Cours VETE0432-4
Partim « Biostatistique »

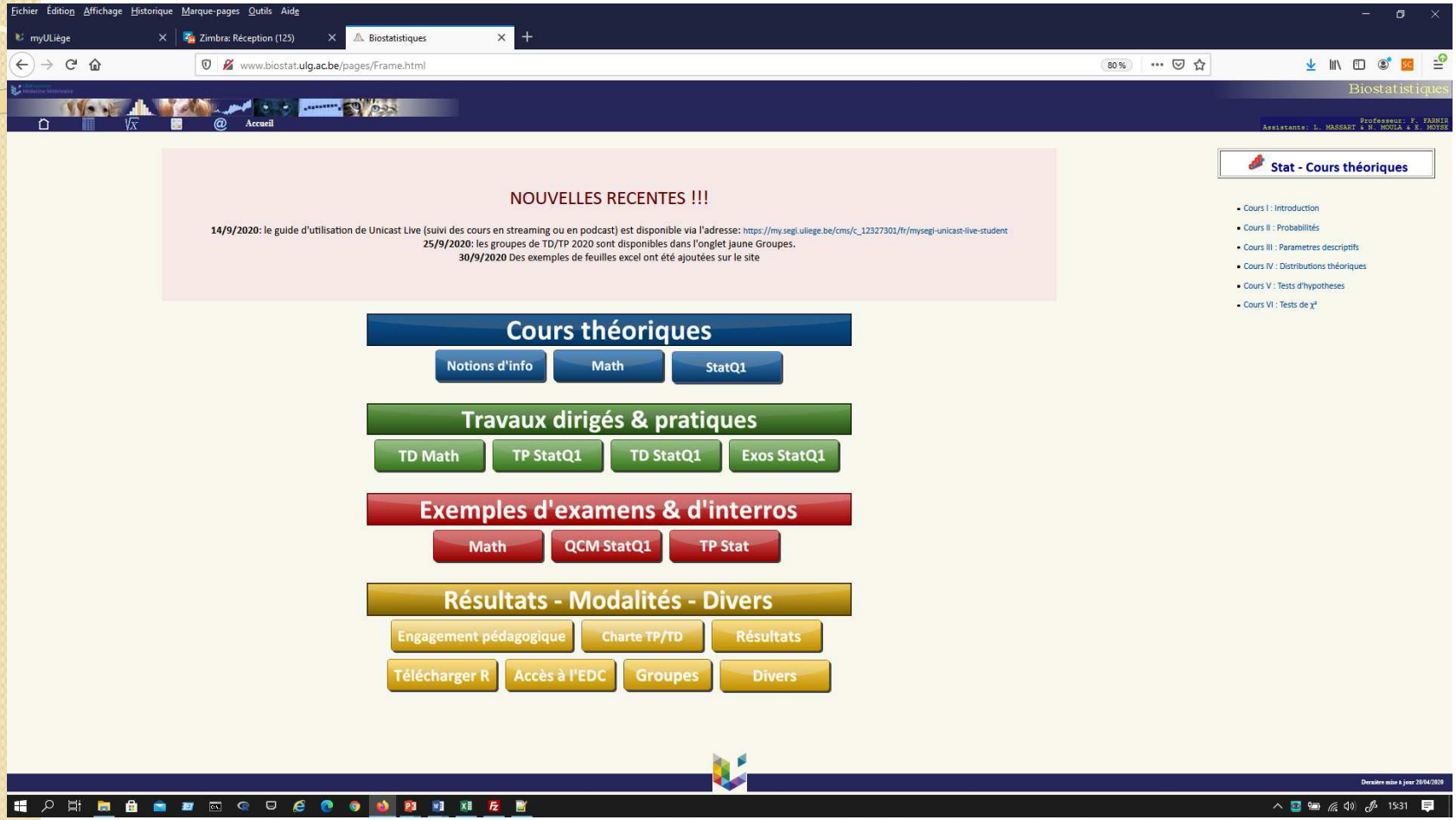
F. Farnir

L. Massart – A. Rives - N. Moula

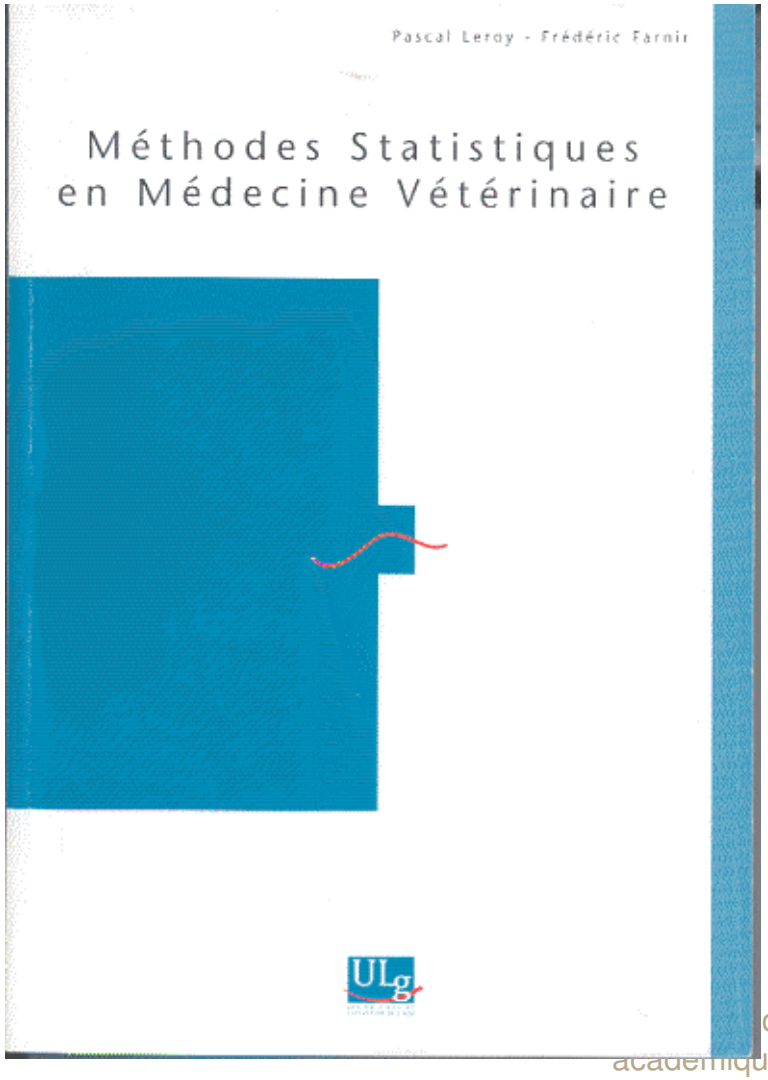
Organisation du cours

- Cours théorique: 28 h
  - Partim I: [Début → χ^2]
  - Partim II: cours VETE2111-1 (bloc 2)
- Séances d'exercices (en amphi): 10 h
- Séances de TP (salle info): 10 h

Site web du cours



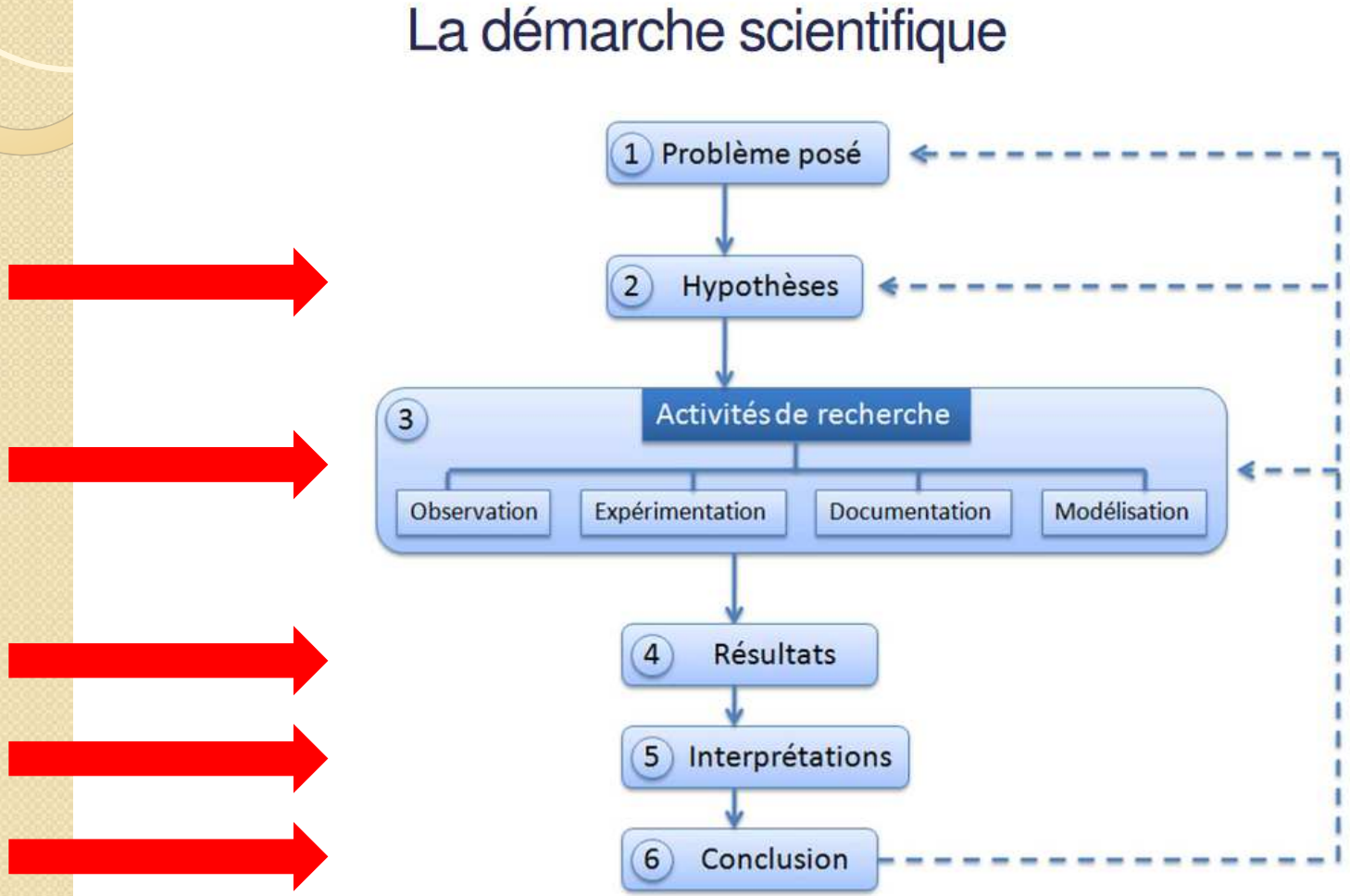
Syllabus du cours



Objectifs du cours (Partim I)

- Statistique descriptive
- Calculs de probabilités
- Distributions théoriques
- Introduction aux tests d'hypothèses
  - Le test de χ^2

Pourquoi enseigner les statistiques ?



Statistique et biostatistique

- Statistique = ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions.
- Biostatistique = statistique appliquée dans le domaine du vivant

Biostatistique en sciences vétérinaires

- Démarche scientifique à acquérir (cfr ↑)!
- Les vétérinaires sont des acteurs du vivant, appliquant une approche scientifique nécessitant:
  - La description de la variabilité importante des phénomènes liés au vivant.
  - L'utilisation d'outils d'investigation de la complexité liée au vivant.
  - L'élaboration et interprétation de tests d'hypothèses *in vivo* (et *in vitro*, voire *in silico*)

Quelques exemples

- Comment s'assurer (« tester ») de l'efficacité d'un nouveau médicament vétérinaire (par exemple, un nouveau vaccin) ?

Des exemples vétérinaires?

- Comment comparer des régimes alimentaires permettant de combattre l'obésité chez le chien ?

Des exemples vétérinaires?

- Comment évaluer les meilleurs reproducteurs dans une espèce utilisée pour la production ?

Des exemples vétérinaires?

- Comment évaluer les performances chez les chevaux trotteurs ?

Des exemples vétérinaires?

- Quels sont les facteurs d'environnement qui influent sur les performances de reproduction chez la brebis ?

Des exemples vétérinaires?

- Comment évaluer l'évolution de la taille de la population d'une espèce en danger ?

Recueil de données et Biostatistique

- Statistique: Ensemble de méthodes mathématiques qui, à partir du recueil et de l'analyse de données réelles, permettent l'élaboration de modèles probabilistes autorisant les prévisions.
- Le premier problème est donc celui de la récolte des données

La statistique descriptive

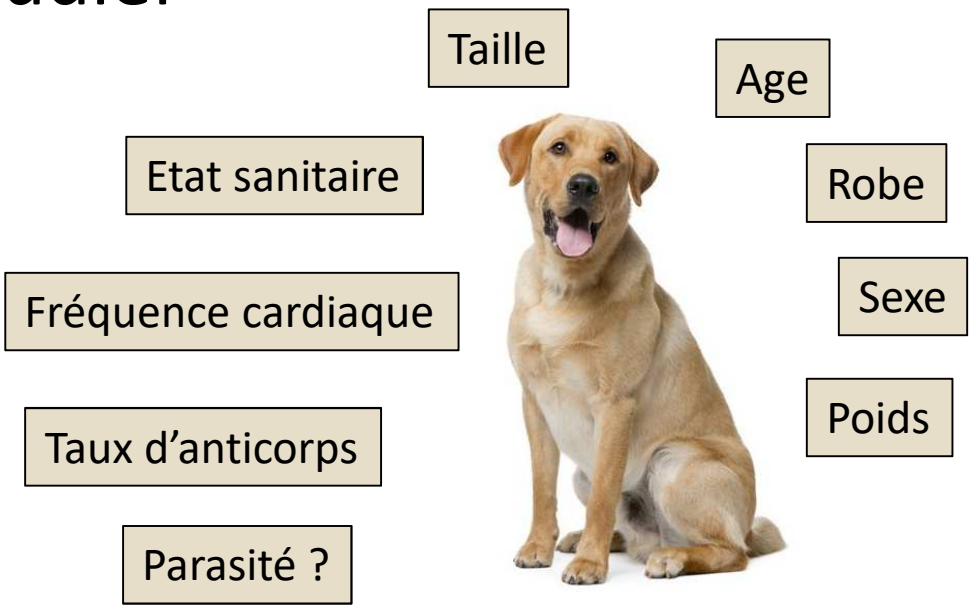
- Statistique descriptive
  - Description des données qu'on souhaite étudier
    - Type de données
    - Quantité de données
  - Résumé des données
    - Graphiques
    - Paramètres descriptifs (position, dispersion, …)

Données

- Que sont « les données » auxquelles il est fait allusion plus haut ?
  - Tout dépend évidemment de l'expérience qui est menée…
    - Ex1: Poids des labradors de 3 ans
    - Ex2: Etat sanitaire des individus (Sain – Malade)
    - Ex3: Couleur de la robe de bovins
    - Ex4: Production quantitative (kgs lait...) ou qualitative (vitesse de traite...)
    - Ex5: Comptages de lymphocytes

Données

- Il existe donc différents types de données, correspondant aux différentes caractéristiques des sujets qu'on souhaite étudier

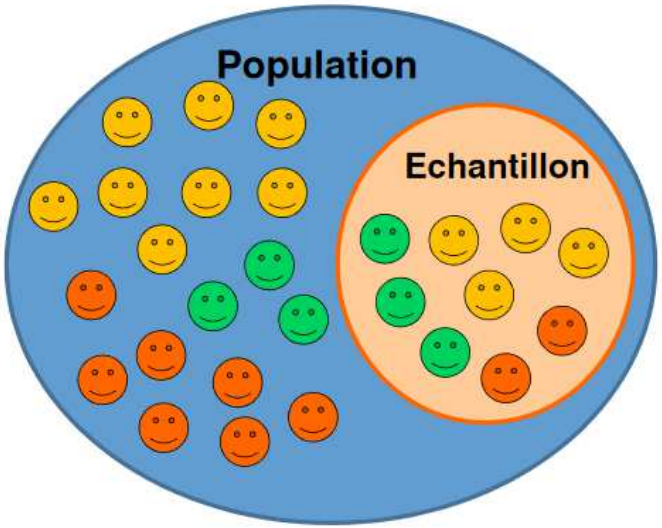


Récolte de données

- L'ensemble de toutes les données qu'il est théoriquement possible de récolter constitue "la population" de mesures
  - Exemple: la population des poids de labradors de 3 ans
  - Remarque: ne pas confondre « population des labradors » et « population des poids de labradors »

Récolte de données

- L'ensemble de toutes les données récoltées constitue « un échantillon » de la population.
- Il s'agit (en général) d'un sous-ensemble de la population.



Echantillonnage

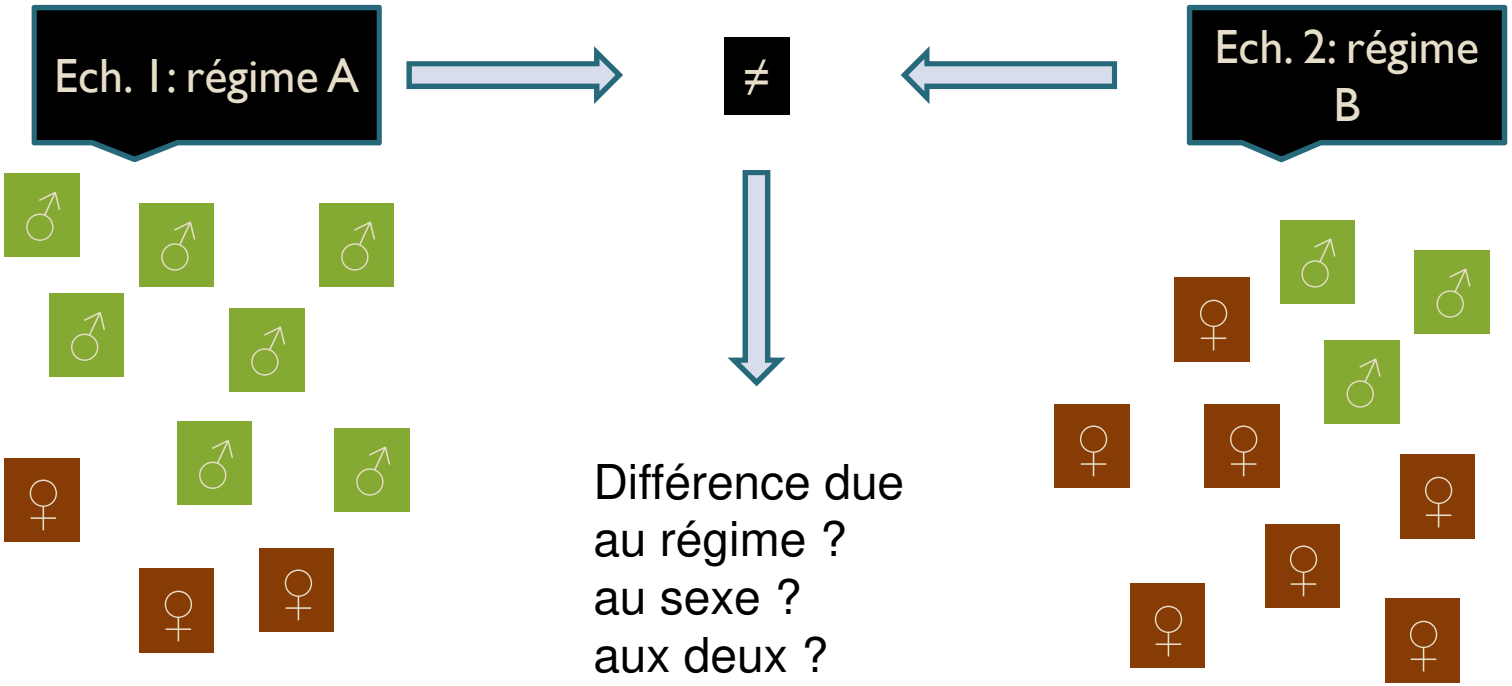
Echantillonnage

- L'échantillon doit être représentatif de la population visée. Il doit donc présenter, pour les caractéristiques qui sont importantes pour l'étude, des propriétés qui soient le plus proche possible de celles de la population dont il est extrait.
- Dans le cas contraire, l'échantillon est « biaisé » et les résultats de l'étude seront faussés.

Echantillonnage

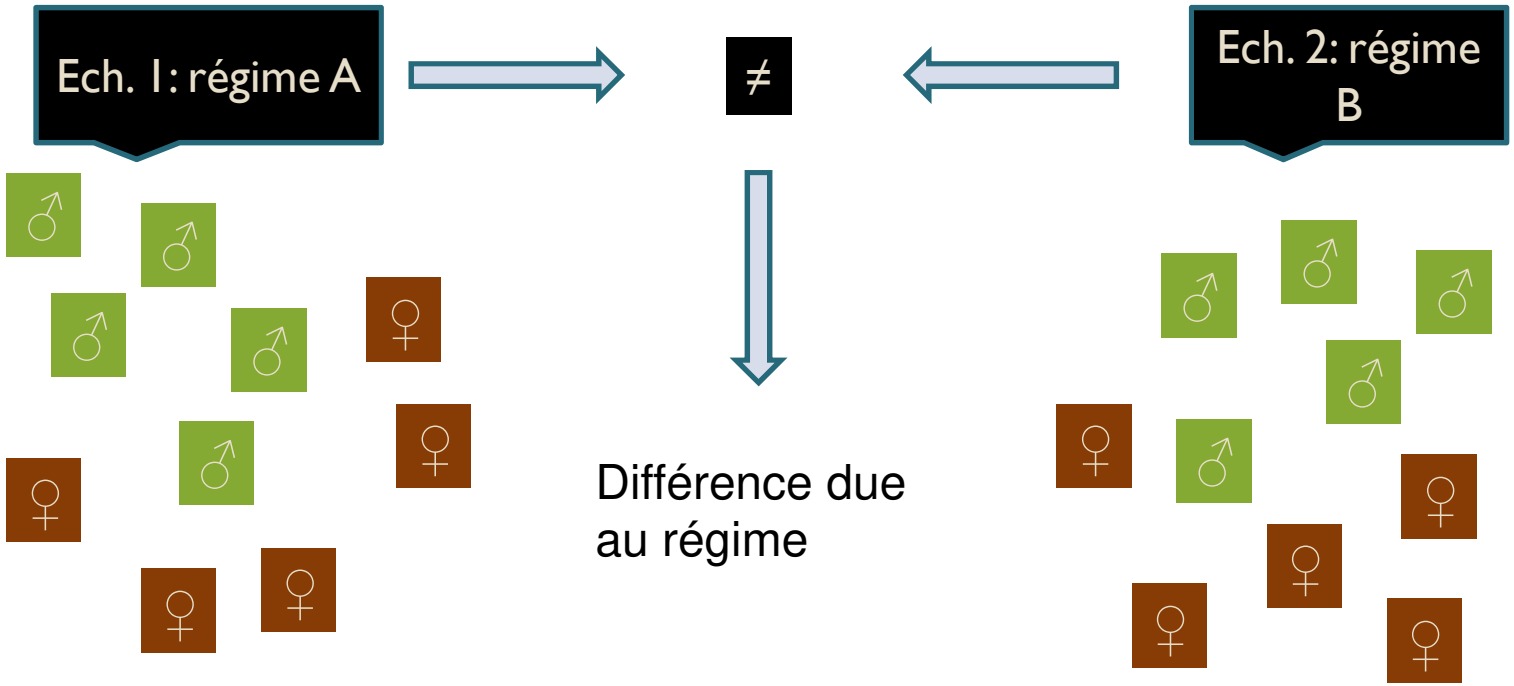
- Echantillonnage
  - L'échantillon doit être représentatif de la population visée
  - Dans le cas contraire, l'échantillon sera « biaisé »
  - Exemple (de biais): comparaison de deux régimes alimentaires chez des moutons.

Un exemple d'étude biaisée



⇒ Confusion entre effets du sexe et du régime !

Un exemple d'étude non biaisée



Comment récolter des données?

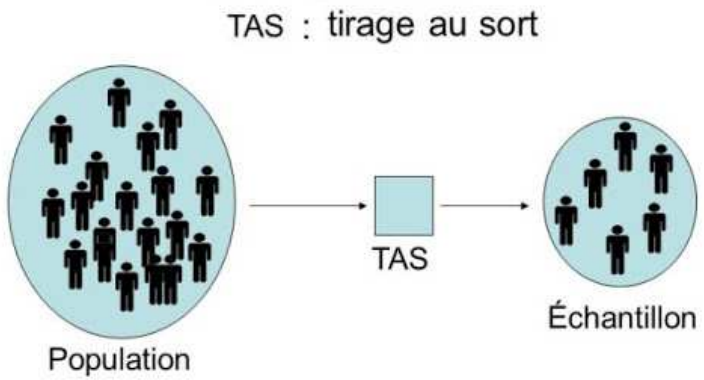
- Echantillonnage
  - Pour que l'échantillon soit représentatif de la population visée, il faut donc:
    - précéder l'échantillonnage de l'identification des facteurs pouvant induire des biais (sexe, âge, …) et/ou de la confusion
    - échantillonner (semi-)aléatoirement les sujets qui constitueront l'échantillon en tenant compte des facteurs identifiés

Comment récolter des données?

- Echantillonnage
  - Il apparaît donc que l'échantillonnage devrait être planifié
    - Que veut-on voir (objectifs) ?
    - Quelle est la « population » visée ?
    - Quels sont les biais potentiels, et comment les éviter ?
  - On parle alors de design expérimental

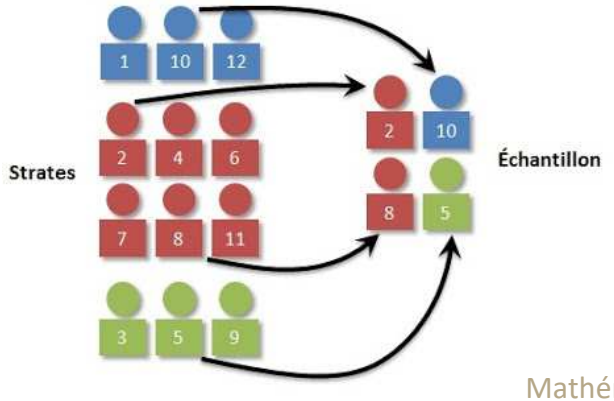
Comment récolter des données?

- Planification de l'échantillonnage:
  - Exemples de planifications:
    - Échantillonnage aléatoire
      - Exemple:



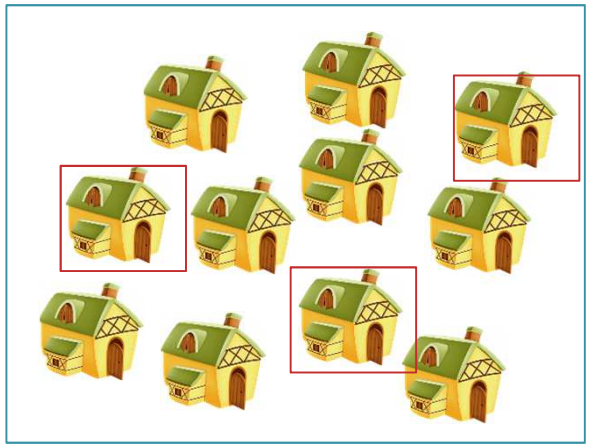
Comment récolter des données?

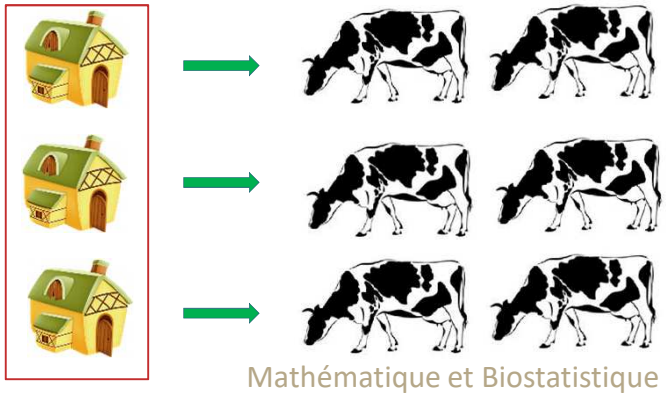
- Planification de l'échantillonnage:
  - Exemples de planifications:
    - Échantillonnage stratifié
      - Exemple: Il faut des chiens avec les 3 robes possibles dans l'étude



Comment récolter des données?

- Planification de l'échantillonnage:
  - Exemples de planifications:
    - Échantillonnage en grappes
      - Exemple: sélection aléatoire de fermes, puis de sélection aléatoire de quelques vaches dans ces fermes





Exemple d'échantillonnage planifié

- Echantillonnage planifié: un exemple.
  - 30 individus, pris au hasard dans la population, doivent être répartis dans 5 groupes qui recevront des traitements différents. On souhaite que les groupes soient homogènes en termes de poids car le poids a une influence potentielle sur le caractère étudié

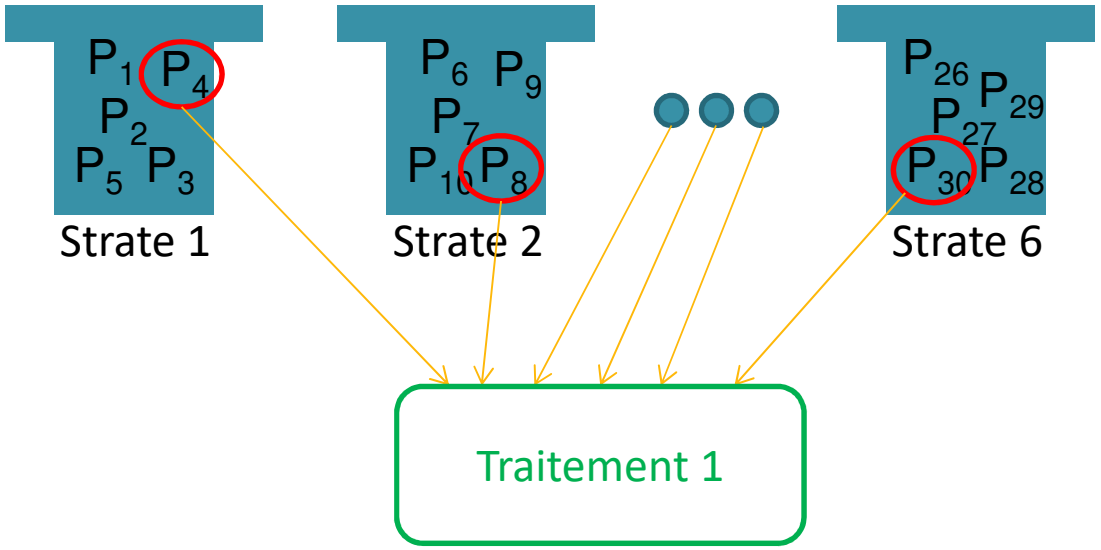
Comment récolter les données ?

- Un exemple de planification d'expérience
  - On peut recourir à un échantillonnage stratifié (sur le poids):
    - 30 individus – 5 traitements => 6 poissons/traitement
    - On crée 6 groupes (= strates) et on prélèvera, pour chaque traitement, 1 individu/strate
    - La strate 1 contient les 5 individus les plus lourds, la strate 2 les 5 individus suivants en termes de poids, …, la strate 6 contient les 5 individus les plus légers.

- Cfr la « méthode des chapeaux » utilisée dans le « tirage au sort » pour les compétitions de football, par exemple

Comment récolter les données ?

- Un exemple de planification d'expérience
  - Pratiquement:
    - On trie les données: $P_1 \geq P_2 \geq \ldots \geq P_{30}$



Comment récolter les données ?

- Un exemple de planification d'expérience
  - Illustration:
    - Supposons qu'on a obtenu le poids des 30 individus à répartir dans un vecteur (appelé, logiquement, poids)

```
> poids
 [1] 213.1 208.3 184.5 203.3 217.9
 [6] 181.1 181.0 188.0 193.7 190.6
[11] 209.7 185.6 189.1 209.3 216.6
[16] 231.2 191.1 163.8 216.1 182.0
[21] 195.2 240.1 203.4 175.9 206.0
[26] 207.3 193.6 191.1 207.5 207.0
>
```

Comment récolter les données ?

- Un exemple de planification d'expérience
  - Illustration:
    - Il est facile, avec la fonction sort de R, de trier ces poids. Le résultat du tri sera mis dans un nouveau vecteur poids.tries:

```
> poids.tries<-sort(poids)
> poids.tries
 [1] 163.8 175.9 181.0 181.1 182.0
 [6] 184.5 185.6 188.0 189.1 190.6
[11] 191.1 191.1 193.6 193.7 195.2
[16] 203.3 203.4 206.0 207.0 207.3
[21] 207.5 208.3 209.3 209.7 213.1
[26] 216.1 216.6 217.9 231.2 240.1
>
```

Comment récolter les données ?

- Un exemple de planification d'expérience
  - Illustration:
    - La fonction sample(v,n) de R permet d'échantillonner n valeurs au hasard dans le vecteur v. Ce qui va nous permettre de « mélanger les chapeaux »:

```
> strate1<-sample(1:5,5)
> strate2<-sample(6:10,5)
> strate3<-sample(11:15,5)
> strate4<-sample(16:20,5)
> strate5<-sample(21:25,5)
> strate6<-sample(26:30,5)
> strate3 # par exemple
[1] 15 12 11 13 14
```

Comment récolter les données ?

- Un exemple de planification d'expérience
  - Illustration:
    - On peut ensuite facilement former les groupes associés à chaque traitement:

```
> trait1<-c(strate1[1],strate2[1],strate3[1],strate4[1],strate5[1],strate6[1])
> trait2<-c(strate1[2],strate2[2],strate3[2],strate4[2],strate5[2],strate6[2])
> trait3<-c(strate1[3],strate2[3],strate3[3],strate4[3],strate5[3],strate6[3])
> trait4<-c(strate1[4],strate2[4],strate3[4],strate4[4],strate5[4],strate6[4])
> trait5<-c(strate1[5],strate2[5],strate3[5],strate4[5],strate5[5],strate6[5])
> trait4 # par exemple
[1] 1 10 13 16 25 27
```

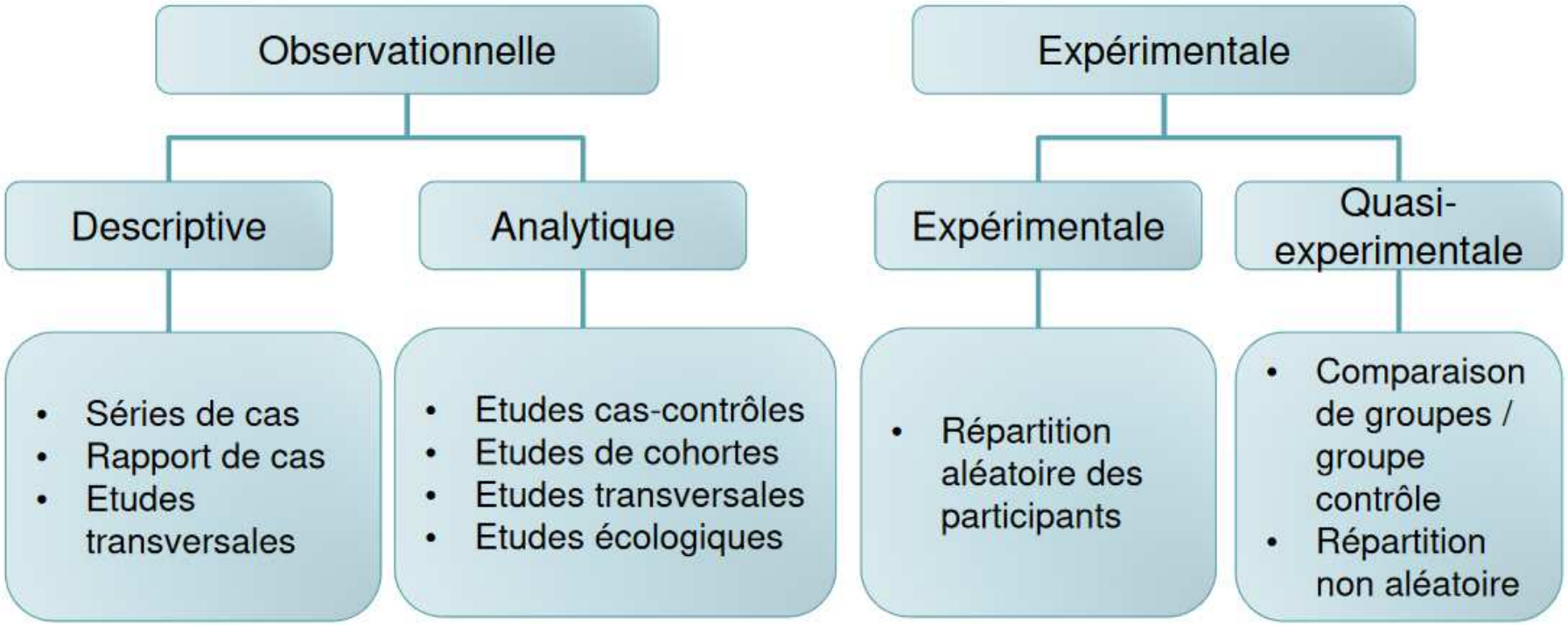
Comment récolter les données ?

- Un exemple de planification d'expérience
  - Illustration:
    - On peut vérifier que les poids sont assez homogènes entre traitements:

```
> mean(poids.tries[trait1])
[1] 198.4333
> mean(poids.tries[trait2])
[1] 198.6667
> mean(poids.tries[trait3])
[1] 201.9667
> mean(poids.tries[trait4])
[1] 196.8333
> mean(poids.tries[trait5])
[1] 201.2833
```

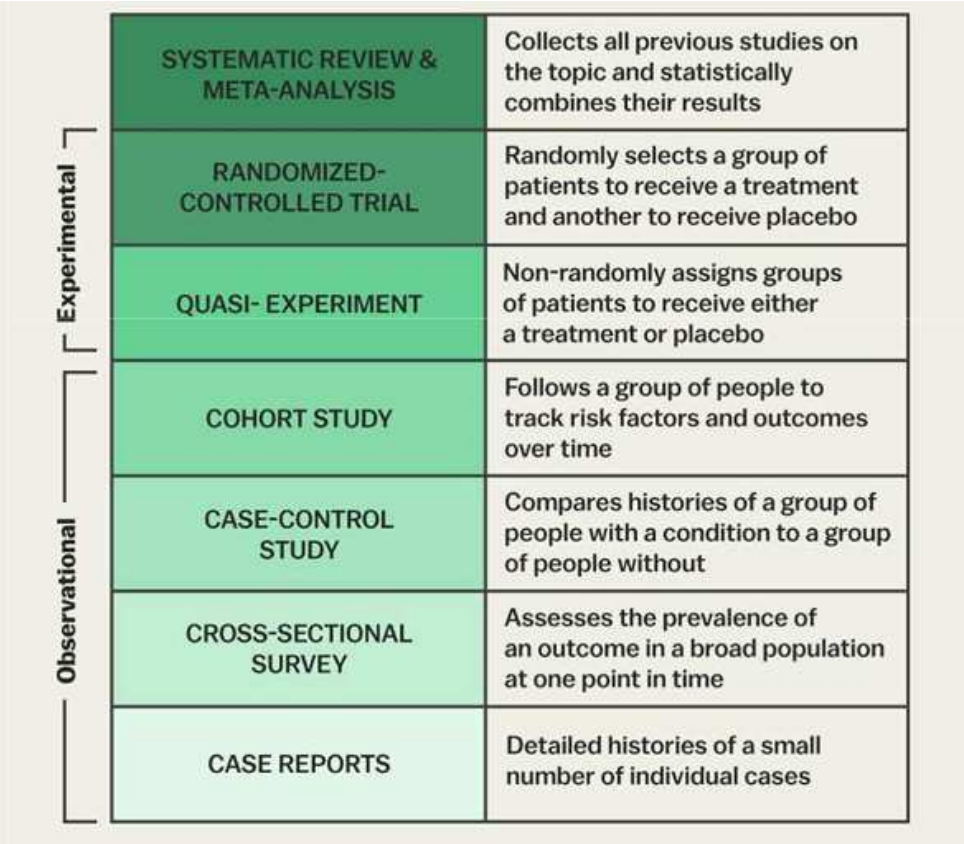
Types d'études

- En médecine (vétérinaire)



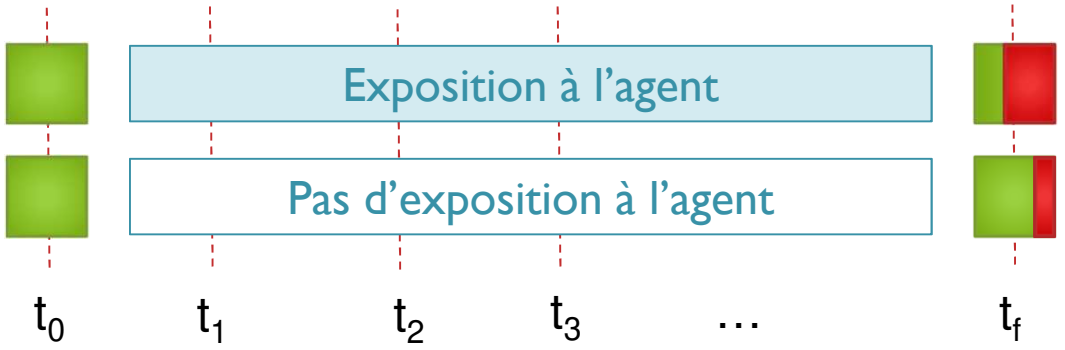
Types d'études

- En médecine (vétérinaire)



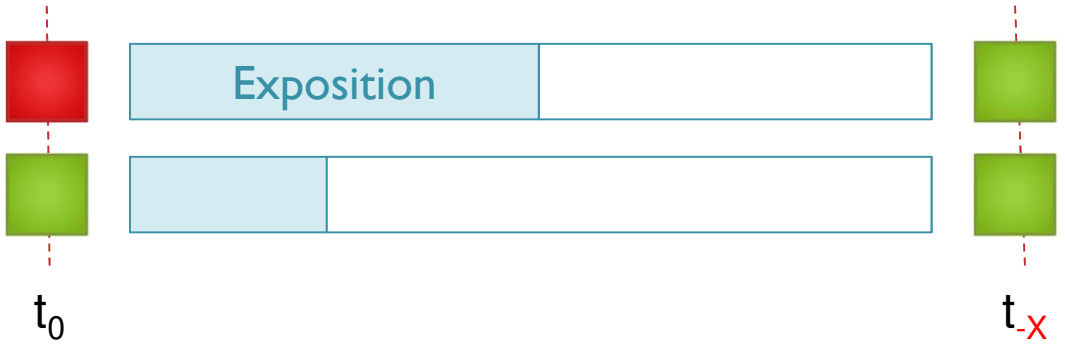
Types d'études

- En médecine (vétérinaire)
  - Etudes prospectives (« cohorts »)
    - Exemple: effet d'un agent chimique sur l'apparition de tumeurs



Types d'études

- En médecine (vétérinaire)
  - Etudes rétrospectives (« case-controls »)
    - Exemple: études « cliniques »

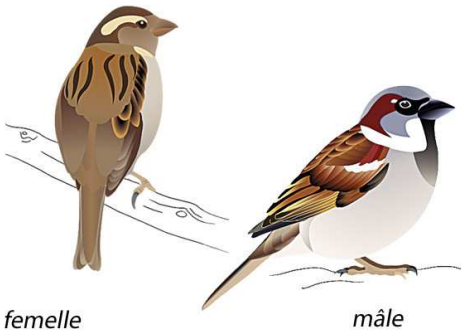


Types d'études

- En médecine (vétérinaire)
  - Etudes transversales (« cross-sectional »)
    - Exemple: enquêtes
      - Questions sur la présence/absence du caractère étudié (maladie, …)
      - Questions sur la présence/absence du facteur étudié (sexe, exposition, …)
      - Autres questions annexes (autres facteurs de confusion…)

Quelles données récolter?

- Cela dépend des objectifs de l'analyse…
- Types de données
  - Discrètes, Nominales (classification)
  - Exemples:
    - Sexe
    - Traitement
    - Race
    - Région d'origine
    - …



Quelles données récolter?

- Types de données
  - Discrètes, Ordinales (classement)
  - Exemples:
    - Année de naissance
    - Mois
    - Sévérité de la pathologie
    - …

Quelles données récolter?

- Types de données
  - Continues
  - Exemples:
    - Poids, Taille, Pression artérielle, Taux de lipides, …
    - Age
    - Nombre de bactéries
    - Durée de survie
    - …

Quelles données récolter?

- Quantité de données
  - Coûts
  - Objectifs de la récolte
  - Taille des effets à mettre en évidence
    - Exemple

      Le régime X augmente-t-il le poids moyen adulte de 1 ou de 5 kilo(s) dans cette race de porcs ?
    - Voir Chebyshev pour la relation N(effet) et voir la notion de « puissance statistique »

Un exemple de récolte

- Un scientifique souhaite étudier l'effet de la leptine sur l'obésité avec un modèle « souris »

Article | Published: 06 March 2019

Dysregulation of a long noncoding RNA reduces leptin leading to a leptin-responsive form of obesity

Olof S. Dallner, Jill M. Marinis, Yi-Hsueh. Lu, Kivanc Birsoy, Emory Werner, Gulya Fayzikhodjaeva, Brian D. Dill, Henrik Molina, Arden Moscati, Zoltán Kutalik, Pedro Marques-Vidal, Tuomas O. Kilpeläinen, Niels Grarup, Allan Linneberg, Yinxin Zhang, Roger Vaughan, Ruth J. F. Loos, Mitchell A. Lazar & Jeffrey M. Friedman

Nature Medicine **25**, 507–516(2019) | Cite this article

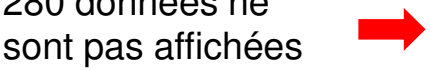
9834 Accesses | **18** Citations | **75** Altmetric | Metrics

Gène de la leptine ok

Gène de la leptine ko !

Un exemple de récolte

280 données ne sont pas affichées



Numéro	Race	Age	Sexe	Poids (g)	Etat sanitaire
1	Lept KO	23	F	76,6	+
2	Lept OK	21	F	24,1	+
3	Lept KO	20	F	82,2	+
4	Lept KO	14	F	74,6	--
5	Lept KO	14	M	90,5	+
6	Lept OK	14	M	31,7	-
7	Lept KO	23	F	71,4	-
8	Lept OK	21	F	34,7	++
9	Lept OK	13	F	23,3	++
10	Lept OK	23	M	37,8	--
291	Lept OK	22	M	31,1	-
292	Lept KO	11	M	98,6	+
293	Lept OK	23	F	7,3	++
294	Lept OK	7	M	39,9	-
295	Lept OK	10	M	31,4	-
296	Lept OK	18	F	14	--
297	Lept OK	16	M	39,1	+
298	Lept KO	12	F	89,4	-
299	Lept OK	6	F	29,8	--
300	Lept OK	17	F	34,4	-

Un exemple de récolte (suite)

- Les questions posées sont les suivantes (entre autres questions possibles...):
  - Y a-t-il des différences de poids entre les souris ayant un gène de la leptine actif et les autres ?
  - Y a-t-il un effet du sexe sur l'état sanitaire ?
  - Y a-t-il un effet du type génétique sur l'état sanitaire ?
  - L'effet éventuel du gène de la leptine est-il le même chez les mâles et les femelles ?
  - ...

Représentation des données

Numéro	Race	Age	Sexe	Poids (g)	Etat sanitaire
1	Lept KO	23	F	76,6	+
2	Lept OK	21	F	24,1	+
3	Lept KO	20	F	82,2	+
4	Lept KO	14	F	74,6	--
5	Lept KO	14	M	90,5	+
6	Lept OK	14	M	31,7	-
7	Lept KO	23	F	71,4	-
8	Lept OK	21	F	34,7	++
9	Lept OK	13	F	23,3	++
10	Lept OK	23	M	37,8	--

R **S** **C**

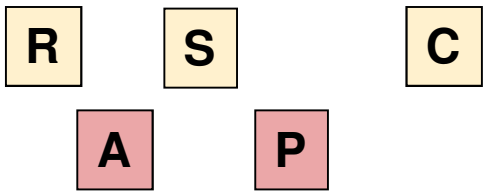
A **P**

Représentation des données

- Dans chaque « colonne » de données, on a:
  - Une mesure de même type (discret, continu),
  - Des valeurs variables de ligne en ligne,
  - Les valeurs de chaque ligne ne peuvent pas être prévues de manière exacte (déterministe), mais bien de manière probabiliste: elles sont aléatoires.
- On parlera dès lors de variables aléatoires, discrètes ou continues.

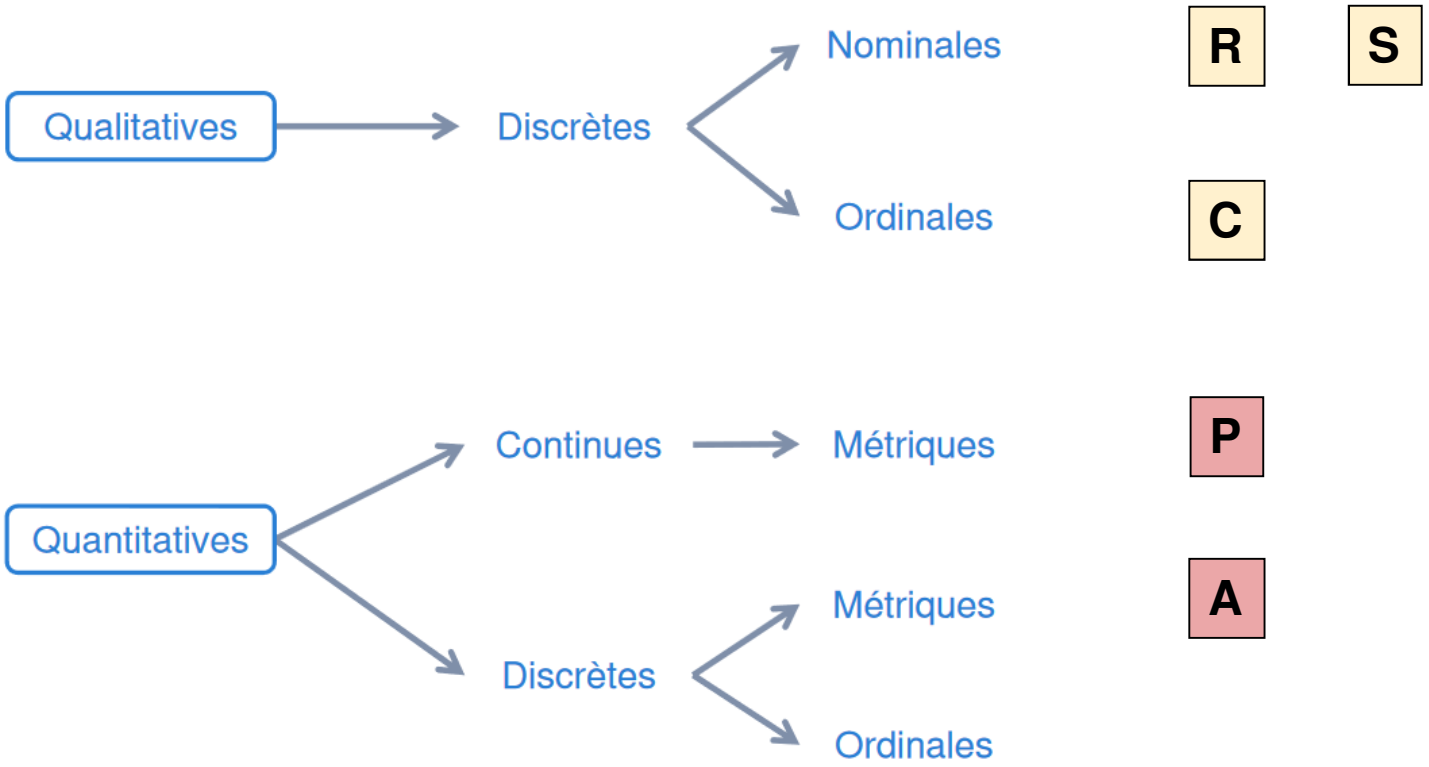
Représentation des données

Numéro	Race	Age	Sexe	Poids (g)	Etat sanitaire
1	Lept KO	23	F	76,6	+
2	Lept OK	21	F	24,1	+
3	Lept KO	20	F	82,2	+
4	Lept KO	14	F	74,6	--
5	Lept KO	14	M	90,5	+
6	Lept OK	14	M	31,7	-
7	Lept KO	23	F	71,4	-
8	Lept OK	21	F	34,7	++
9	Lept OK	13	F	23,3	++
10	Lept OK	23	M	37,8	--



- On décrit donc cet ensemble de données via 5 variables aléatoires (3 discrètes, 2 continues)

Classification des variables aléatoires



Comment caractériser les variables aléatoires ?

- 4) Par l'ensemble Ω (fini ou infini) des valeurs qu'elles peuvent prendre
  - Race: Ω = (BBB, Charolais, Croisé)
  - Sexe: Ω = (Male, Femelle)
  - Conformation: Ω = (- - , - , + , ++)
  - Taille: $\Omega = R_+$
  - Poids: $\Omega = R_+$
  - Age: $\Omega = R_+$

Pourquoi caractériser les variables aléatoires ?

- Les procédures à mettre en place pour manipuler les variables aléatoires varient en fonction du type de variable
  => nécessité d'identifier les variables aléatoires du problème et leur type.
  - Exemple: représentation des « distributions » de variables aléatoires (voir dias suivantes)

Distributions

- Une distribution est une description complète des variables aléatoires (cfr plus loin).
  - On donne une description complète d'une variable aléatoire en spécifiant une fonction (distribution) qui associe à chacune de ses valeurs dans Ω une (densité de) probabilité (voir plus loin)

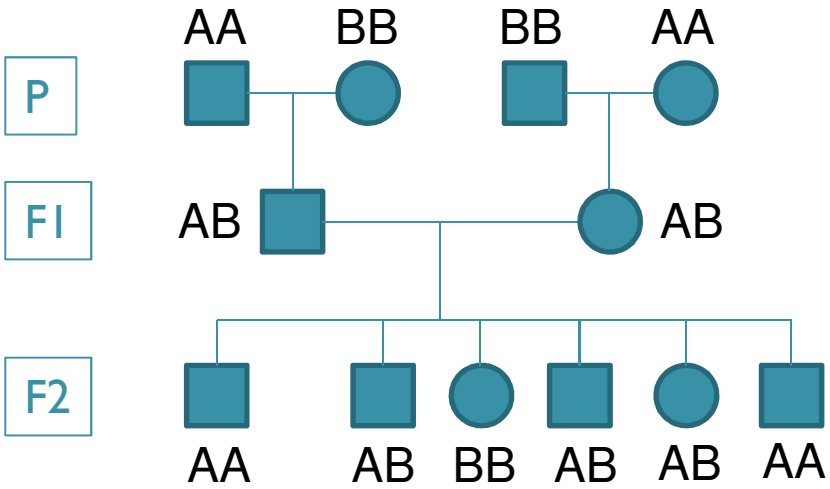
    $d: \Omega \rightarrow R: X \rightarrow Proba(X)$

Distributions théoriques et empiriques

- Une distribution obtenue sur base de considérations théoriques est appelée « distribution théorique ». Elle inclut en général des paramètres qui seront supposés connus ou devront être estimés (cfr suite).
- Une distribution obtenue sur base d'un échantillon est appelée « distribution empirique »

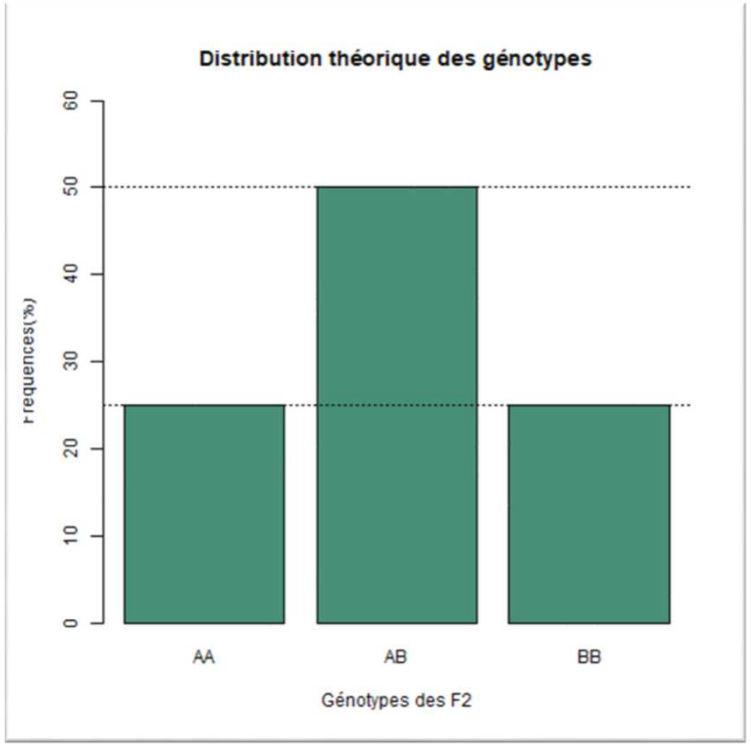
Des exemples de distributions

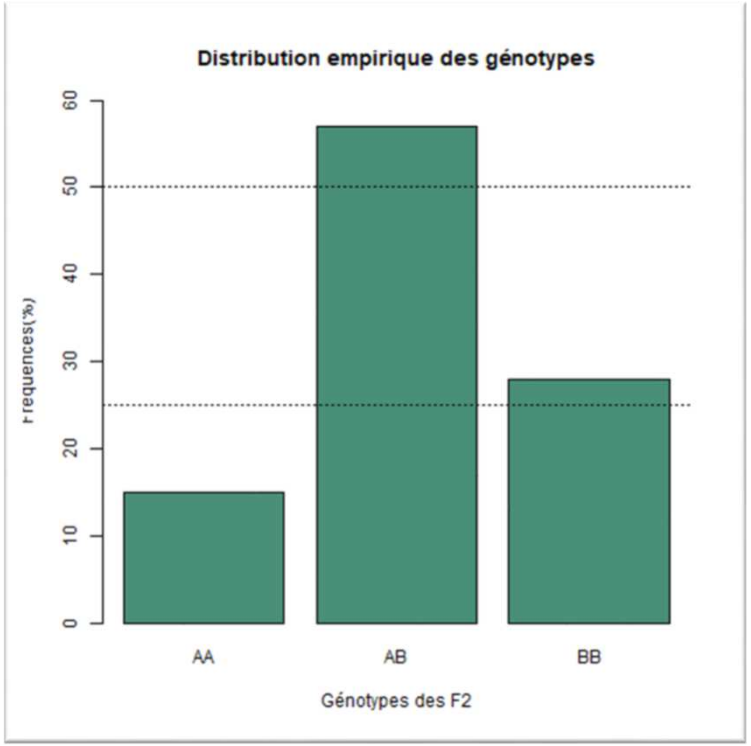
- Croisement F2



Des exemples de distributions

- Croisement F2: distributions



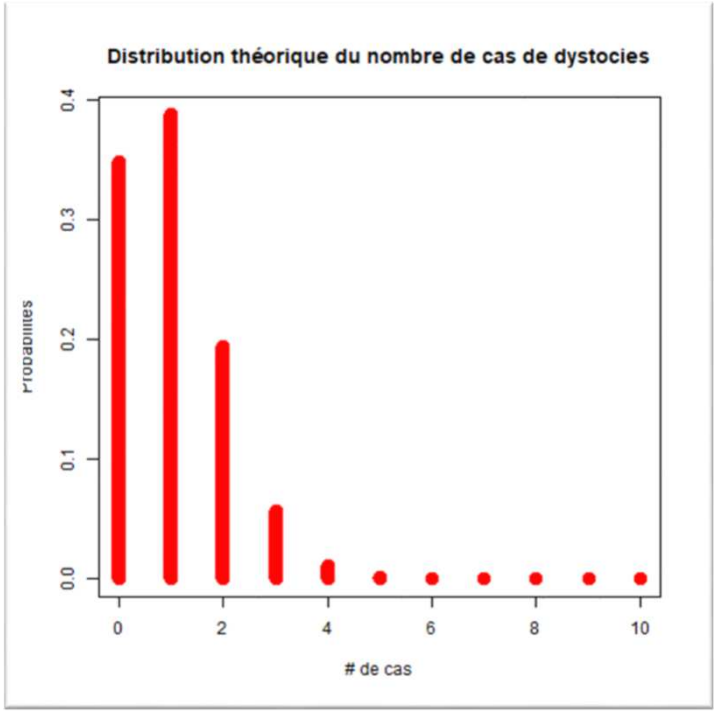


Des exemples de distributions

- Dystocies lors de 10 vêlages
  - On a estimé le nombre moyen de cas de dystocies à 1 sur 10 dans cette population
  - On peut en déduire la distribution théorique (voir, plus loin, la distribution binomiale)
  - On peut aussi obtenir une distribution empirique en examinant le résultat des 10 derniers vêlages dans 100 exploitations, par exemple.

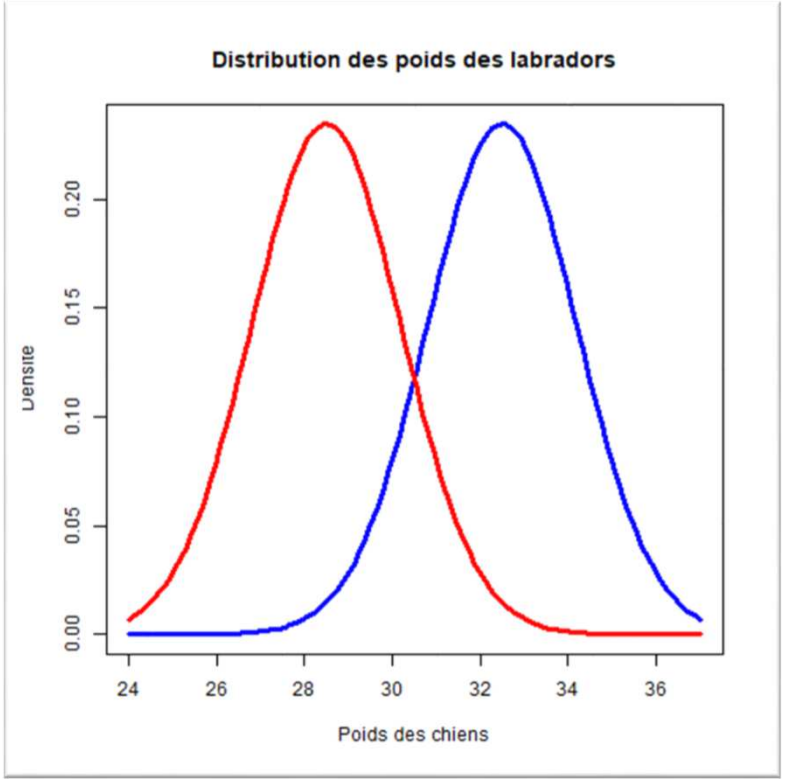
Des exemples de distributions

- Dystocies lors de 10 vêlages.



Des exemples de distributions

- Autre exemple de distribution
  - Poids des chiens d'une race donnée à un âge fixé



Mâles
Femelles

Représentation des variables aléatoires (données) ?

- Représentation synthétique
  - Tables de fréquences
- Représentation graphique
  - Diagrammes de fréquences
- Paramètres descriptifs
  - Position
  - Dispersion
  - Aplatissement, asymétrie, …

Tables de fréquences

- Table à N entrées , dont les cases contiennent les effectifs correspondants.
  - N est le nombre de variables utilisées
  - Exemple: Table de fréquences de la robe dans un échantillon de 200 labradors

(table à 1 entrée)

Chocolat	Dorée	Noire
24	114	62

Comment obtenir la table de fréquences ?

```
# Répertoire de travail
setwd("d:/docsusers/cours/2020-2021/bmv1/stats")
# Lecture des données
t<-read.table(file="labradors.txt",
              head=T,sep="\t",dec=« .")
names(t)
[1] "robe" "sexe" "poids" "confo"
# Table de fréquences de la robe
table(t$robe)

Chocolat    Dorée    Noire
      24      114       62
```

Table de fréquences

– Exemple: Table de fréquences de la robe X sexe (table à 2 entrées)

	Chocolat	Dorée	Noire	
♂	12	53	28	93
♀	12	61	34	107
	24	114	62	200

Comment obtenir la table de fréquences ?

```
#
# Table de fréquences de robe X sexe
#
table(t$sexe,t$robe)

   Chocolat Dorée Noire
F       12    53    28
M       12    61    34
```

Table de fréquences

– Exemple: Table de fréquences de la robe X sexe X conformation (table à 3 entrées)

♀ / ♂	Chocolat	Dorée	Noire	
- -	1	14	6	21
-	2	14	12	28
+	3	18	12	33
+ +	6	15	4	25
	12	61	34	107

Table de fréquences cumulées

– Exemple: Table de fréquences (simples et cumulées) de l'état sanitaire (table à 1 entrée)

	Simples	Cumulées
- -	87	87
-	85	172
+	87	259
+ +	41	300
	300	

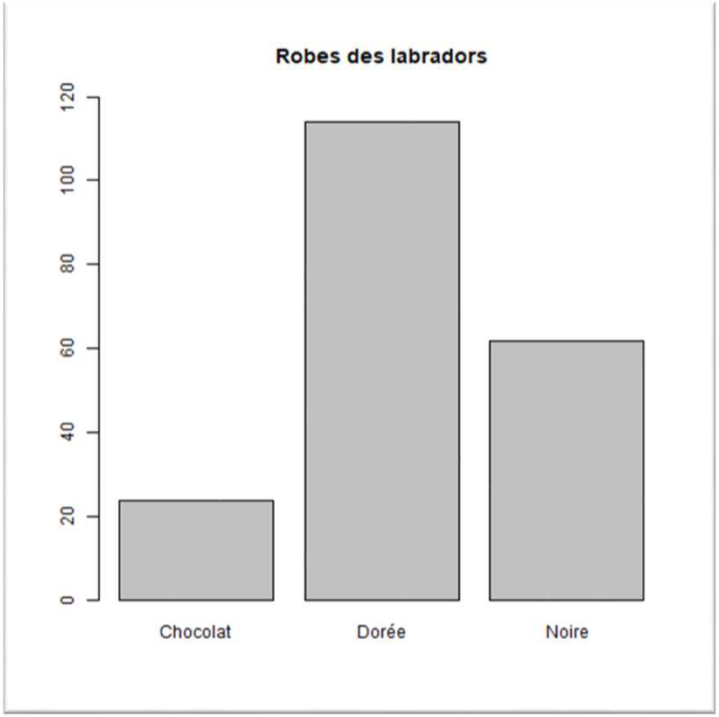
Comment obtenir la table de fréquences cumulées ?

```
#
# Table de fréquences cumulées
#
> l<-read.table(file="leptine.txt",
+ header=T,sep="\t")
> tb<-table(l$Etat_sanitaire)
> tc<-tb
> for (i in 2:4) { tc[i]<-tb[i]+tc[i-1] }
> tc

  -  --   +  ++
 87 172 259 300
>
```

Représentations graphiques

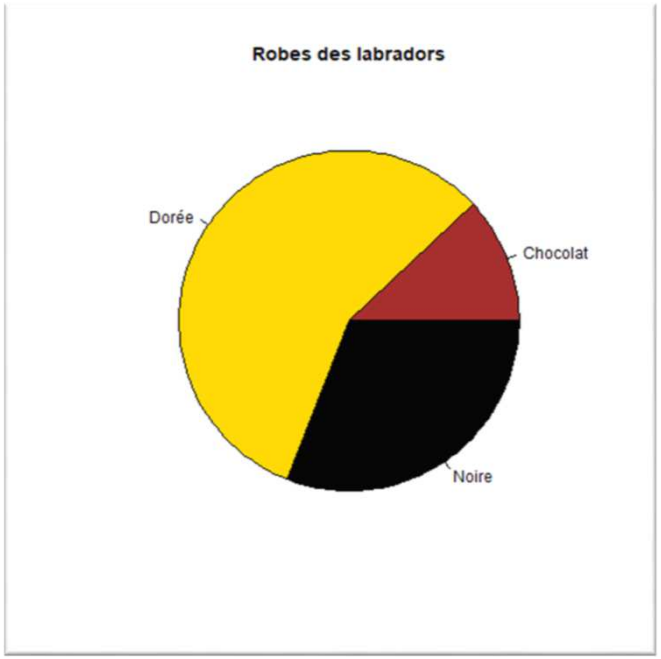
– Exemple: Diagrammes à bâtons de la fréquence des différentes robes de labradors.




```
plot(t$robe,ylim=c(0,120),
main="Robes des labradors")
```

Existe-t-il d'autres représentations graphiques ?

- Oui, beaucoup d'autres… (voir notes).
  - Exemples: Diagramme en tarte (*Pie chart*)




```
pie(table(t$robe),
main="Robes des labradors",
col=c("brown","gold","black"))
```

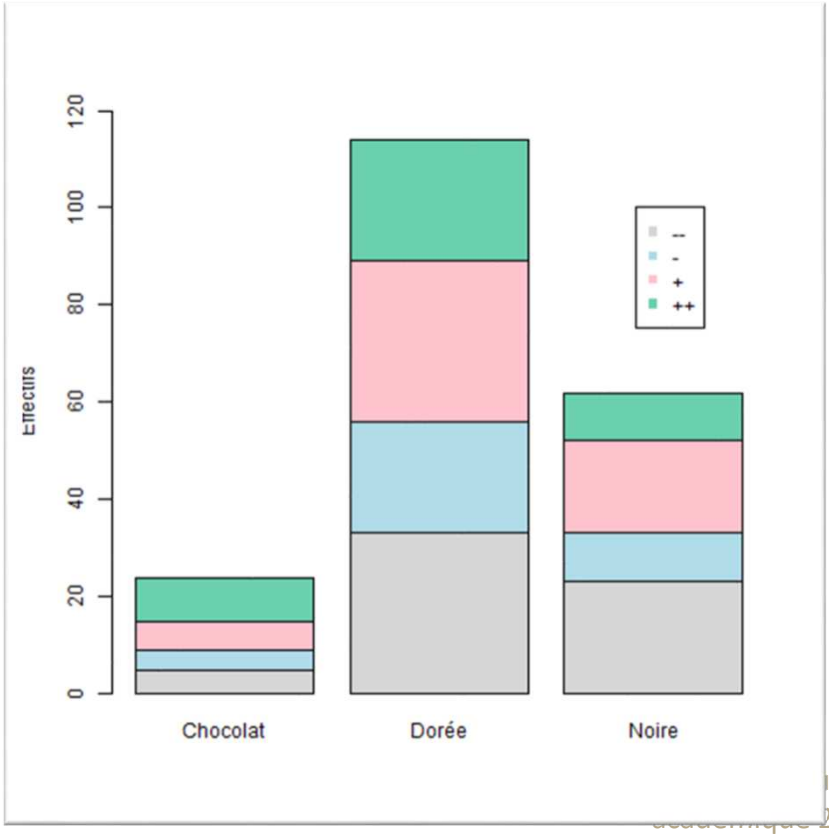
Existe-t-il d'autres représentations graphiques ?

– Exemples: Diagrammes en tuyaux d'orgue

```
> barplot(
+ apply(table(t$confo,t$robe,t$sexe),c(1,2),sum),
+ col=c("lightgrey","lightblue","pink","aquamarine3"),
+ ylim=c(0,120), ylab="Effectifs")
> legend(3,100,
+ col=c("lightgrey","lightblue","pink","aquamarine3"),
+ legend=c("--","-","+","++"),
+ pch=15)
```

Existe-t-il d'autres représentations graphiques ?

– Exemples: Diagrammes en tuyaux d'orgue



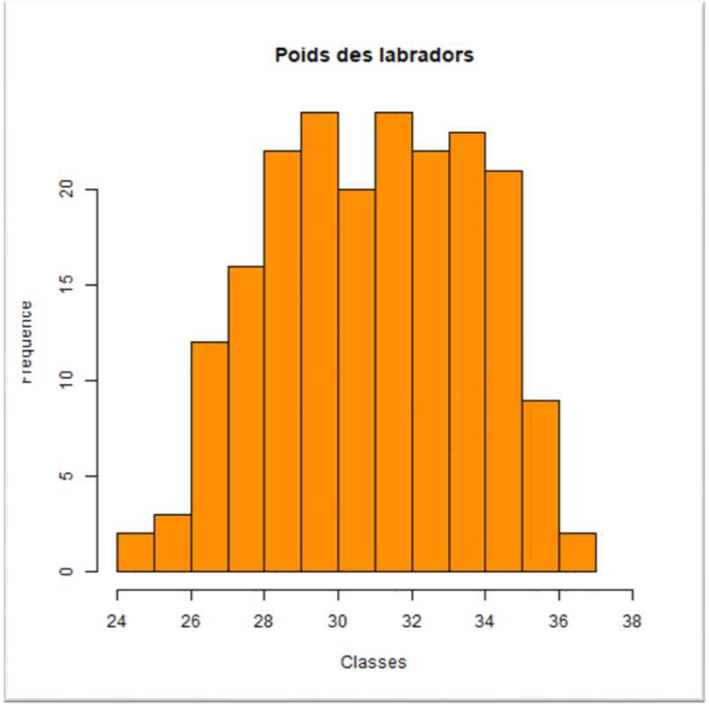
Comment procéder avec des variables continues ?

- Discrétisation
  - Exemple I: l'âge exprimé en mois est une discrétisation de l'âge.
  - Exemple II: création de classes de poids
    - Classe I: Poids entre 24 et 26 kilos
    - Classe II: Poids entre 26 et 28 kilos
    - Classe III: Poids entre 28 et 30 kilos
    - …
    - Classe VII: Poids entre 36 et 38 kilos

Comment procéder avec des variables continues ?

– Exemple II (suite): représentation graphique

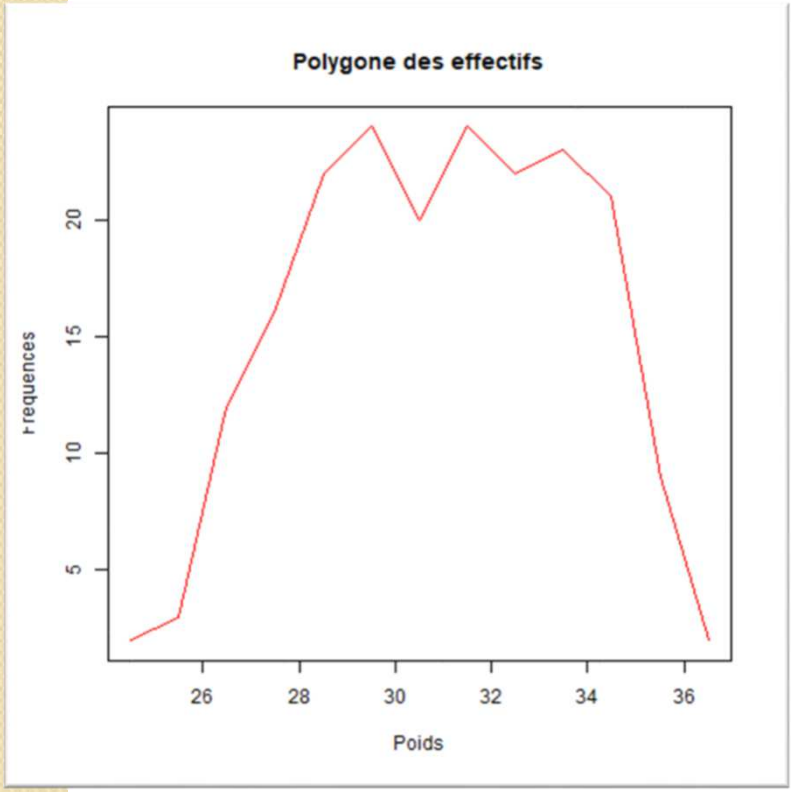
Histogramme




```
hist(df$poids,
main="Poids des labradors",
xlab="Classes",
xlim=c(24,38),
ylab="Fréquence",
col="darkorange")
```

D’autres représentations de variables continues ?

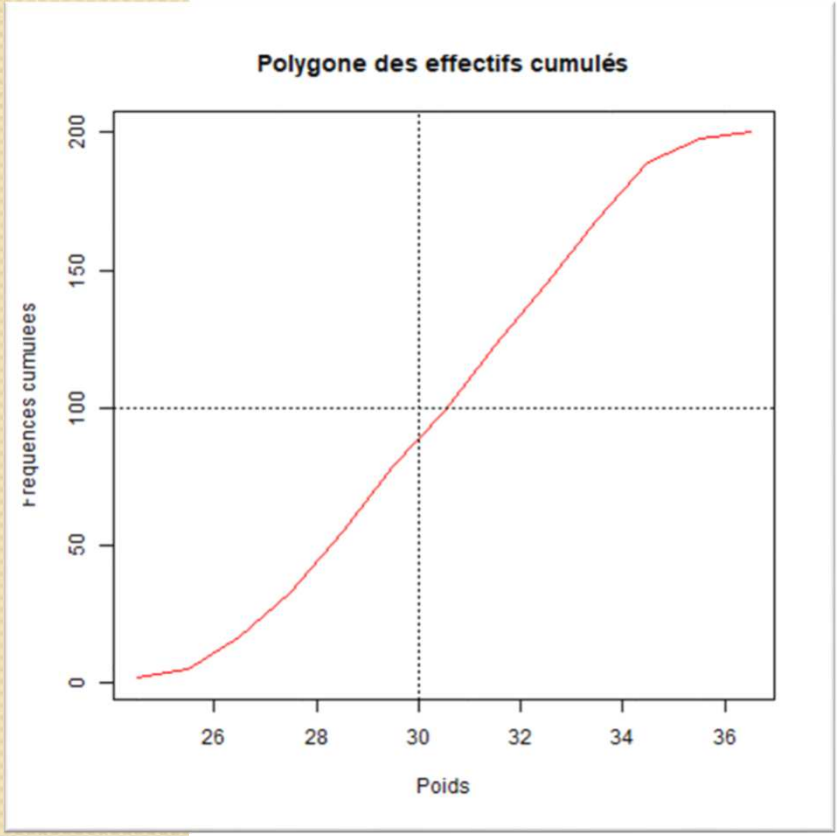
– Exemple III : Polygone des effectifs




```
h<-hist(df$poids,plot=FALSE)
plot(h$mids,h$counts,type="l",
 col="red",xlab="Poids",
 ylab="Fréquences",
 main="Polygone des effectifs")
```

D’autres représentations de variables continues ?

– Exemple IV : Polygone des effectifs cumulés




```
n<-length(h$mids)
cumul<-rep(0,n)
cumul[1]<-h$counts[1]
for (i in 2:n) {
 cumul[i]<-cumul[i-1]+h$counts[i]
}
plot(h$mids,cumul,type="l",
 col="red",xlab="Poids",
 ylab="Fréquences cumulées",
 main="Polygone des eff cumulés")
abline(v=30,lty=3)
abline(h=100,lty=3)
```

D’autres représentations de variables continues ?

– Exemple V : Diagramme « *stem-leaves* »

Stem	Leaves
000	444555899900
100	1112344444445556677777888999 90
700	1123

Le même diagramme, avec

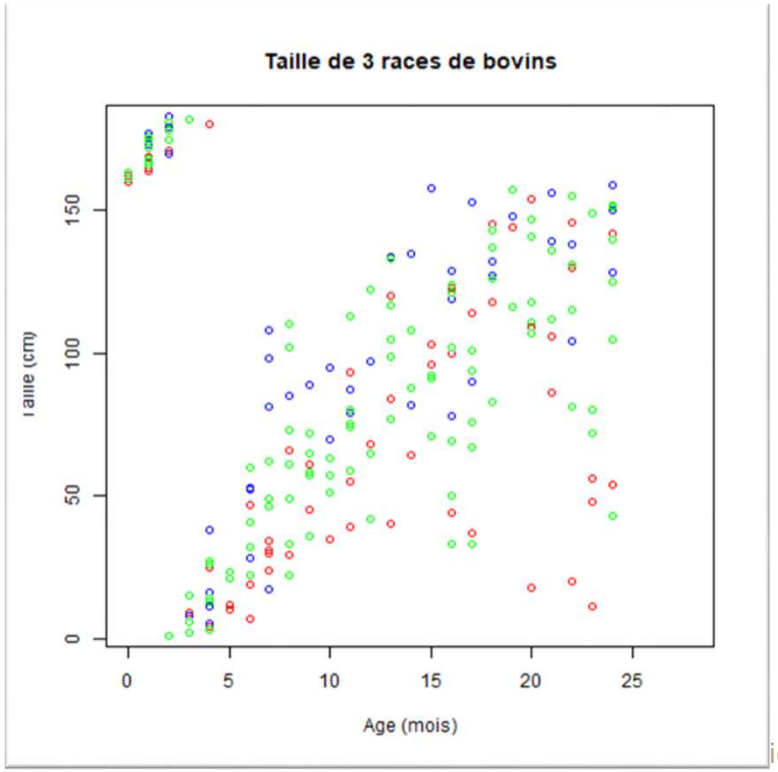
```
> stem(t$Poids)

 The decimal point is at the |

  24 | 8
  25 | 0579
  26 | 233445579
  27 | 0001134466799
  28 | 000000122233455667788889999
  29 | 00111122334455556888999
  30 | 0011112223333334566689
  31 | 12223334445556677 78
  32 | 000001111233345555678889 99
  33 | 01233333344445566777788
  34 | 0111111233444566668999
  35 | 134444568
  36 | 36
```

Des représentations avec plusieurs variables continues ?

– Exemple VI : Diagramme de dispersion (*scatter plot*)



En résumé

- Les statistiques servent à interpréter les données récoltées dans une expérience.
- Les données sont représentées par des variables aléatoires
- Une description des données débute très souvent par une représentation de l'échantillon récolté, sous forme de graphiques ou de tables de fréquences