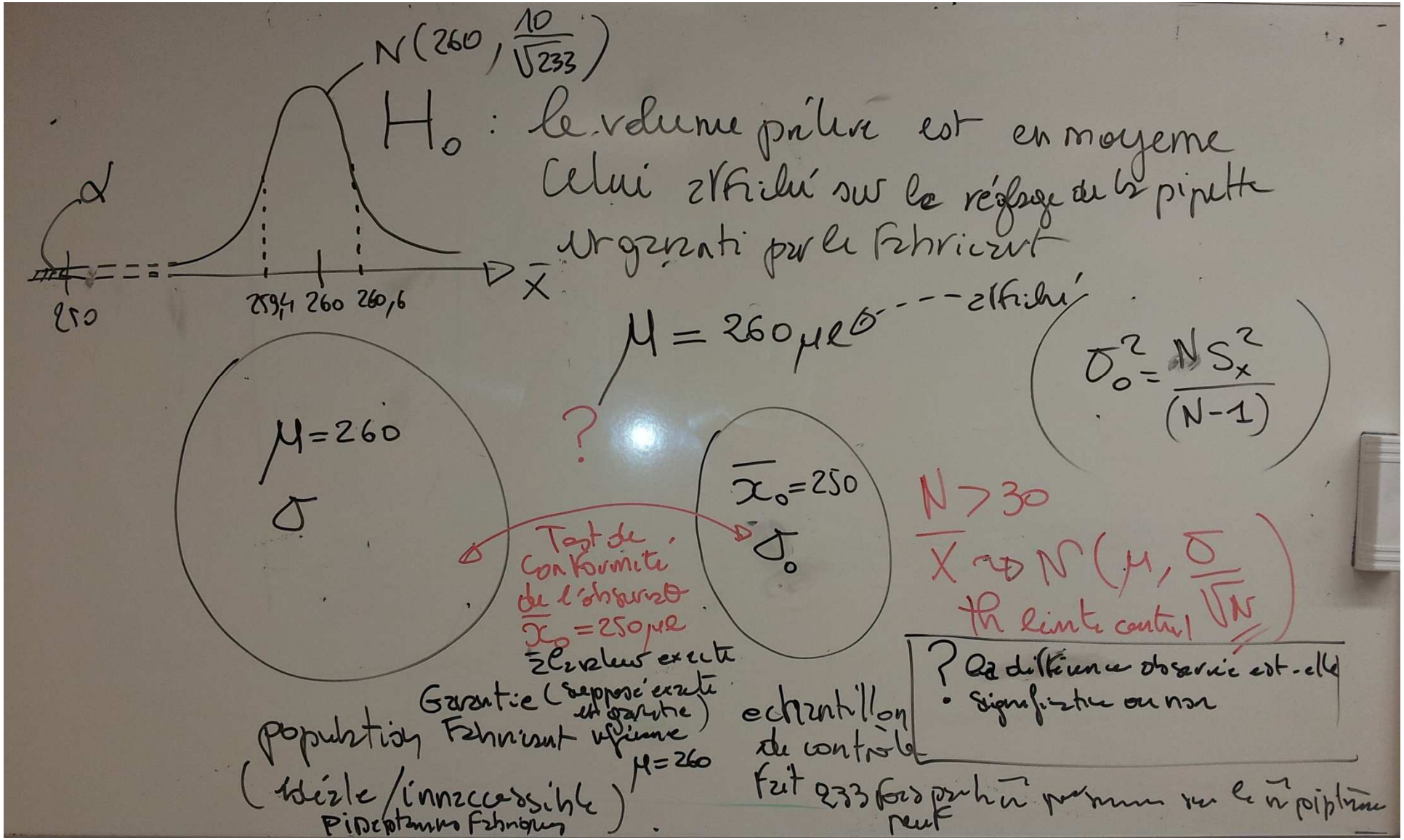
  celles liées à l'**expérimentateur**, celles liées aux **conditions thermodynamiques** de l'expérience, celles liées aux **appareillages** utilisés.
- Ces erreurs peuvent se cumuler de façon plus ou moins aléatoire dans l'acquisition des données expérimentales
- Il faut en réduire la portée
- Tenir compte des erreurs lors de la représentation graphique

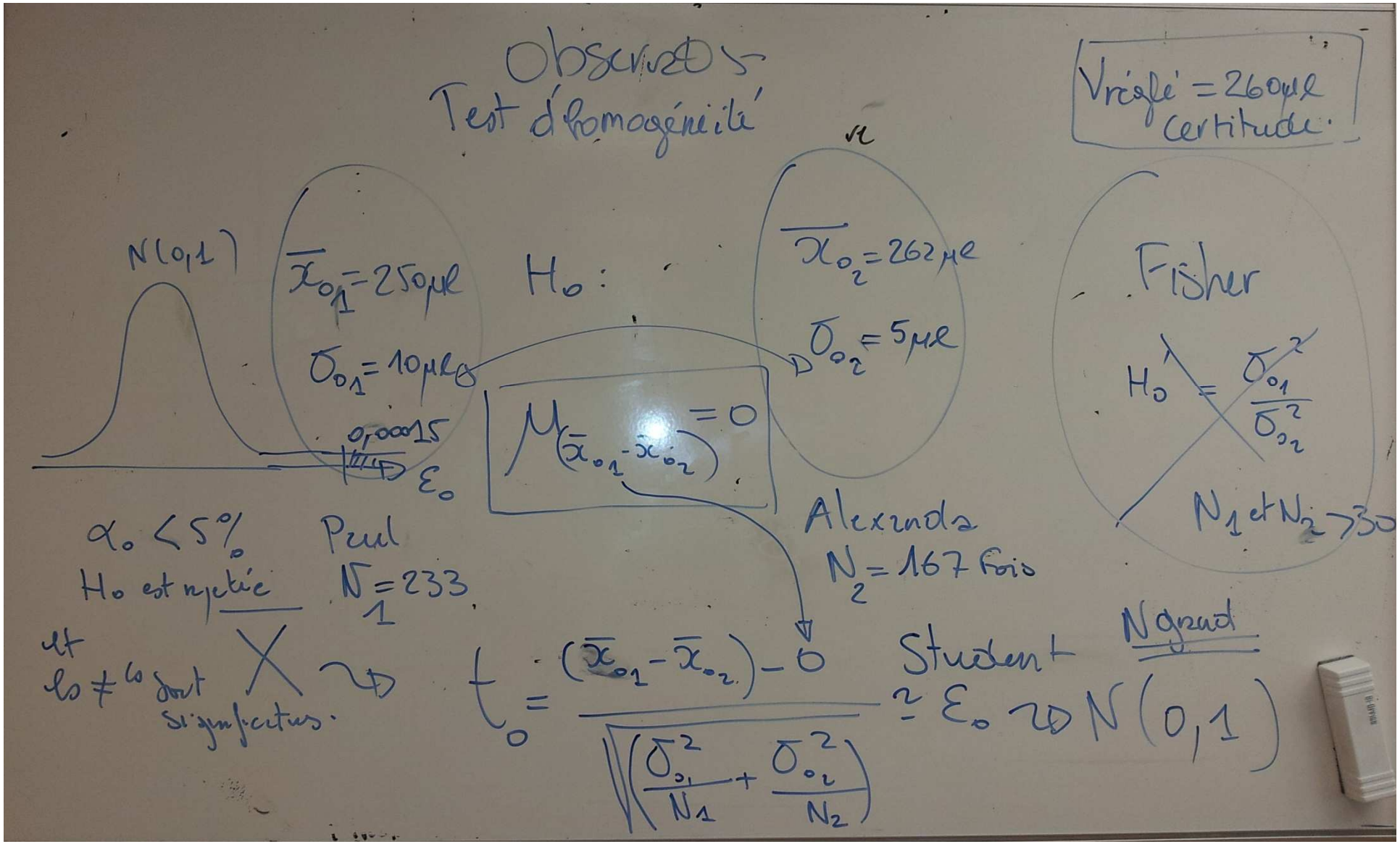
  > gestion des « barres d'erreur »

Acquisition et analyse des données expérimentales

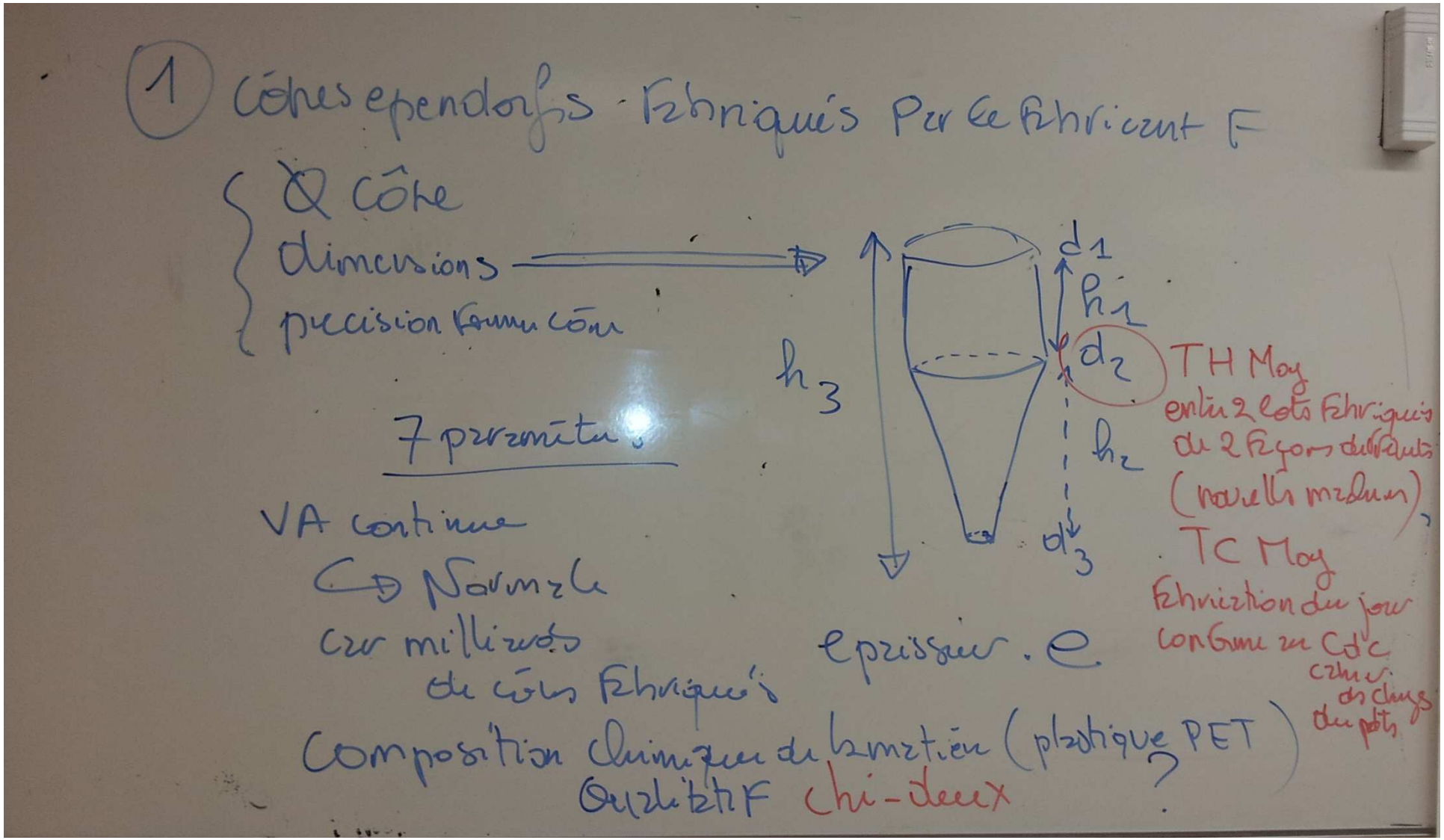
1. Fiabilité des données expérimentales et gestion des erreurs



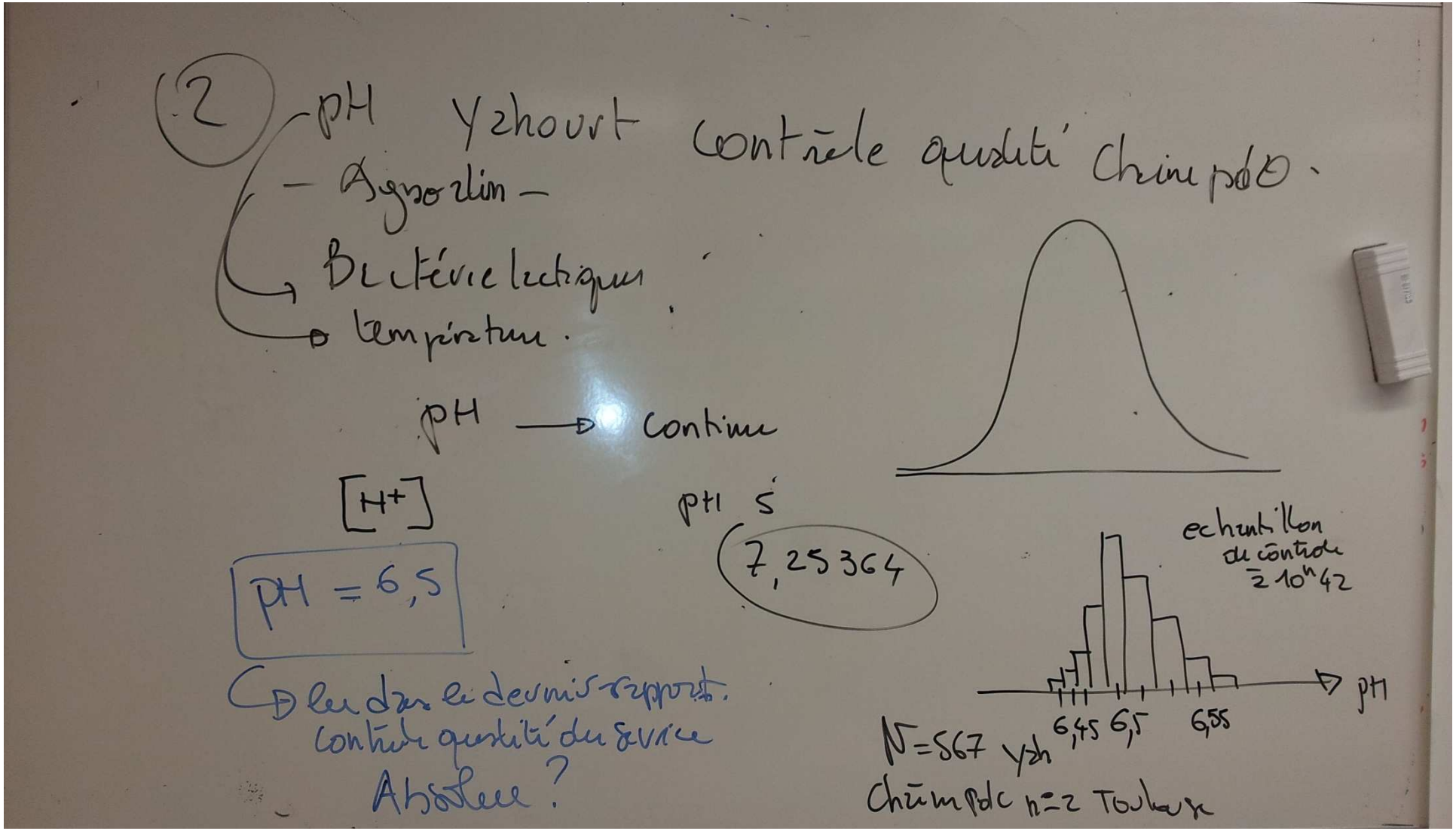

N(260, 10/√233)
H0 : le volume prélevé est en moyenne celui affiché sur le réglage de la pipette urgzrnati par le fabricant
α
250
259,4 260 260,6
x̄
μ = 260 µl ---- affiché
σ0² = N Sx² / (N-1)
μ=260
σ
?
x̄0=250
σ0
Test de conformité de l'observé x̄0 = 250 µl
Garantie (supposé exacte et garantie)
population fabricant infinie
μ=260
(théorique / inaccessible)
pipettes fabriquées
échantillon de contrôle
fait 233 fois pour chaque prélèvement sur le n° pipette neuf
N > 30
X̄ → N(μ, σ/√N)
th limite centrale
? La différence observée est-elle significative ou non


Observées
Test d'homogénéité
Vréglé = 260 µl
certitude
N(0,1)
$\bar{x}_{o_1} = 250\,\mu l$
$\sigma_{o_1} = 10\,\mu l$
H_0 :
$\bar{x}_{o_2} = 262\,\mu l$
$\sigma_{o_2} = 5\,\mu l$
Fisher
$H_0 = \frac{\sigma_{o_1}^2}{\sigma_{o_2}^2}$
N_1 et $N_2 > 30$
0,00015
ε_0
$\mu_{(\bar{x}_{o_1} - \bar{x}_{o_2})} = 0$
$\alpha_0 < 5\%$
H_0 est rejetée
Paul
$N_1 = 233$
Alexandra
$N_2 = 167$ fois
et les $\neq$ ⁰ sont significatives.
Student
Normal
$t_0 = \frac{(\bar{x}_{o_1} - \bar{x}_{o_2}) - 0}{\sqrt{\left(\frac{\sigma_{o_1}^2}{N_1} + \frac{\sigma_{o_2}^2}{N_2}\right)}} \simeq \varepsilon_0 \rightarrow N(0,1)$

Acquisition et analyse des données expérimentales
1. Fiabilité des données expérimentales et gestion des erreurs

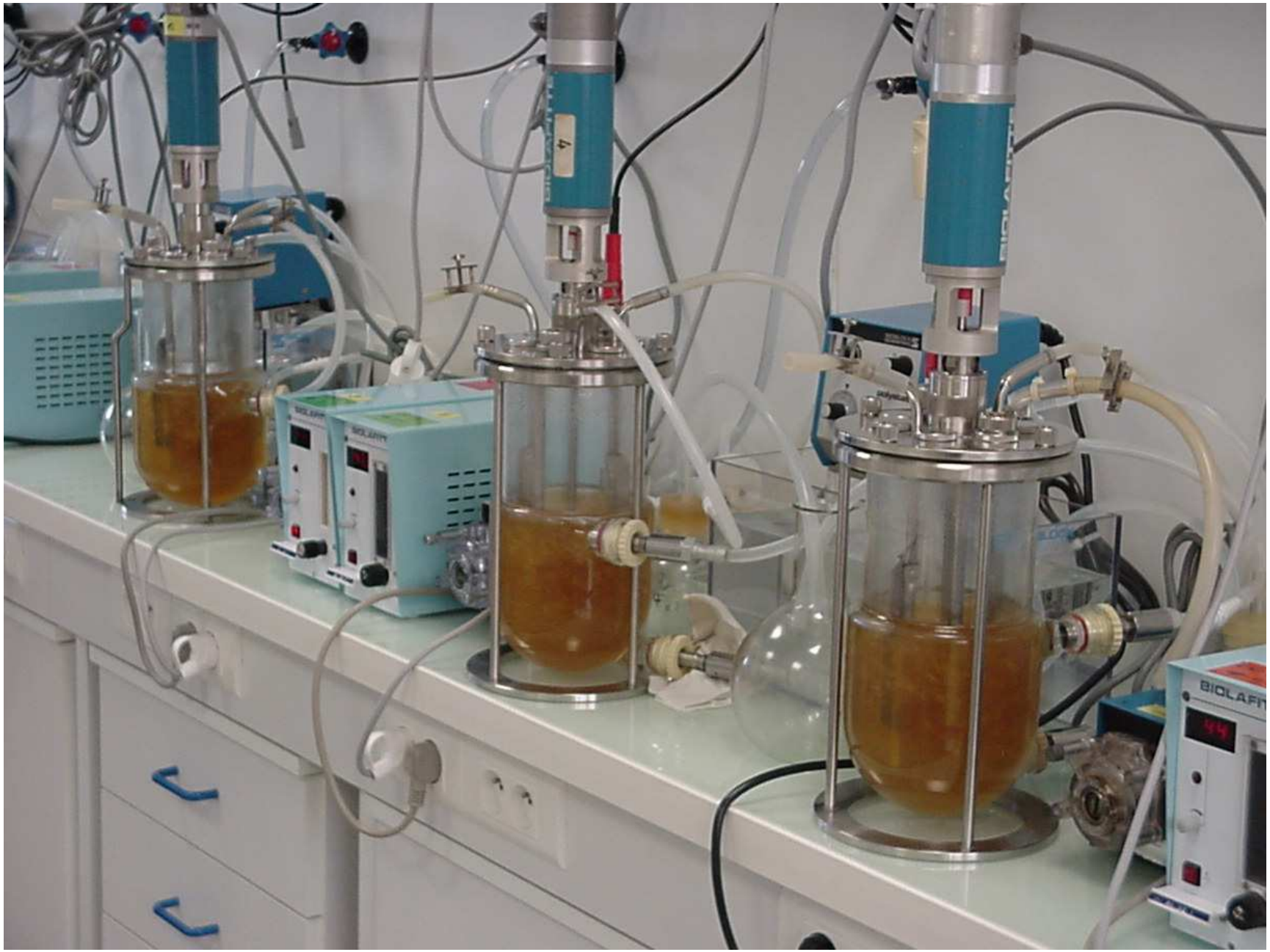


Acquisition et analyse des données expérimentales
1. Fiabilité des données expérimentales et gestion des erreurs



Acquisition et analyse des données expérimentales

1. Fiabilité des données expérimentales et gestion des erreurs



Acquisition et analyse des données expérimentales

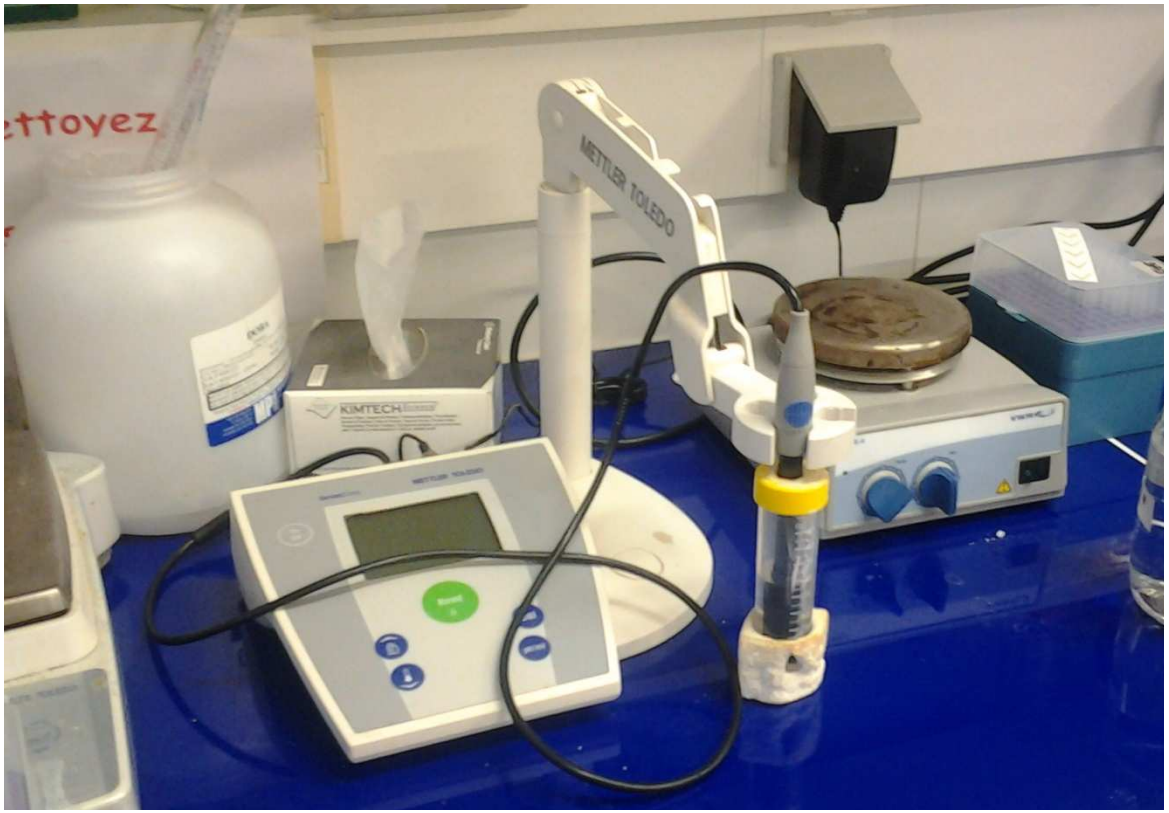
1. Fiabilité des données expérimentales et gestion des erreurs

Source des erreurs

Q1 - Identifiez les sources d'erreurs associées à cet appareillage

Q2 - Indiquez la variable aléatoire associée

Q3 - Pouvons nous obtenir une valeur absolue/précise ? Commentez ….



Acquisition et analyse des données expérimentales
1. Fiabilité des données expérimentales et gestion des erreurs

- Identifiez les sources d'erreurs associées à cet appareillage

- Indiquez la variable aléatoire associée

Acquisition et analyse des données expérimentales
1. Fiabilité des données expérimentales et gestion des erreurs

- Identifiez les sources d'erreurs associées à cet appareillage

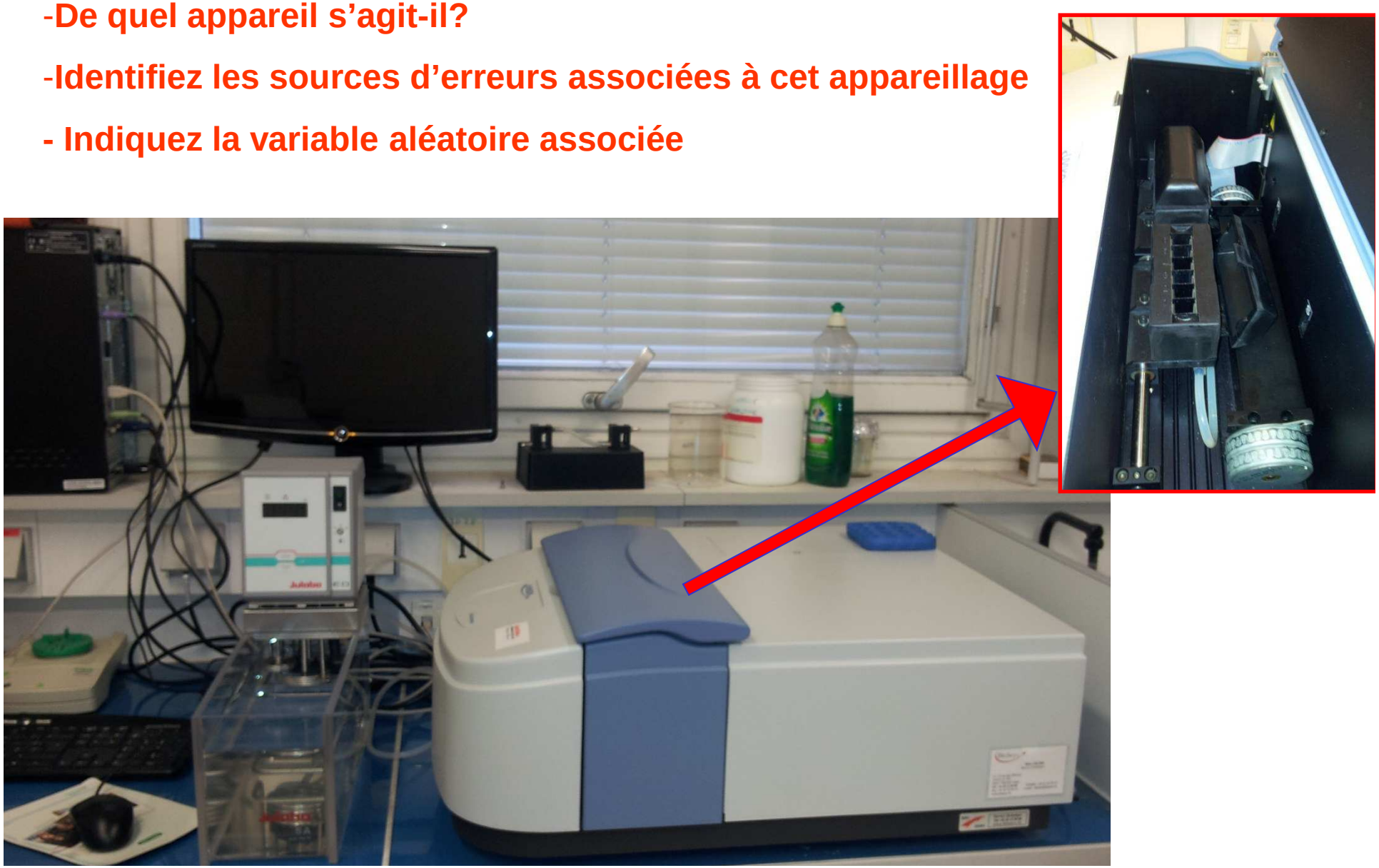
- Indiquez la variable aléatoire associée

Acquisition et analyse des données expérimentales
1. Fiabilité des données expérimentales et gestion des erreurs

-De quel appareil s'agit-il?

-Identifiez les sources d'erreurs associées à cet appareillage

- Indiquez la variable aléatoire associée



Source des erreurs

- La gestion des erreurs et une partie importante du traitement des données
- Nous allons donc voir dans les cours suivants comment nous allons inclure les barres d'erreur dans une courbe expérimentale.
- Les **4 étapes** indispensables à la réalisation d'un graphe :

  1/ **l'acquisition** des données,

  2/ l'édition ou la récupération de ces **données dans un fichier**

  3/ le **tracé du graphe** expérimental avec ces données

  4/ la gestion des **barres d'erreurs** dans le graphe.

Source des erreurs

- Il existe différentes façons d'inclure les barres d'erreur dans un graphe. Nous allons le voir aussi bien avec le logiciel R qu'avec le logiciel Excel. Pour tracer les barres d'erreur, nous aurons à considérer soit les écarts minima et maxima, soit la dispersion des données (lorsque nous en avons suffisamment dans un modèle de distribution normalienne) à une distance d'1, 2 ou 3 fois l'écart type. Mais le plus rigoureux sera le traitement par intervalle de confiance

- Notre apprentissage se fera essentiellement avec le logiciel R. Nous aurons à nous familiariser avec certaines des instructions de ce logiciel notamment pour les parties graphique et statistiques

Population : ensemble total d'objets ou d'individus à étudier, à partir duquel sont extraits des échantillons.

La moyenne μ et l'écart-type σ de la population

sont des **constantes** (généralement inconnues),

exemples de **paramètres** fixes de la population ou **objectifs**.

⇨ Les probabilités **P(X)** sont utilisées dans le calcul de μ et σ

Echantillon : Sous ensemble de la population. Un échantillon **représentatif** est un sous ensemble **choisi au hasard dans la population**.

La moyenne $\overline{X}$ et l'écart-type **s** de l'échantillon sont des **variables aléatoires**, variant d'un échantillon à l'autre, et sont appelées **statistiques d'échantillon**, statistiques aléatoires ou **estimateurs** (ici respectivement de μ et σ)

Remarque : la médiane m_e peut être dans certains cas un meilleur estimateur de μ que $\overline{x}$

Les fréquences relatives f/n sont utilisées dans le calcul de $\overline{x}$ et s

⇨[Un estimateur T est dit biaisé si son espérance E(T) est différente de sa cible θ dans la population : biais = $E(T)-\theta$]

Tirage d'un échantillon **avec remise** : **important** pour garantir **l'indépendance** des n observations qui le constituent (surtout dans les **petites populations**).

La proportion comme la moyenne changeant sinon à chaque tirage!

Tirage d'un échantillon **sans remise** : **sans importance dans les grandes populations**, aucune différence pratiquement que l'on remette ou non chaque individu avant le tirage suivant. Pour l'essentiel les observations sont indépendantes. Ce n'est pas le cas pour une petite population

Déduction :

prédire, à partir d'une population connue ou supposée connue,

les caractéristiques des **échantillons** qui en seront prélevés

Induction (inférence) :

prédire les caractéristiques d'une **population** inconnue à partir des statistiques déterminées dans un échantillon représentatif de cette population.

⇨ *Extrapolation des observations réalisées dans un échantillon à l'ensemble de la population*