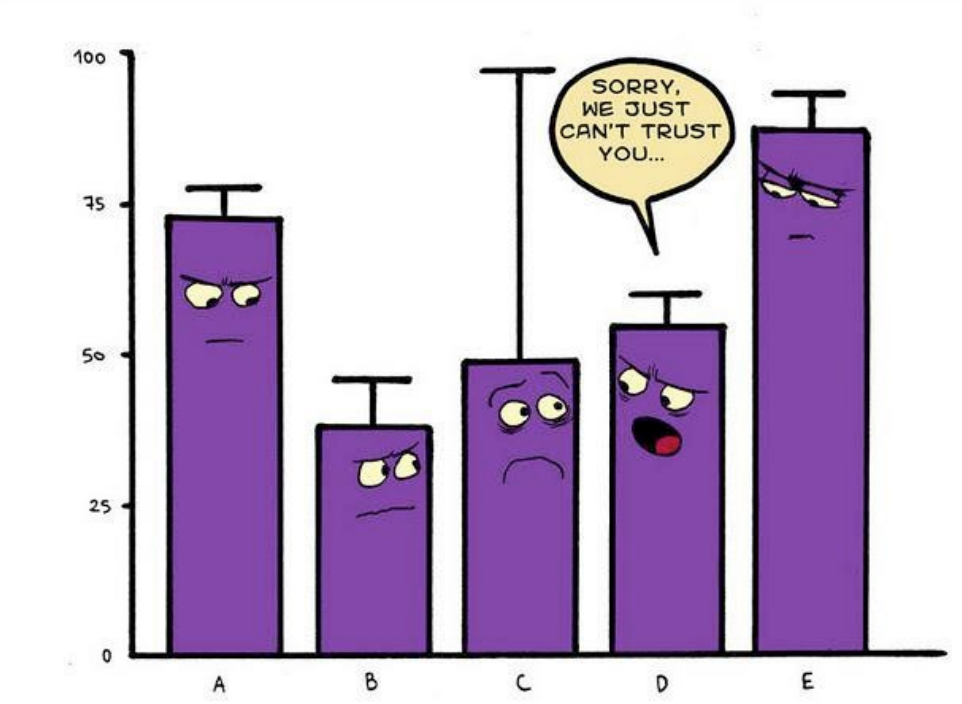
Statistiques : quelques notions de base



http://andrewgelman.com/2011/12/16/suspicious-histograms/

G. San Martin

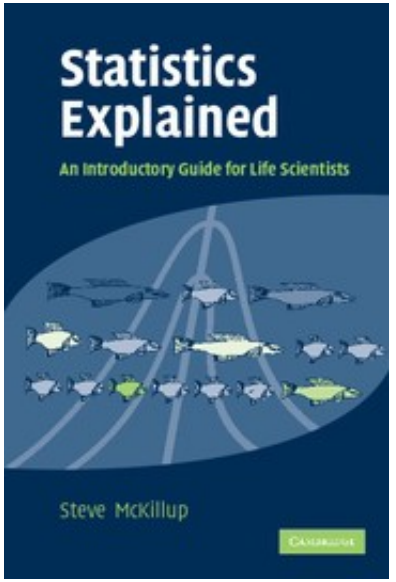
gilles.sanmartin@gmail.com
Centre Wallon de Recherche Agronomique

cra-w

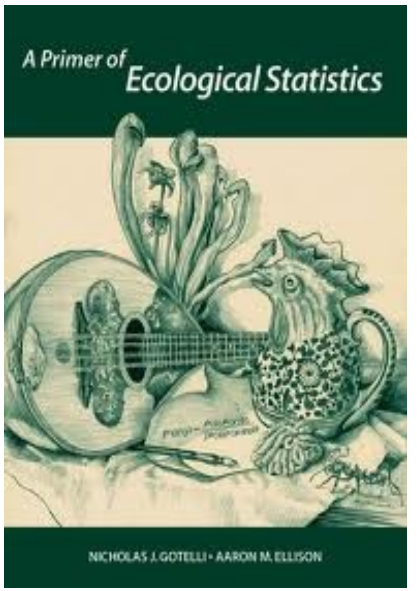
Quelques livres

Il existe de nombreux livres d'introduction aux stats avec des approches très différentes.
3 principaux livres utilisés ici :

Sokal & Rohlf
Très complet
tous les calculs détaillés
didactique mais long

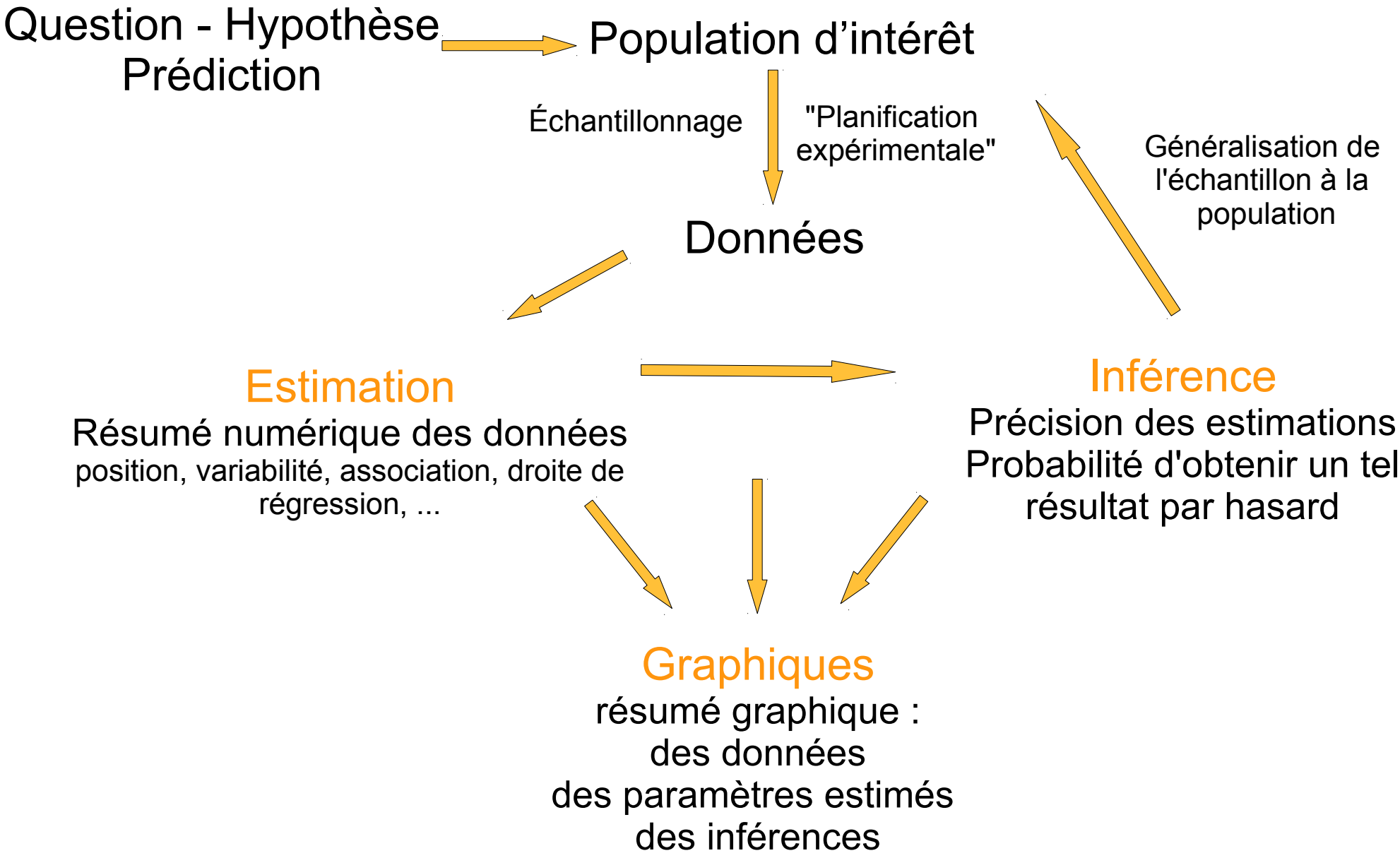


McKillup
Similaire à Sokal & Rohlf
sans les calculs détaillés
axé sur la compréhension
court



Gotelli & Ellison
En particulier pour la partie
sur les designs expérimentaux

Les Statistiques : Pourquoi ?



Partie 1 :
Introduction aux concepts de base

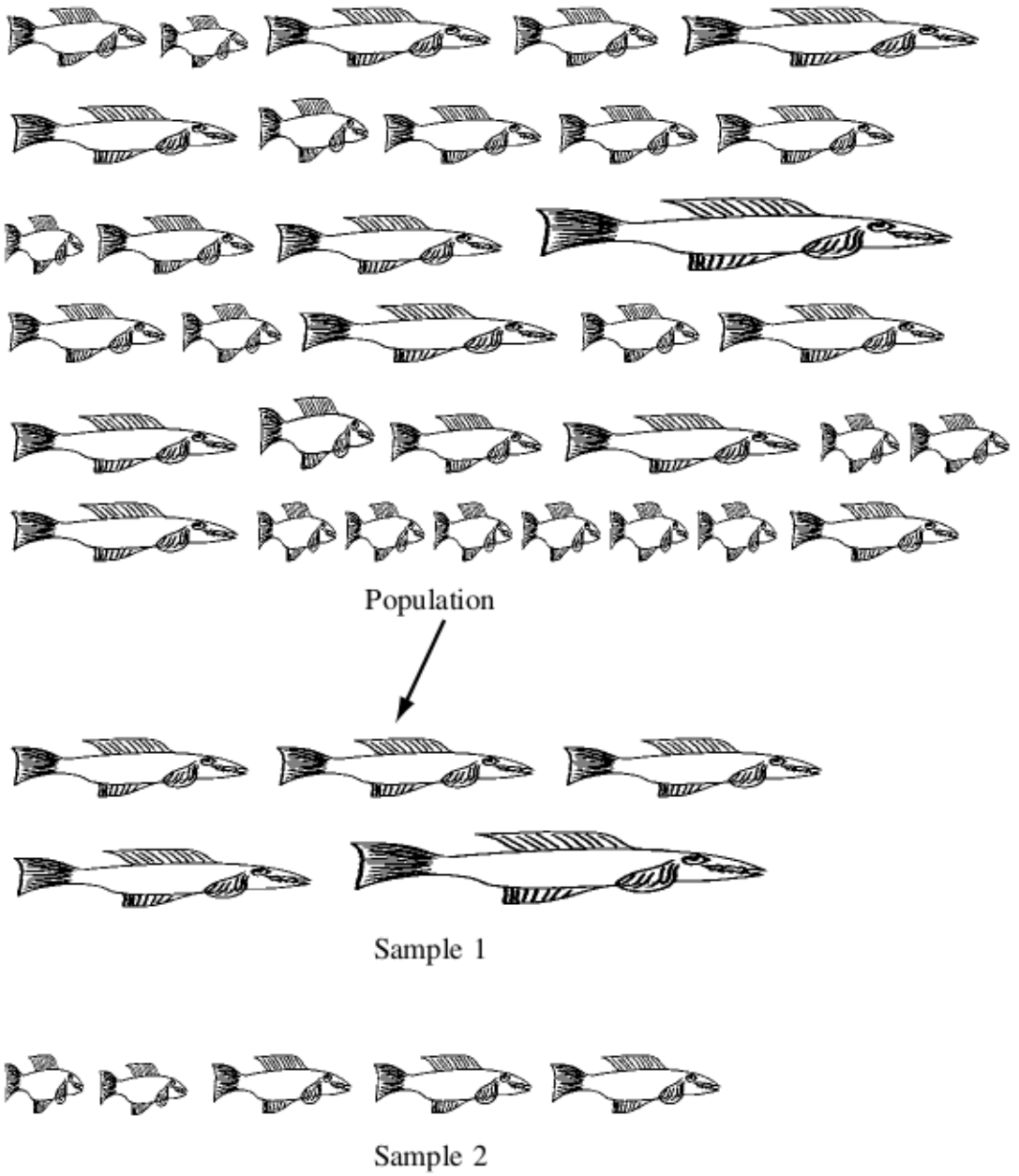
Quelques définitions

Population statistique :
ensemble des individus /objets sur lesquels porte l'étude

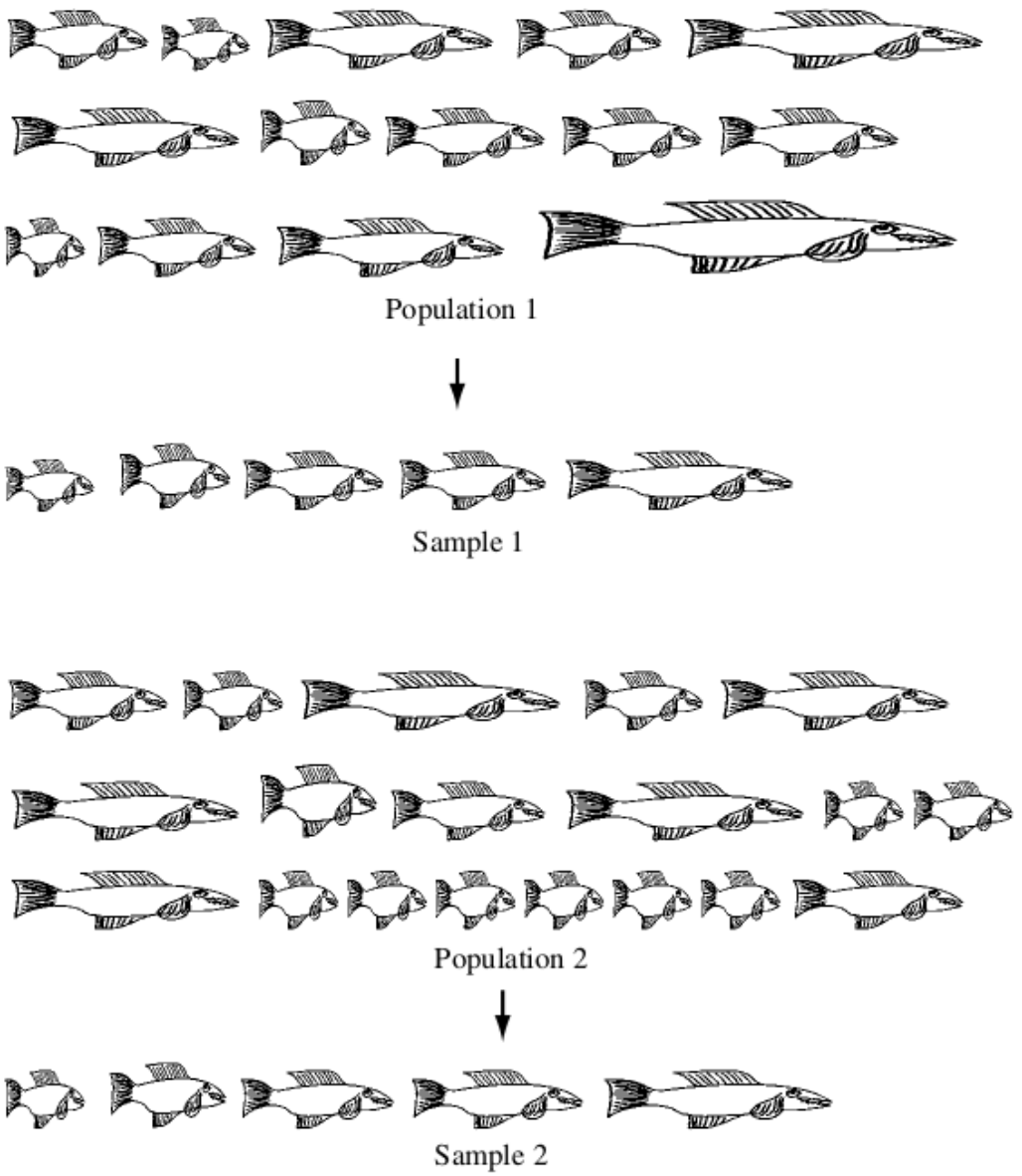
On a jamais accès à toute la population
--> on extrait un échantillon
--> on essaye d'inférer les caractéristiques de la population totale sur base de son échantillon

Il y a presque toujours une différence entre les paramètres estimés sur base de l'échantillon et le paramètre réel de la population

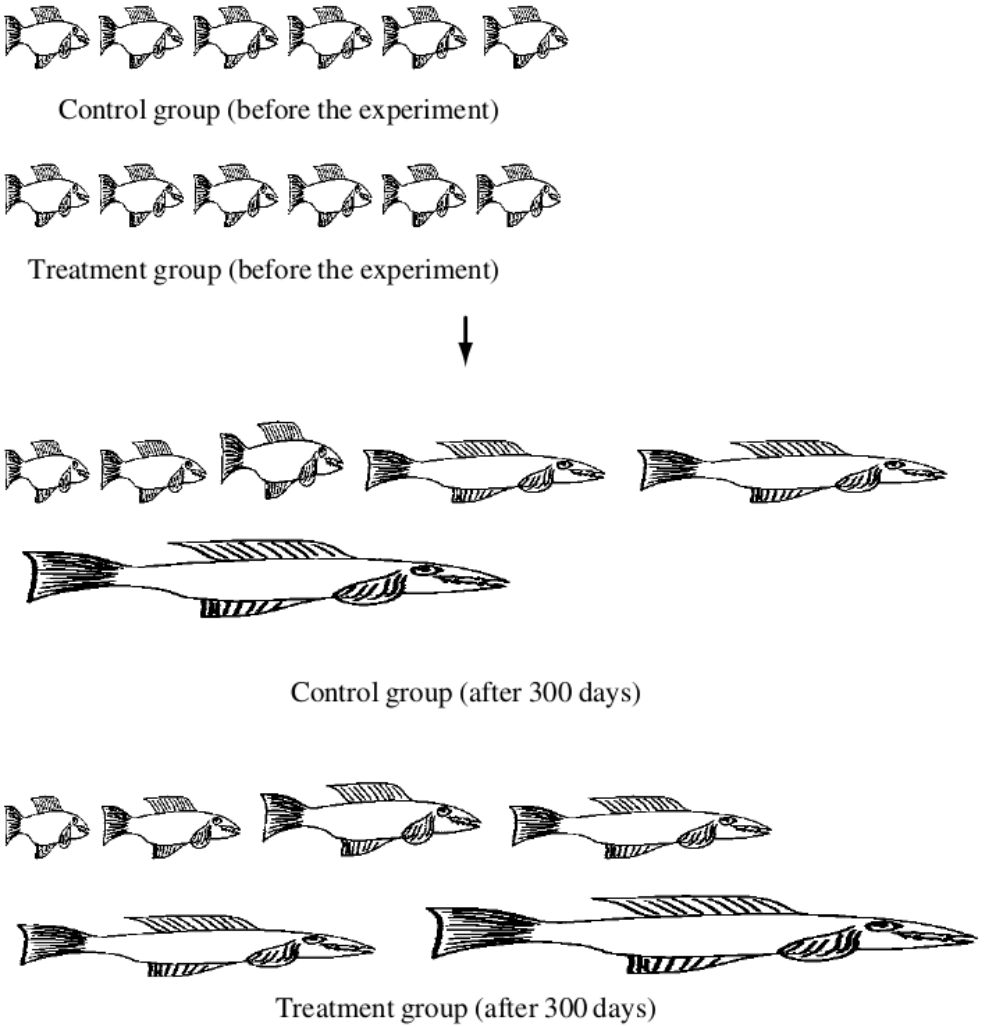
Il y a presque toujours une différence entre plusieurs échantillons



2 échantillons issus de la même population peuvent être différents



2 échantillons issus de 2 populations différentes peuvent être similaires



La variabilité entre individus peut masquer l'effet d'un traitement

Quelques définitions

Variable statistique = l'attribut que l'on étudie / mesure

On distingue en général :

Variables quantitatives (= "numériques") :
représentées par un nombre
Variables qualitatives (= "nominales"):
pas représentées par un nombre (pex couleur)
Variables ordinales :
les valeurs peuvent uniquement servir à ordonner les observations (établir leur ordre)

Variable quantitative continue :
peut prendre toutes les valeurs entre 2 nombres
Variable quantitative discrète :
ne peut pas prendre toutes les valeurs

Quelques définitions

Probabilité :
2 approches / définitions (parmi d'autres)

Approche "fréquentiste"
La probabilité d'un événement est la fréquence à laquelle il va apparaître si on recommence l'expérience un grand nombre de fois
(pex : lancé de dés)

Approche "Bayésienne"
La probabilité ("posterior probability") est plutôt vue comme le degré de certitude que l'on a sur un événement en se basant à la fois sur ce qu'on sait déjà (prior probability) et sur les données nouvellement acquises (likelihood).

Estimation

On calcule une valeur qui décrit les propriétés de la population qui nous intéresse

On appelle cette valeur :
"paramètre", "statistique", "estimateur", "coefficient", etc..

Il en existe une infinité (on peut en inventer)...
L'estimation est souvent un simple calcul algébrique.
Dans certains cas l'estimation est plus complexe
(pex : quelle est la meilleure droite passant par un nuage de points ?)

On va passer ici en revue très rapidement quelques estimateurs fréquemment utilisés

Estimation : position

"Location estimators"

Valeur unique donnant une approximation de la "grandeur" de la variable mesurée

Moyenne

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

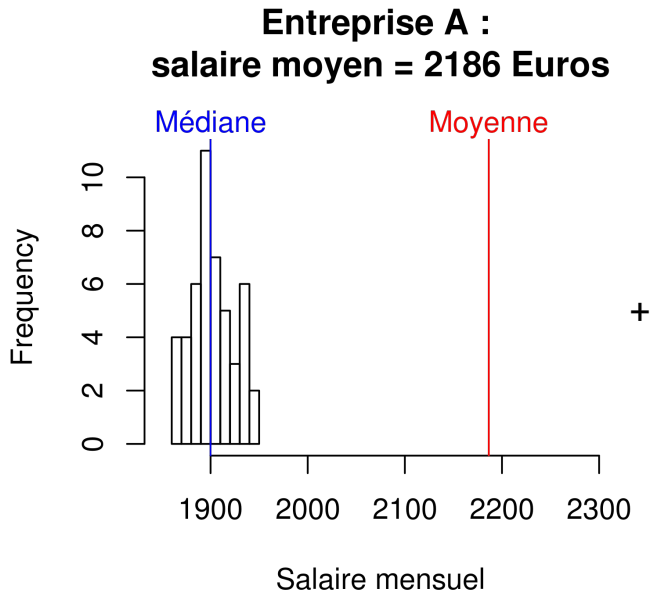
Médiane

La moitié des données ont une valeur plus petite/plus grande que la médiane

Mode

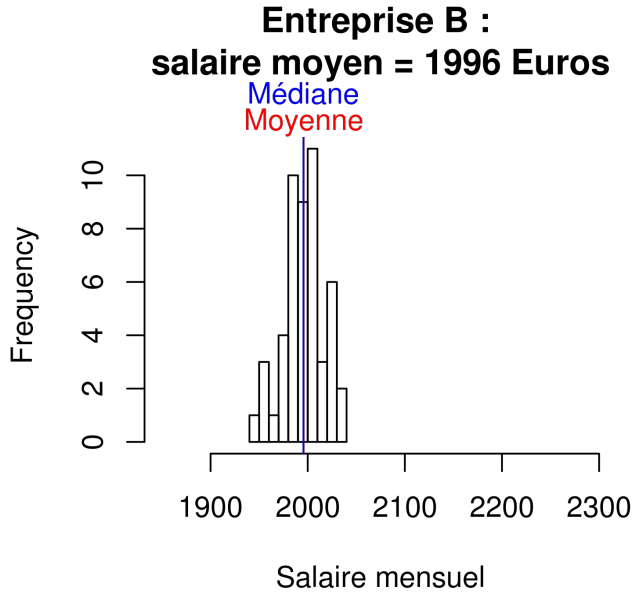
Valeur la plus fréquente

Estimation : position



Attention : la moyenne est très sensible aux valeurs extrêmes

+ Le patron : 15000 Euros
+ Le frère du patron : 10000 euros



Inspiré de Normand Baillargeon : Petit cours d'autodéfense intellectuelle

Estimation : variabilité

"scale", "dispersion", "spread" estimators

Description de la variabilité des données/de la population

Basés sur l'écart des observations par rapport à la moyenne

Variance

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

Ecart type

= "Standard deviation"

≠ "Standard error" !!!!

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

si 1/n : estimateur biaisé

"Range"

= valeurs minimales et maximales

Quantiles

Sur le même principe que la médiane pour n'importe quel % des données (percentiles, déciles, quartiles,...)

pex : premier décile : 10 % des données sont en dessous de cette valeur et 90 % au dessus

Estimation : association

Décrit le degré d'association, de dépendance entre 2 variables

Covariance

$$Cov_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

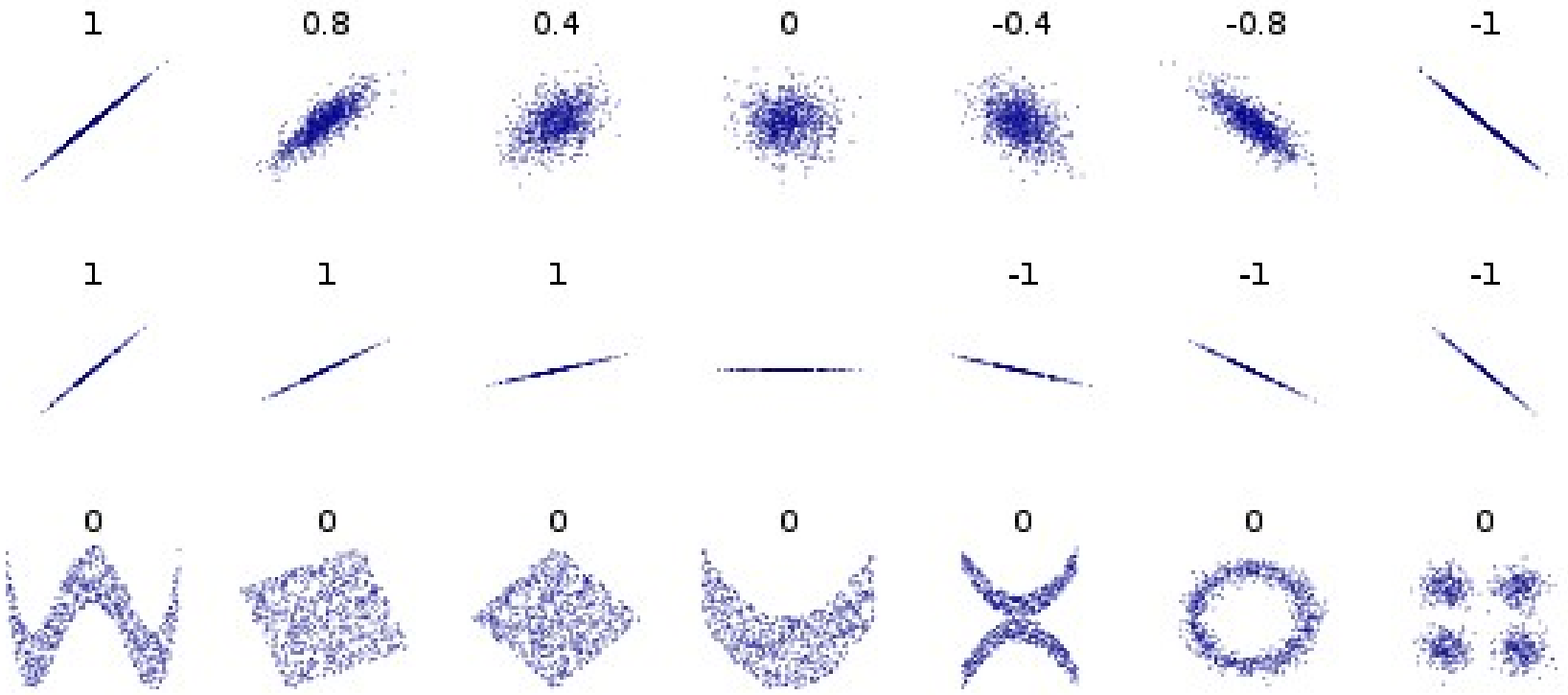
Mesurent l'association <u>linéaire</u> entre 2 variables

Corrélation de Pearson

= version standardisée de la covariance (sans unités)
= covariance divisée par le produit des écarts types
compris entre -1 et +1
Ne pas confondre avec le R^2 (coefficient de détermination) même si ils sont liés

$$R_{xy} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Estimation : association

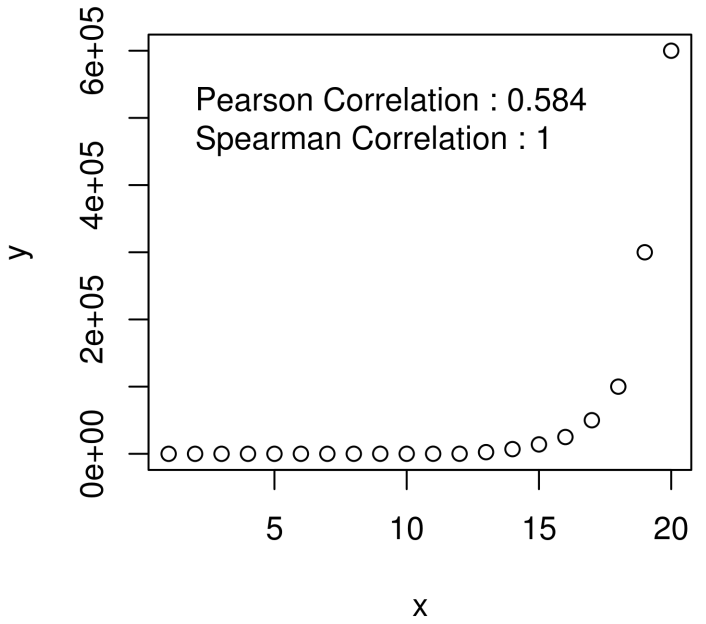


Estimation : association

Corrélation de rang de Spearman

Relation monotone (pas obligatoirement linéaire) entre 2 variables
Équivaut à calculer la corrélation de Pearson entre les rangs des 2 variables

```
> x <- c(1:20)
> y <- c( 3, 5, 7, 32, 38, 40, 41, 82, 82, 84, 99, 100, 2500, 7000, 14000, 25000,
    50000, 100000, 300000, 600000)

> cor(x, y)
[1] 0.5837021

> cor(x, y, method = "spearman")
[1] 0.999624

> cor(rank(x), rank(y))
[1] 0.999624
```



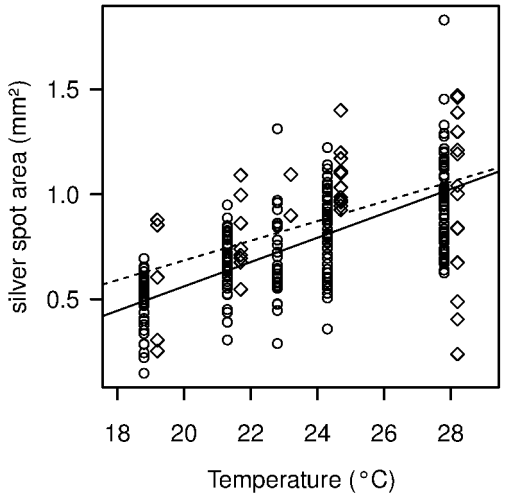
NB : il existe de nombreuses autres statistiques d'association ea pour mesurer l'association entre variables nominales

Estimation : régression

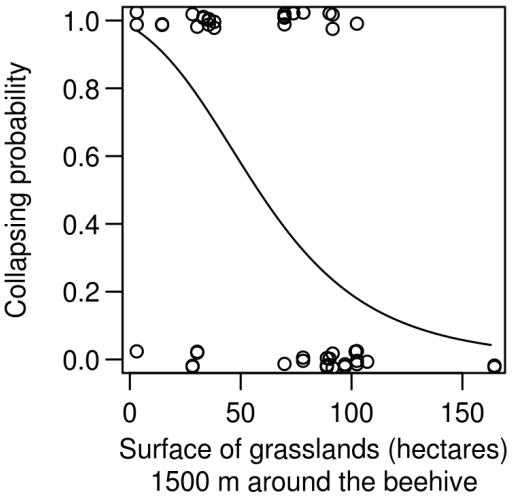
Le but de la régression est de prédire la valeur d'une variable en fonction d'une (ou plusieurs) autre(s)

Cela revient à faire passer une droite
(ou un plan, ou une courbe) par un nuage de points
On estime les paramètres (pente, intercept,...) qui décrivent cette droite (ou plan, ou courbe,...)

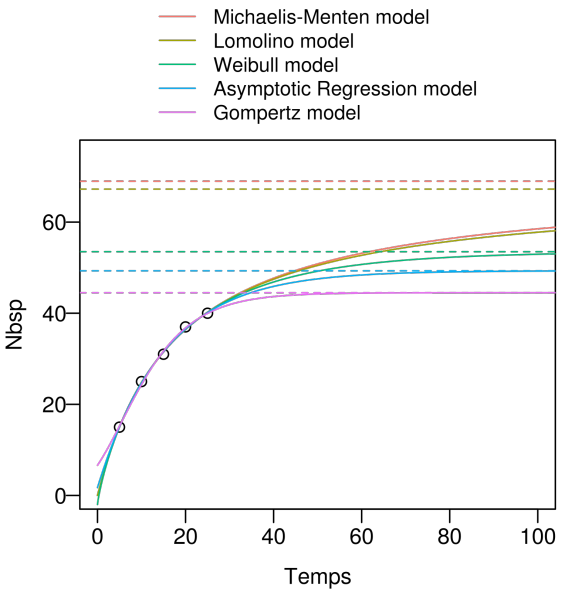
Deux méthodes principales d'estimation :
Moindres carrés
Maximum de Vraisemblance



Régression linéaire



Régression logistique



Régressions non linéaires

Inférence : précision des paramètres

On utilise les **paramètres calculés sur les échantillons**
pour estimer les **paramètres de la population**

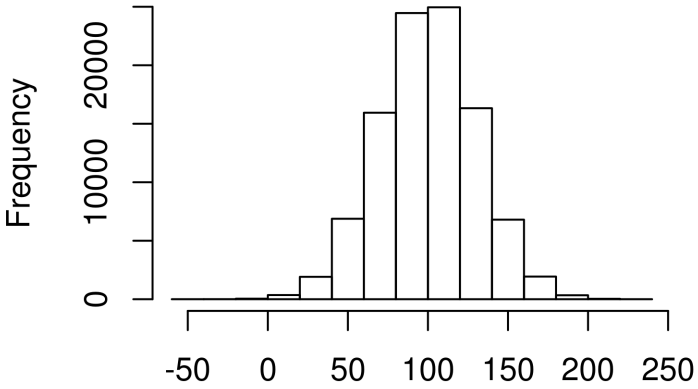
MAIS :

Les valeurs estimées dans les échantillons sont variables
(à cause du hasard de l'échantillonnage)
On a besoin d'estimer leur précision pour savoir
à quel point on peut avoir confiance en ces estimations
et à quel point on peut les extrapoler à l'ensemble de la population

-> c'est le rôle de :
l'**erreur standard**
l'**intervalle de confiance**

Inférence : précision des paramètres

On génère une population de 100000 individus
de moyenne "mu" = 100 et d'écart-type "sigma" = 30




```
> pop <- rnorm(100000, 100, 30)
>
> # 5 échantillons de taille = 5
> ech1 <- sample(pop, 5)
> ech2 <- sample(pop, 5)
> ech3 <- sample(pop, 5)
> ech4 <- sample(pop, 5)
> ech5 <- sample(pop, 5)
>
```

On prélève 5 échantillons
aléatoires de 5 individus

```
> mean(ech1) ; mean(ech2) ; mean(ech3) ; mean(ech4) ; mean(ech5)
[1] 99.10138
[1] 87.84037
[1] 101.8613
[1] 87.04577
[1] 109.5411
> sd(ech1) ; sd(ech2) ; sd(ech3) ; sd(ech4) ; sd(ech5)
[1] 36.63038
[1] 24.44459
[1] 29.94871
[1] 42.03631
[1] 33.99683
```

On peut calculer la moyenne
et l'écart-type
de chaque échantillon

On obtient 5 valeurs
différentes entre échantillons
et différents des "vraies" moyennes
et écart-types de la population

Inférence : précision des paramètres

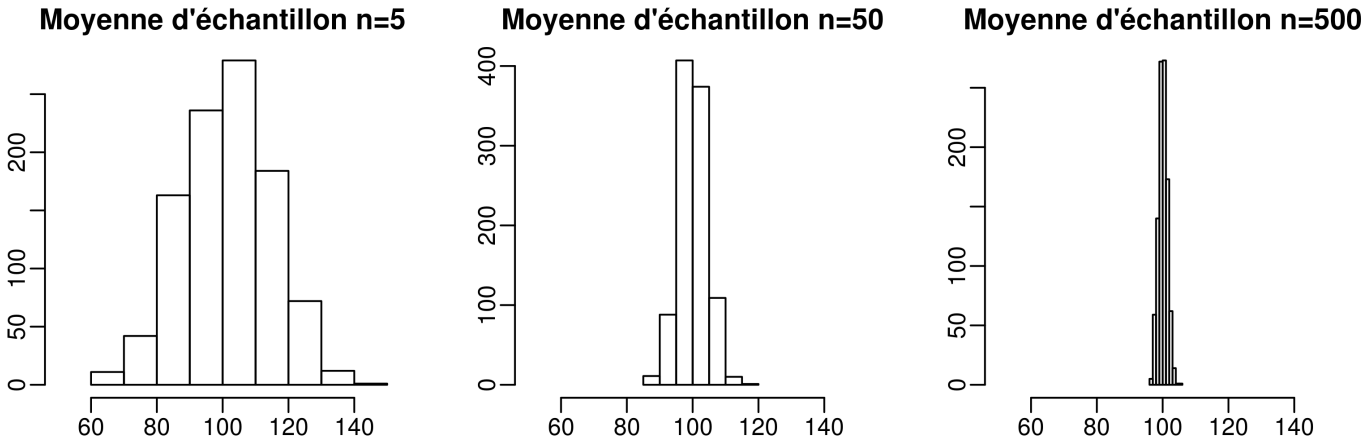
Erreur standard

Il s'agit de l'écart type des moyennes des échantillons

et de manière plus générale l'écart-type de n'importe quel paramètre estimé

La précision dépend de la variabilité de la population
et de la taille des échantillons

```
# 1000 échantillons de taille différentes
mu_hat_5 <- replicate(n = 1000, mean(sample(pop,5)))
mu_hat_50 <- replicate(n = 1000, mean(sample(pop,50)))
mu_hat_500 <- replicate(n = 1000, mean(sample(pop,500)))

> sd(mu_hat_5) ; sd(mu_hat_50) ; sd(mu_hat_500)
[1] 13.38564
[1] 4.182749
[1] 1.323881
```



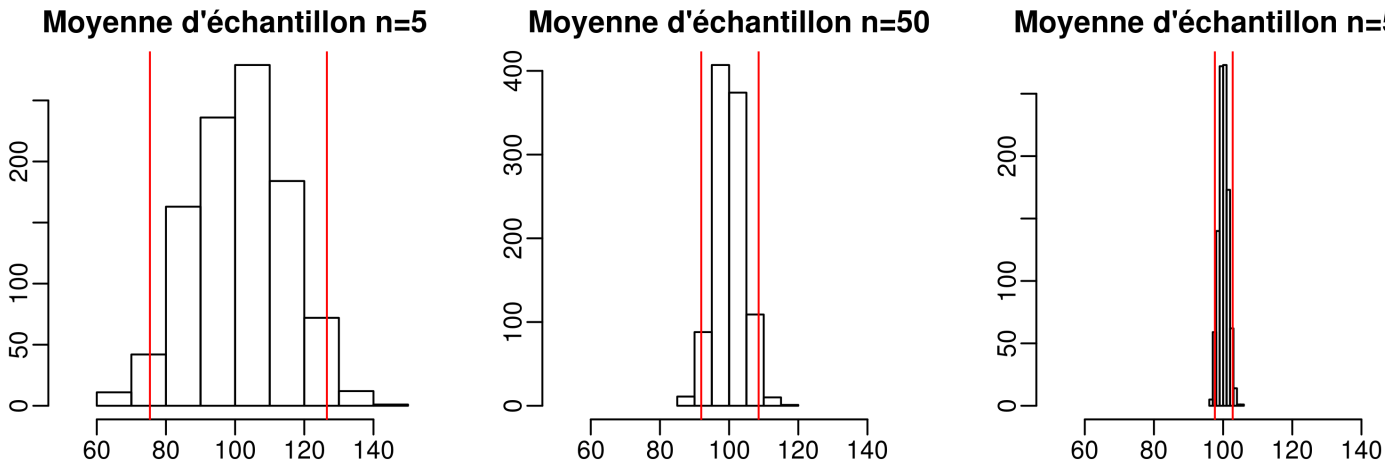
Inférence : précision des paramètres

Intervalle de confiance à 95 %

Représente les bornes (quantiles) entre lesquelles se trouvent 95 % des moyennes d'échantillons après élimination des 2.5 % des valeurs les plus petites et 2.5 % des valeurs les plus grandes

```
> quantile(mu_hat_5, probs = c(0.025, 0.975))
     2.5%     97.5%
 75.36634 126.53596
> quantile(mu_hat_50, probs = c(0.025, 0.975))
     2.5%     97.5%
 91.93753 108.53296
> quantile(mu_hat_500, probs = c(0.025, 0.975))
     2.5%     97.5%
 97.61462 102.77148
```

Un intervalle de confiance de [75.36, 126.53] signifie que la moyenne d'un échantillon (de taille 5 pour cette population) a 95 % de chance de se trouver entre 75.36 et 126.53



Inférence : précision des paramètres

En Pratique, on ne peut pas échantillonner 1000 fois la population.
Comment fait-on en réalité?

2 approches les plus connues :

Non parametric bootstrap :
Approche de rééchantillonnage par ordinateur
(computer intensive method, resampling)

Lois de probabilité théoriques
Approche purement mathématique
Historiquement la première, et encore la plus utilisée aujourd'hui

Autre méthode très utile :
"Fake data simulation" = "Parametric bootstrap" = "Monte-Carlo simulations"

Inférence : précision des paramètres

Bootsrap

Méthode a priori saugrenue mais très simple qui consiste à rééchantillonner aléatoirement (à l'aide d'un ordinateur) son échantillon !

1) on a par exemple un échantillon réel de 50 unités **1**
2) on prélève aléatoirement 50 unités dans cet échantillon avec la possibilité de prélever plusieurs fois la même unité **2**
3) on calcule le paramètre d'intérêt et on le stocke **3**
4) on recommence un grand nombre de fois (1000) **4**
5) On peut alors calculer l'écart-type (= erreur standard du paramètre) et/ou l'intervalle de confiance sur ces 1000 valeurs* **5**

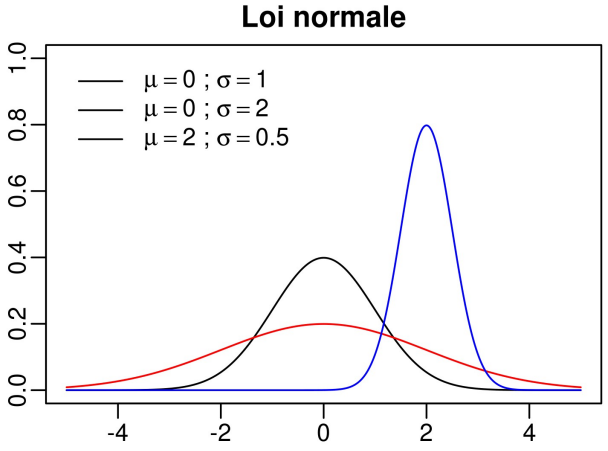
```
> ech <- sample(pop, 50) # échantillon réel 1
> mu_boot <- replicate(1000, mean(sample(ech,size=50, replace=TRUE)))
> sd(mu_boot)                    4        3                  2
[1] 4.071419
> quantile(mu_boot, probs = c(0.025, 0.975)) 5
    2.5%    97.5%
 91.20069 107.13916
```

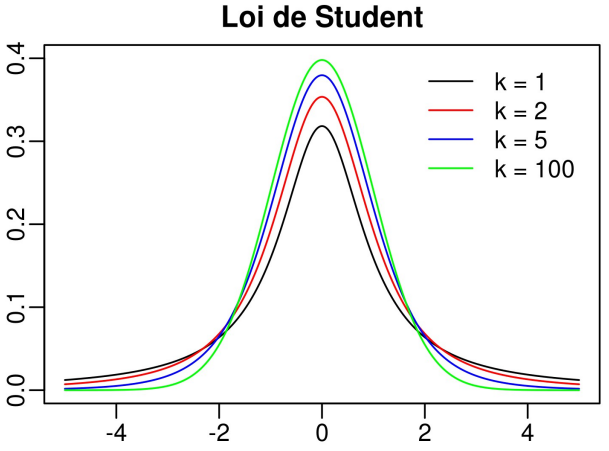

*NB : il existe des méthodes améliorées en particulier pour les intervalles de confiance

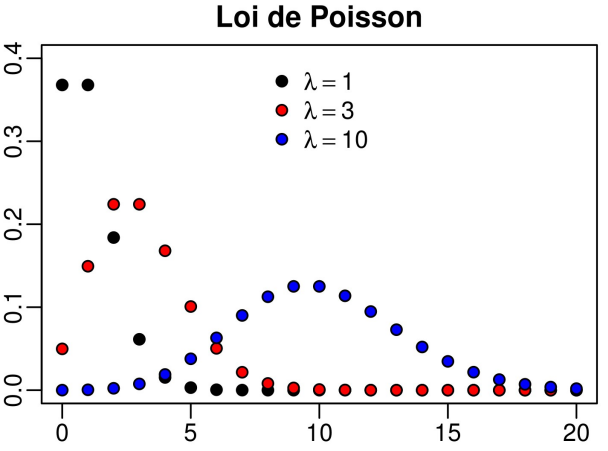
Inférence : précision des paramètres

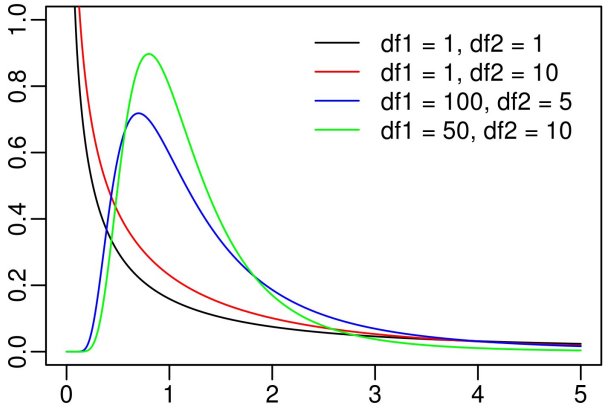
Lois de Probabilité théoriques

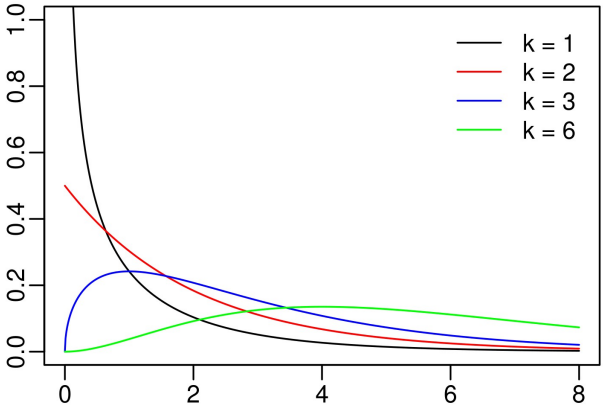
Les lois de probabilités sont caractérisées notamment par une "fonction de densité de probabilité" qui décrit la probabilité d'obtenir chaque valeur d'une variable aléatoire pour certaines valeurs des paramètres caractérisant la loi

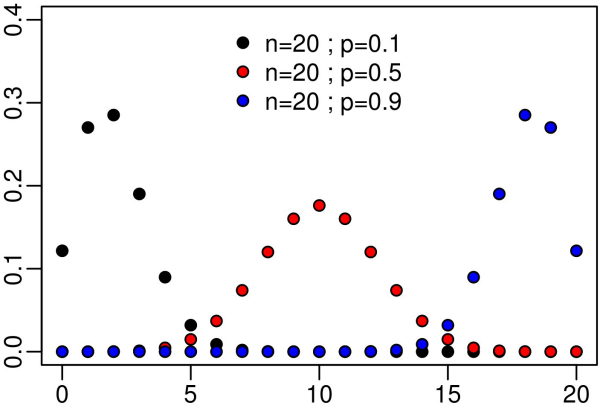










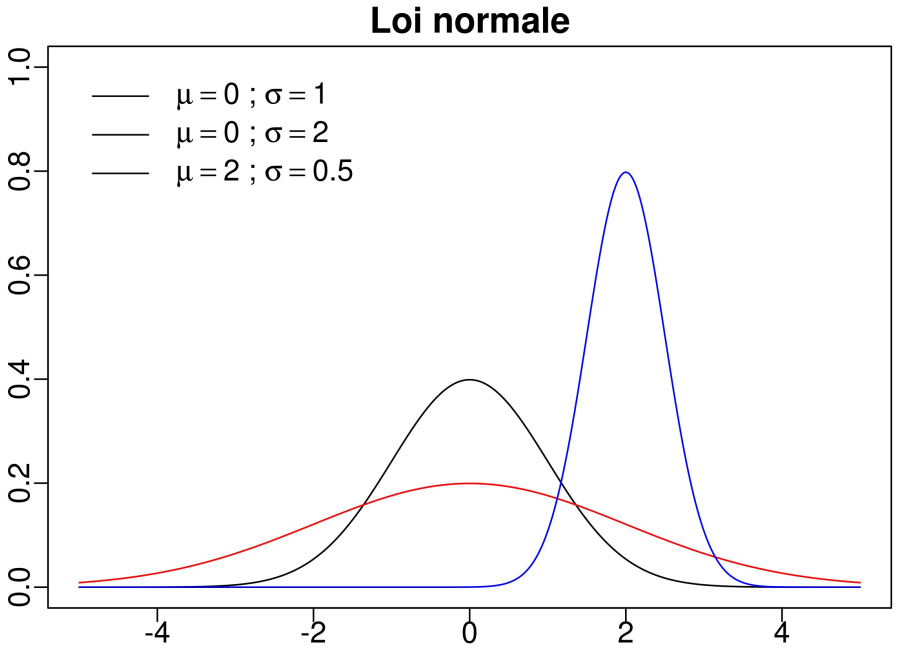


Inférence : précision des paramètres

Loi Normale

De loin la plus importante parce que :
- de nombreux phénomènes naturels sont distribués selon cette loi
- de nombreuses distributions tendent vers la loi normale dans certaines conditions (ea quand la taille d'échantillon augmente)

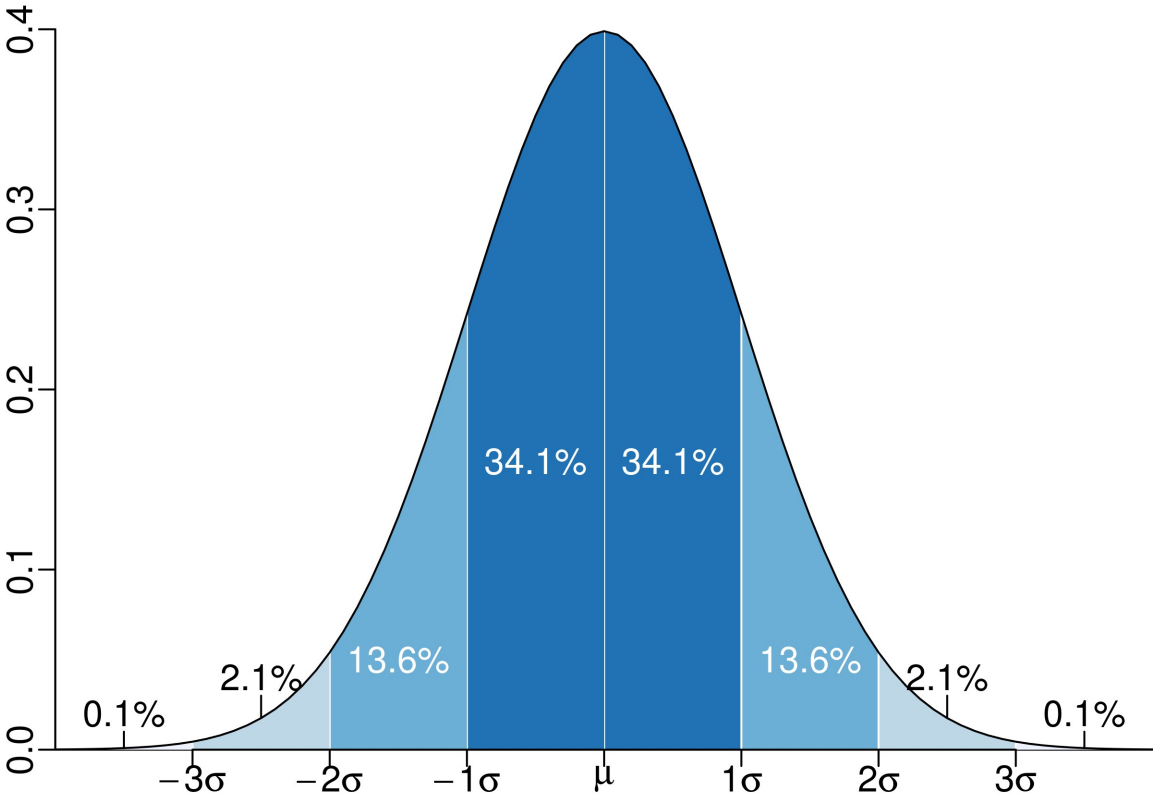
Caractérisée uniquement par 2 paramètres :
mu : la moyenne
sigma : l'écart-type



Inférence : précision des paramètres

Loi Normale

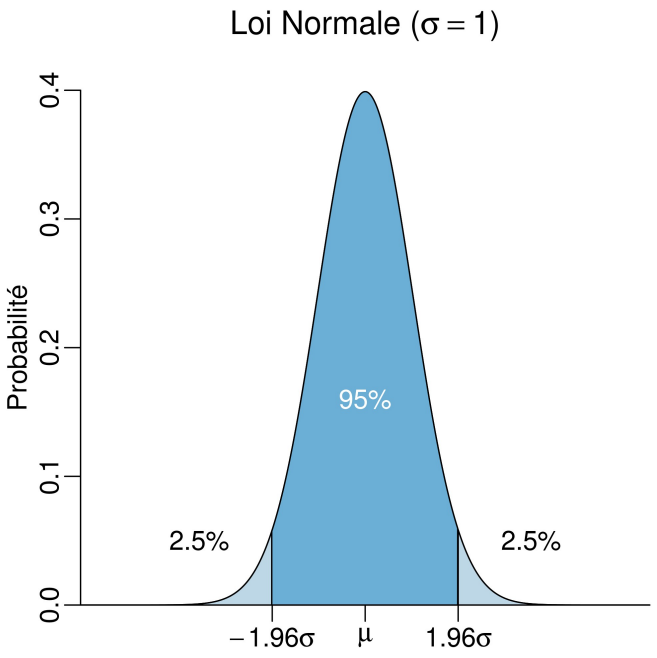
La courbe est symétrique autour de la moyenne
Et les valeurs comprises entre
mu +- 2 sigma correspondent à 95.45 % des observations
quelles que soient les valeurs de mu et sigma

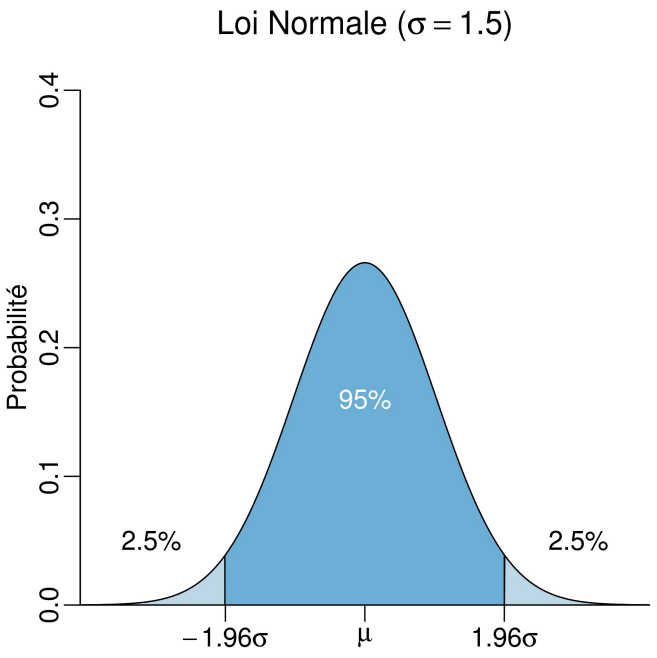


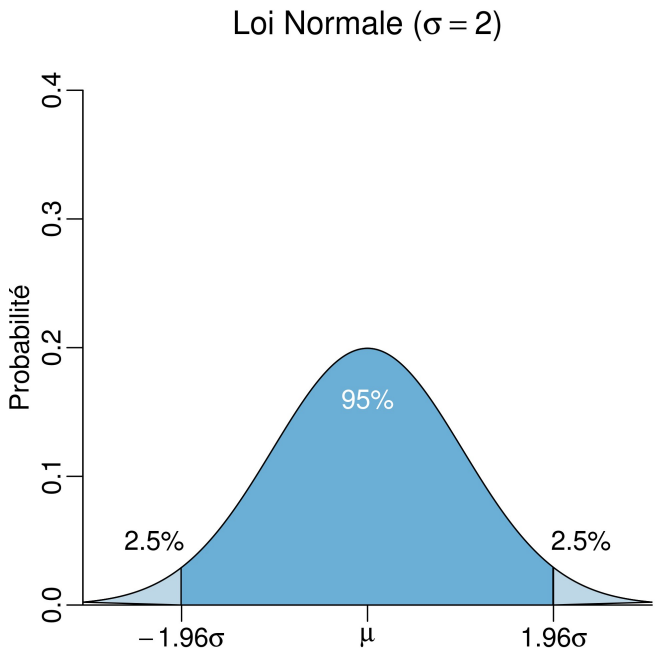
Inférence : précision des paramètres

Loi Normale

Les valeurs comprises entre
mu +- 1.96 sigma correspondent à 95 % des observations
quelles que soient les valeurs de mu et sigma







Inférence : précision des paramètres

Lois de Probabilité théoriques

On peut montrer que :
si la population d'origine a une distribution normale
et que l'échantillon est suffisamment grand*,

alors les moyennes d'échantillon
ont également une **distribution normale**
avec une **erreur standard**
égale à l'écart-type divisé par la racine carré de l'effectif :

$$se_{\bar{x}} = \frac{s}{\sqrt{n}}$$

L'**intervalle de confiance** à 95 % est donc :

Moyenne de l'échantillon

$$\bar{x} - 1.96\, se_{\bar{x}} \leqslant \mu \leqslant \bar{x} + 1.96\, se_{\bar{x}}$$

Moyenne de la population

*ou que l'on connaît la variance réelle de la population (pas de l'échantillon)

Inférence : précision des paramètres

Lois de Probabilité théoriques

On peut également calculer d'autres intervalles de confiance par exemple à 90 % ou 99 % (soit 0.9 ou 0.99 en proportion)

L'**intervalle de confiance** à 100(1-α) % peut se calculer comme :

$$\bar{x} - z_{\alpha/2}\, se_{\bar{x}} \leqslant \mu \leqslant \bar{x} + z_{\alpha/2}\, se_{\bar{x}}$$

on utilise α/2 car on élimine la moitié des valeurs dans la partie basse de la cloche et l'autre moitié dans la partie haute
z représente les "quantiles" de la loi normale c'est à dire les bornes entre lesquelles on a une proportion de 1-α des données

Dans R ces quantiles s'obtiennent avec la fonction `qnorm()`

```
> qnorm(p= 0.05/2, mean = 0, sd = 1, lower.tail=FALSE)
[1] 1.959964
> qnorm(p= 0.1/2, mean = 0, sd = 1, lower.tail=FALSE)
[1] 1.644854
> qnorm(p= 0.01/2, mean = 0, sd = 1, lower.tail=FALSE)
[1] 2.575829
```

Inférence : précision des paramètres

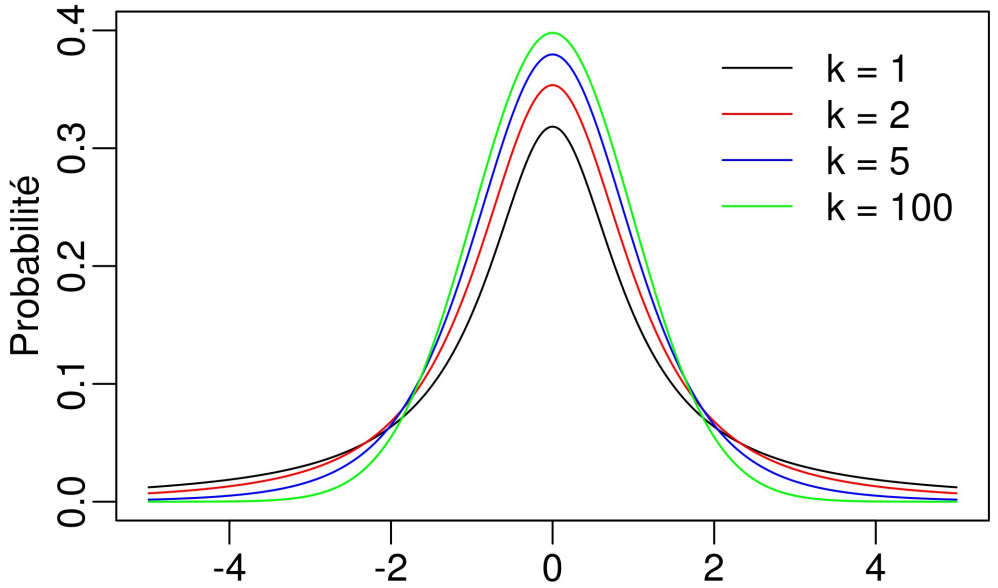
Lois de Probabilité théoriques

Si l'échantillon n'est pas suffisamment grand, il faut se référer à une autre loi proche de la normale : la loi t de Student

Cette loi a une forme différente en fonction du nombre de "degrés de liberté" qui est le nombre d'observations (n) -1

Lorsque n est très grand la distribution t de Student est identique à la normale

Loi de Student



Inférence : précision des paramètres

Lois de Probabilité théoriques

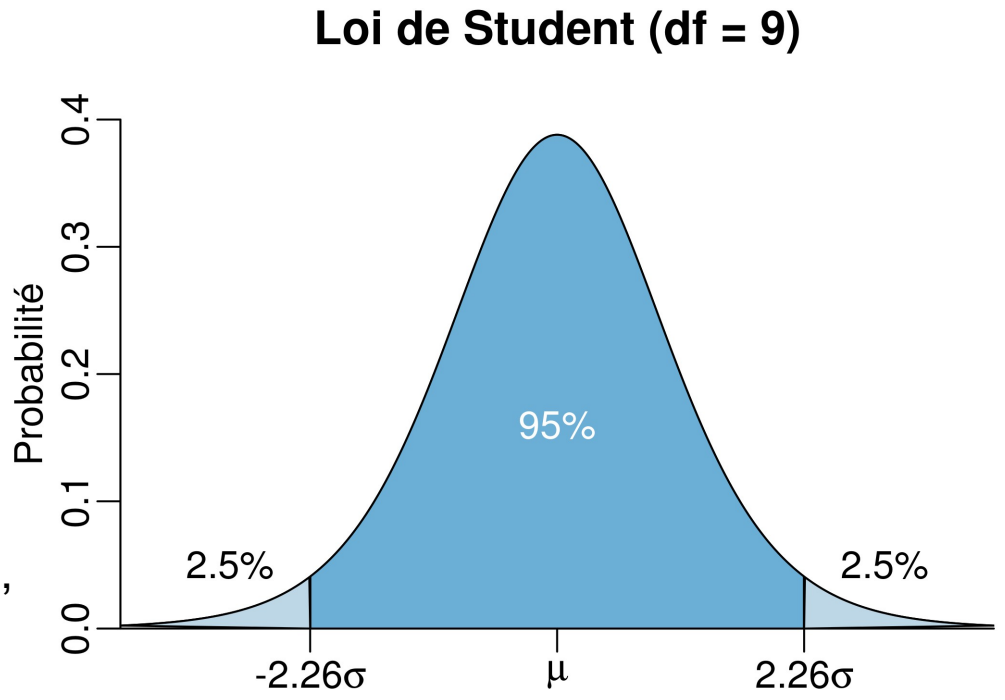
L'**intervalle de confiance** à 100(1-α) % devient donc :

$$\bar{x} - t_{\alpha/2[n-1]} se_{\bar{x}} \leqslant \mu \leqslant \bar{x} + t_{\alpha/2[n-1]} se_{\bar{x}}$$

Dans R on peut calculer les quantiles de cette loi avec la fonction `qt()` *

```
> qt(p= 0.05/2, df=9)
[1] -2.262157
> qt(p= 0.1/2, df=9)
[1] -1.833113
> qt(p= 0.01/2, df=9)
[1] -3.249836
```

Attention !
sans indiquer `lower.tail = FALSE`,
on obtient des valeurs négatives



*voir ?Distributions pour les autres distributions possibles

Inférence : précision des paramètres

En pratique :

```
> set.seed(123)
> d <- rnorm(15, mean = 100, sd = 10)
> d
 [1]  94.39524  97.69823 115.58708 100.70508 101.29288 117.15065
 [7] 104.60916  87.34939  93.13147  95.54338 112.24082 103.59814
[13] 104.00771 101.10683  94.44159

> t.test(d, conf.level= 0.99)

t = 46.5139, df = 14, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
  95.02642 108.02127
sample estimates:
mean of x
 101.5238

> n <- length(d)
> mean(d) - qt(0.01/2, df = n-1, lower.tail=FALSE) * sd(d)/sqrt(n)
[1] 95.02642
> mean(d) + qt(0.01/2, df = n-1, lower.tail=FALSE) * sd(d)/sqrt(n)
[1] 108.0213
```

quantile de la loi t de Student

Erreur standard

Inférence : précision des paramètres

On peut comparer 3 approches
(loi Student, bootstrap, simulation de données) :

```
> t.test(d)

95 percent confidence interval:
  96.84251 106.20518

> n <- length(d)
> mean(d) - qt(0.05/2, df = n-1, lower.tail=FALSE) * sd(d)/sqrt(n)
[1] 96.84251
> mean(d) + qt(0.05/2, df = n-1, lower.tail=FALSE) * sd(d)/sqrt(n)
[1] 106.2052
>
> bootmean <- replicate(10000, mean(sample(d, replace=TRUE)))
> quantile(bootmean, probs = c(0.025, 0.975))
     2.5%     97.5%
 97.51355 105.87046
>
> pop <- rnorm(100000, mean=mean(d), sd = sd(d))
> simvar <- replicate(10000,mean(sample(pop, n)))
> quantile(simvar, probs = c(0.025, 0.975))
     2.5%     97.5%
 97.34556 105.86454
```

Bootstrap

"fake data simulation"

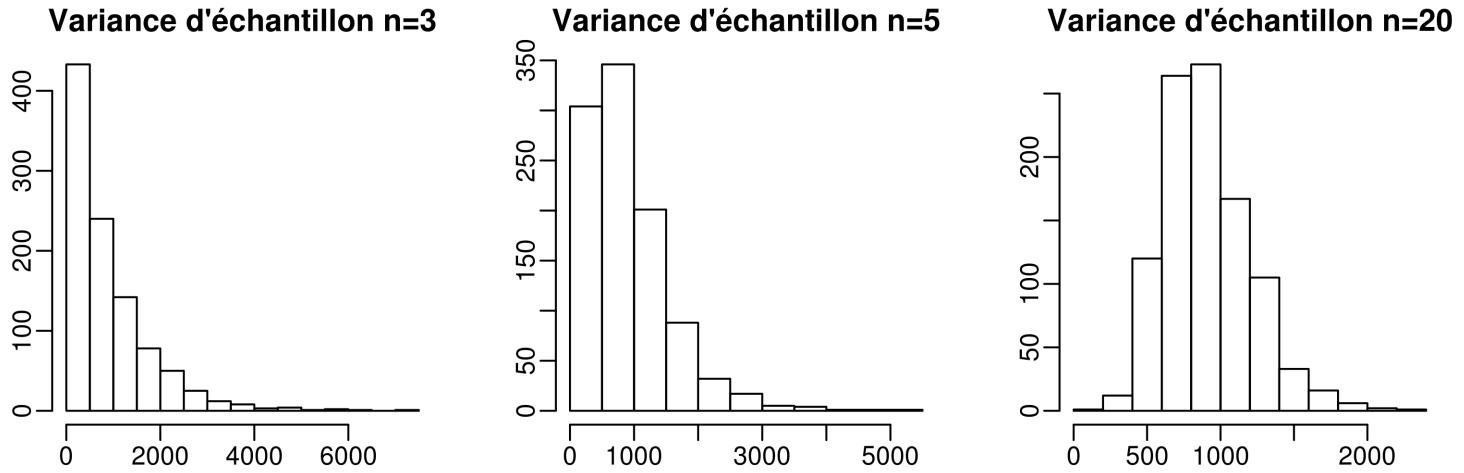
Inférence : précision des paramètres

Lois de Probabilité théoriques

Avec l'approche classique on a donc besoin pour chaque type de paramètre estimé de la théorie mathématique reliant le paramètre à une loi de probabilité.

La variance n'est par exemple pas du tout distribuée selon une loi Normale et son erreur standard ne peut pas se calculer de la même manière que la moyenne

Si on rééchantillonne 1000 fois notre population de 100000 individus avec des tailles d'échantillons différentes voici ce qu'on obtient :

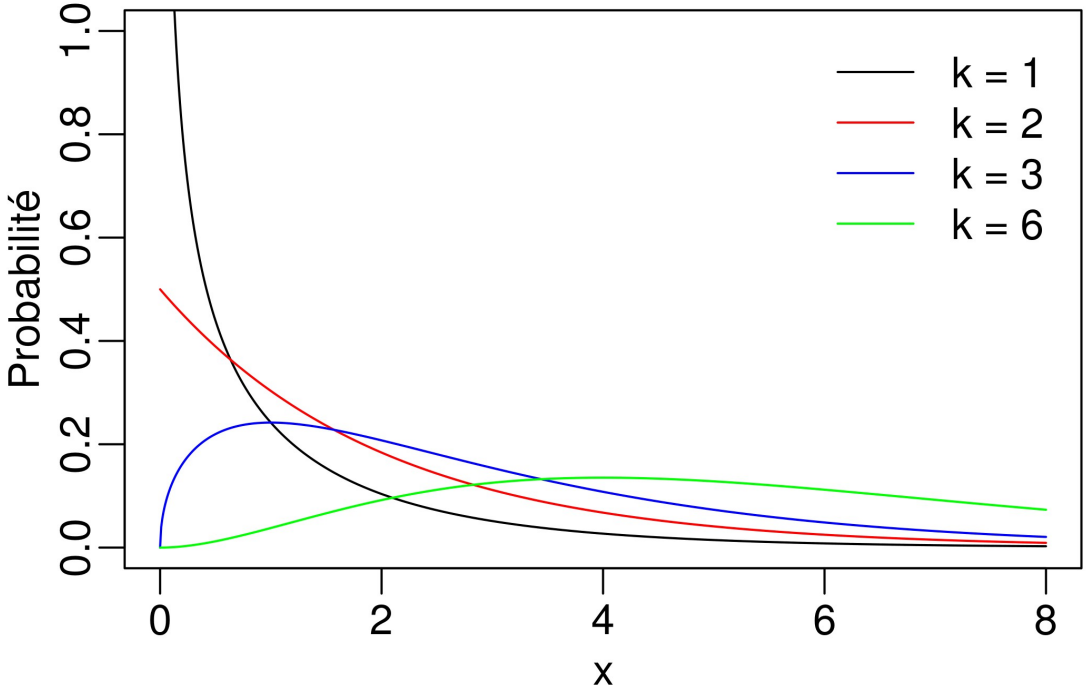


Inférence : précision des paramètres

Lois de Probabilité théoriques

La variance est liée de manière indirecte à la loi Chi carré qui est une loi asymétrique qui dépend comme la Student du nombre de degrés de liberté (k)

Loi Chi carré



Inférence : précision des paramètres

Lois de Probabilité théoriques

On peut démontrer que l'intervalle de confiance à 100(1-α) % de la variance issue d'une population normale peut se calculer comme ceci :

(Sokal & Rohlf 1995 p 155)

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2)[n-1]}} \leqslant \sigma^2 \leqslant \frac{(n-1)s^2}{\chi^2_{(1-(\alpha/2))[n-1]}}$$

Dans R on pourrait le calculer comme ceci :

```
> d <- c(9.51, 12.001, 14.058, 16, 19.1) # jeu de données exemple
> var(d)
[1] 13.51999
> n <- length(d)
>
> ((n-1)*var(d)) / qchisq(0.05/2, df = n-1, lower.tail = FALSE)
[1] 4.853142
> ((n-1)*var(d)) / qchisq(1-(0.05/2), df = n-1, lower.tail = FALSE)
[1] 111.6389
```

Important à cause de l'asymétrie

Inférence : précision des paramètres

On peut comparer 3 approches
(loi chi carré, bootstrap, simulation de données) :
Avec un échantillon de 50 unités : résultats consistants

```
> set.seed(123)
> d <- rnorm(50, mean = 100, sd = 3)
> var(d)
[1] 7.715117
> n <- length(d)
>
> ((n-1)*var(d)) / qchisq(0.05/2, df = n-1, lower.tail = FALSE)
[1] 5.383477
> ((n-1)*var(d)) / qchisq(1-(0.05/2), df = n-1, lower.tail = FALSE)
[1] 11.98041
>
> bootvar <- replicate(10000,var(sample(d, replace=TRUE)))
> quantile(bootvar, probs = c(0.025, 0.975))
    2.5%    97.5%
 5.02463 10.34479
>
> pop <- rnorm(100000, mean=mean(d), sd = sd(d))
> simvar <- replicate(10000,var(sample(pop, n)))
> quantile(simvar, probs = c(0.025, 0.975))
     2.5%     97.5%
 4.983863 11.035615
```

Inférence : précision des paramètres

Avec un échantillon de 5 unités
Qui croire ?

```
> set.seed(123)
> d <- rnorm(5, mean = 100, sd = 3)
> var(d)
[1] 5.919808
> n <- length(d)
>
> ((n-1)*var(d)) / qchisq(0.05/2, df = n-1, lower.tail = FALSE)
[1] 2.124977
> ((n-1)*var(d)) / qchisq(1-(0.05/2), df = n-1, lower.tail = FALSE)
[1] 48.88176
>
> bootvar <- replicate(10000,var(sample(d, replace=TRUE)))
> quantile(bootvar, probs = c(0.025, 0.975))
      2.5%      97.5%
 0.2038835 10.4320152
>
> pop <- rnorm(100000, mean=mean(d), sd = sd(d))
> simvar <- replicate(10000,var(sample(pop, n)))
> quantile(simvar, probs = c(0.025, 0.975))
      2.5%      97.5%
 0.7132527 16.1905818
```

Attention : beaucoup de distributions théoriques de statistiques sont valides uniquement asymptotiquement i.e. que pour des grands échantillons

Bootsrap par percentiles loin d'être infaillible !
En général peu adapté à des petits échantillons
Il existe d'autres approches plus robustes

Inférence : précision des paramètres

Variance - Ecart type - Erreur Standard - Intervalle de Confiance

Souvent confondus !

La variance (σ^2) et l'écart type (σ) sont des caractéristiques de la population que l'on estime au moyen de la variance (s^2) et de l'écart type (s) d'un échantillon.
On peut mesurer/calculer directement la variance/écart type des données au sein d'un échantillon

L'erreur standard et l'intervalle de confiance sont une caractéristique des paramètres d'un échantillon particulier (ie la précision avec laquelle on les estime), pas de la population. Ils représentent la variabilité de ce paramètre entre les échantillons si on avait pu récolter plusieurs échantillons similaire
On ne peut pas les calculer directement (il faut les inférer).

Inférence : précision des paramètres

Variance - Ecart type - Erreur Standard - Intervalle de Confiance

L'écart type et la variance ne dépendent pas de la taille de l'échantillon mais uniquement de la variabilité (écart type et variance) de la population.
L'erreur standard et l'intervalle de confiance dépendent à la fois de la variabilité de la population et de la taille de l'échantillon.

L'erreur standard de la moyenne est la plus connue et la plus facile à calculer mais tous les paramètres peuvent avoir une erreur standard

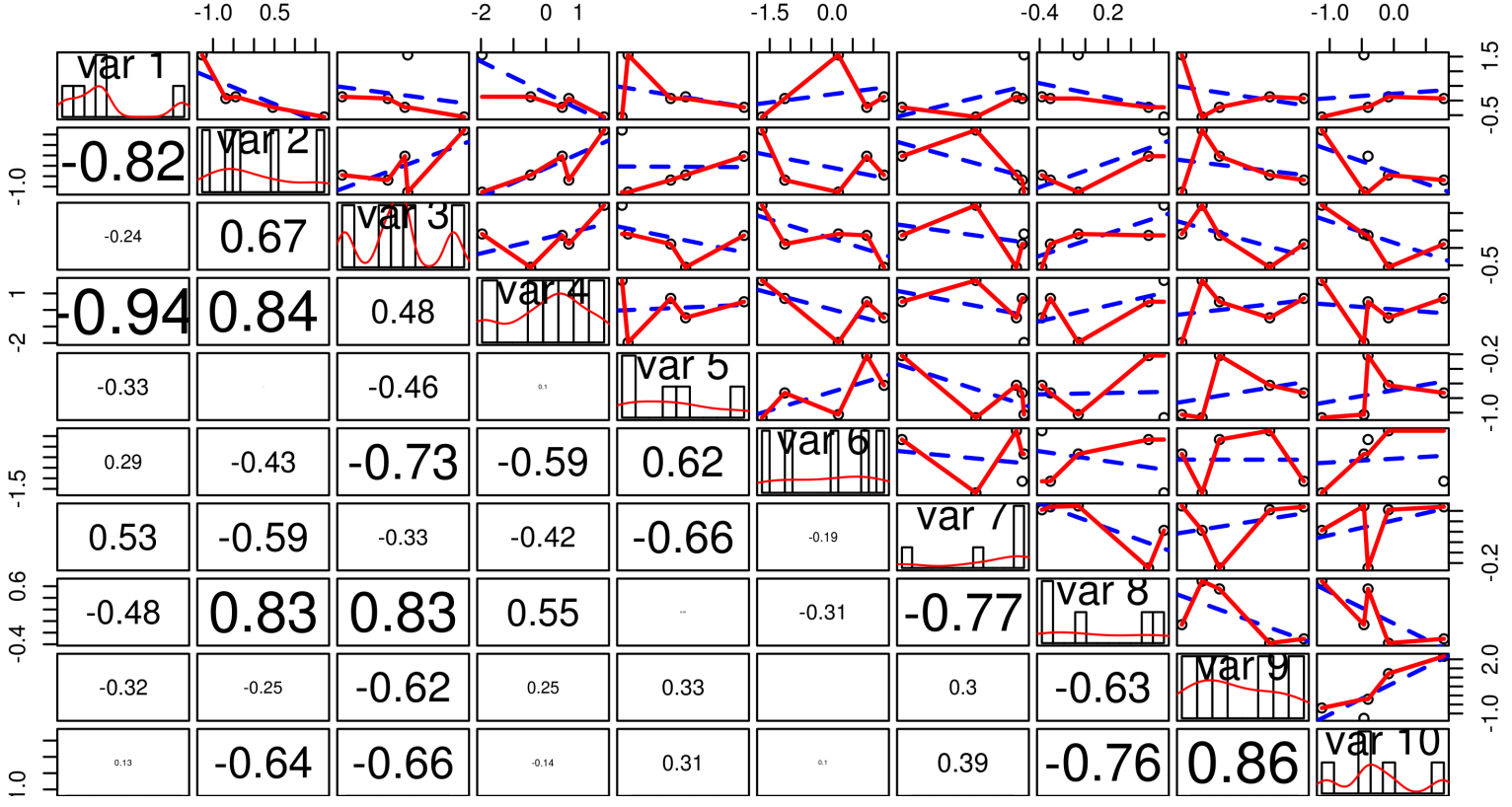
Sur les graphiques on les représente souvent par des barres d'erreur. Il faut bien préciser ce qu'on représente !
Représenter la variance de cette manière n'a en général pas beaucoup de sens

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le problème

On génère 10 variables complètement aléatoires (donc indépendantes)
Chaque variable contient 5 éléments
On calcule la corrélation entre chaque paire de variable

Certaines corrélations peuvent être extrêmement élevées
mais c'est uniquement dû au hasard



Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le raisonnement

Si je mesure deux variables réelles et que j'observe une corrélation de 0.8, comment savoir si il y a réellement une relation entre ces variables ou si ce résultat est dû uniquement au hasard ?

C'est pour répondre à cette question qu'on estime une **"p valeur"** qui est la probabilité d'obtenir un tel résultat (ou une corrélation encore plus forte) dans l'hypothèse (la fameuse **hypothèse nulle**) où il n'y aurait en fait aucune relation entre les deux variables

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le raisonnement

Si cette probabilité est trop élevée (en général p>0.05) on considère qu'on aurait très bien pu obtenir le même résultat uniquement par hasard sans qu'il y ait de relation réelle entre les variables.

Il y a peut-être une relation entre ces variables mais **nos données sont en tous cas insuffisantes** pour l'affirmer.

Classiquement on dit qu'on "ne peut pas rejeter l'hypothèse nulle (H0)" qui est dans ce cas que la corrélation est 0.

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le raisonnement

Si cette probabilité est faible (en général $p<0.05$) on considère qu'il est très peu probable d'avoir obtenu de tels résultats uniquement par hasard et donc que **nos données sont suffisantes pour supporter le résultat observé**.

Classiquement on dit qu'on "rejette l'hypothèse nulle" et par conséquent on accepte l'hypothèse alternative (H1) qui dans ce cas est que la corrélation est différente de 0

NB : très peu probable ne veut pas dire impossible, on se trompe donc parfois...

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : le raisonnement

NB : Le seuil de 0.05 est une convention purement arbitraire à ne certainement jamais prendre au pied de la lettre !

NB : le raisonnement est souvent considéré comme tordu/compliqué car :
on estime la probabilité d'obtenir nos données sachant que H0 est vraie alors que ce qui nous intéresse en réalité est de calculer la probabilité que H1 soit vraie sachant nos données...

Inférence : tests d'hypothèse nulle

Tests d'Hypothèse nulle : en pratique

2 approches fréquentes :

- tests par permutation (ou randomisation)
(une autre méthode de rééchantillonnage par ordinateur)

- tests classiques basés sur des lois de probabilité

Inférence : tests d'hypothèse nulle

Tests par permutation

1) On calcule le paramètre d'intérêt sur notre échantillon (valeur observée)

2) On mélange les données aléatoirement pour éliminer les différences/relations entre les données et on calcule le paramètre d'intérêt sur ces données mélangées

3) On recommence l'opération 2 un grand nombre de fois

4) On regarde combien de fois on obtient une valeur plus grande ou égale à la valeur observée
(selon deux modalités possibles)

Inférence : tests d'hypothèse nulle

Tests unilatéral vs test bilatéral

On peut en fait poser la question de deux manières :

a) Test unilatéral = "One sided test" :
Est-ce que le paramètre est plus grand que 0 ? *

b) Test bilatéral = "Two sided tests" :
Est-ce que le paramètre est différent de 0 ?

Dans ce dernier cas on compare les valeurs absolues du paramètre et des paramètres obtenus par permutation

On utilise en général un test bilatéral sauf si on a de bonnes raisons *a priori* de tester le paramètre uniquement dans une direction.
Les "two sided p-values" sont toujours plus grandes

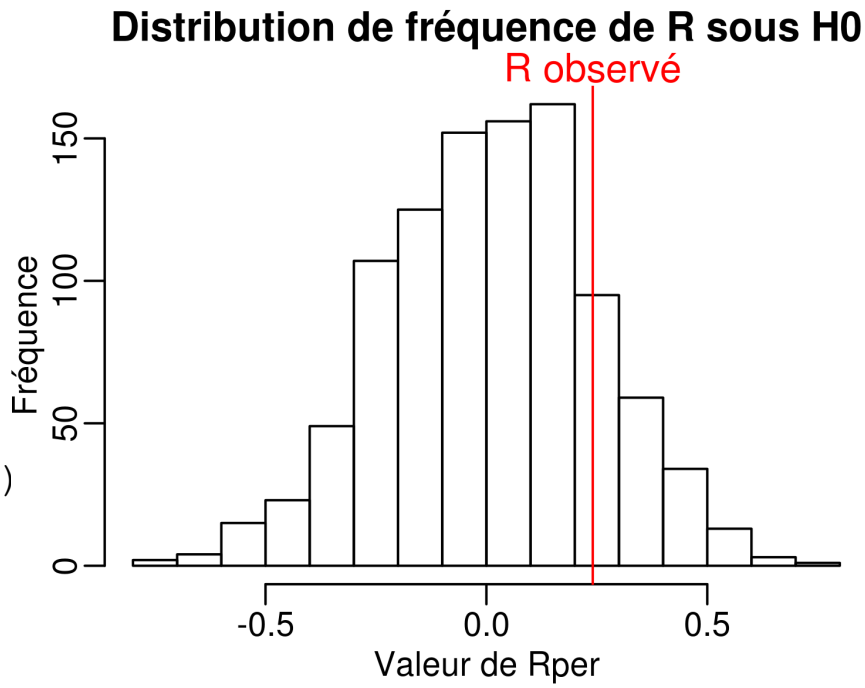
* ou bien : "Est-ce que R est plus petit que 0 ?"

Inférence : tests d'hypothèse nulle

Tests par permutation

Pour mélanger les données, on utilise la fonction `sample()` sans spécifier l'argument `size` (par défaut il prend toutes les valeurs) et avec `replace = FALSE`

```
> # on crée un jeu de données
> set.seed(123)
> x <- rnorm(20)
> set.seed(12)
> y <- 1.5*x + rnorm(20, 0, 4)
>
> (Robs <- cor(x,y))
[1] 0.2407653
> Rper <- replicate(999,
    cor(sample(x, size=20, replace=FALSE), y))
> # évite des p valeurs = 0
> Rper <- c(Rper, Robs)

> # "one sided test"
> sum(Rper >= Robs) /1000
[1] 0.159
# "two sided test"
> sum(abs(Rper) >= abs(Robs)) /1000
[1] 0.308
```



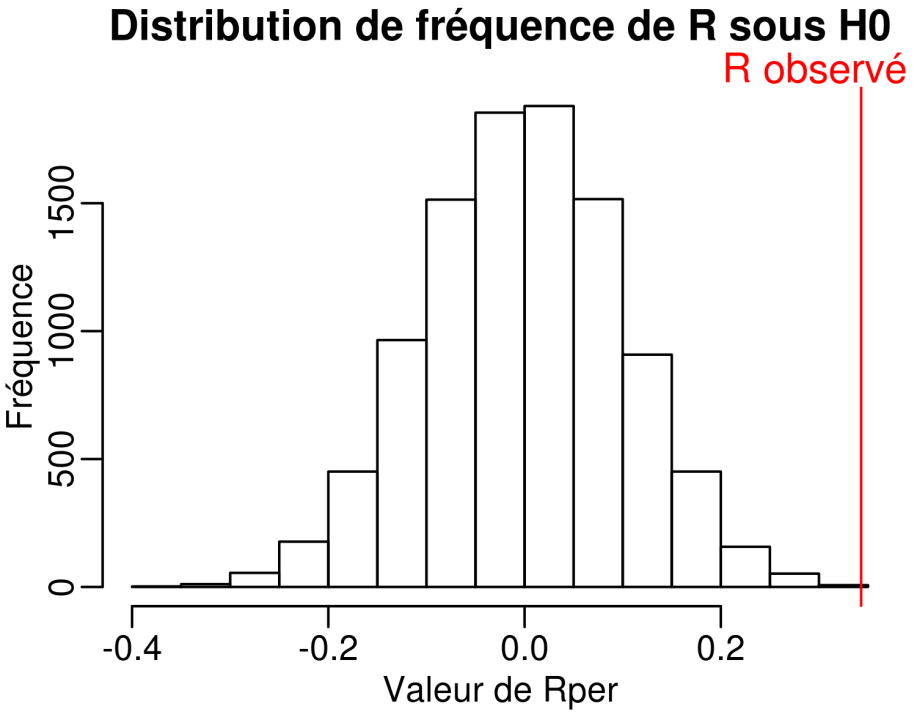
Conclusion : on a 30.8% de chance d'obtenir un tel résultat par hasard
--> les données sont insuffisantes pour affirmer qu'il y a bien une corrélation

Inférence : tests d'hypothèse nulle

Tests par permutation

On reprend exactement les mêmes données mais avec un échantillon plus grand (100 ou lieu de 20)

```
> set.seed(123)
> x <- rnorm(100)
> set.seed(12)
> y <- 1.5*x + rnorm(100, 0, 4)
>
> (Robs <- cor(x,y))
[1] 0.3429773
> Rper <- replicate(9999,
    cor(sample(x, replace=FALSE), y))
> Rper <- c(Rper, Robs)
>
> # évite notation scientifique
> options(scipen=10)
> sum(Rper >= Robs) /10000
[1] 0.0001
> sum(abs(Rper) >= abs(Robs)) /10000
[1] 0.0005
```



Conclusion : on a 0.05% de chance d'obtenir un tel résultat par hasard --> les données sont suffisantes pour affirmer qu'il y a bien une corrélation

Inférence : tests d'hypothèse nulle

Tests par permutation

On ne doit faire aucune hypothèse sur la distribution de la population. La distribution de fréquence sous H0 est simplement simulée à partir de l'échantillon.

A priori on peut appliquer les tests par permutation à n'importe quelle statistique mais idéalement elle doit être "pivotale" c'est à dire une statistique dont la distribution garde la même forme sous l'hypothèse nulle quelles que soient les valeurs testées

pex pour comparer les moyennes de deux groupes il vaut mieux utiliser la statistique pivotale utilisée dans le test de student (avec n1 + n2 -2 degrés de liberté):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

que la statistique plus simple et plus intuitive :

$$\bar{x}_1 - \bar{x}_2$$

Inférence : tests d'hypothèse nulle

Lois de Probabilité

On peut montrer que :
si les deux variables sont distribuées selon une loi normale et que leur corrélation est 0 (Hypothèse nulle), alors la statistique suivante est distribuée selon une loi t de Student avec n-2 degrés de liberté (le dénominateur représente l'erreur standard de R):

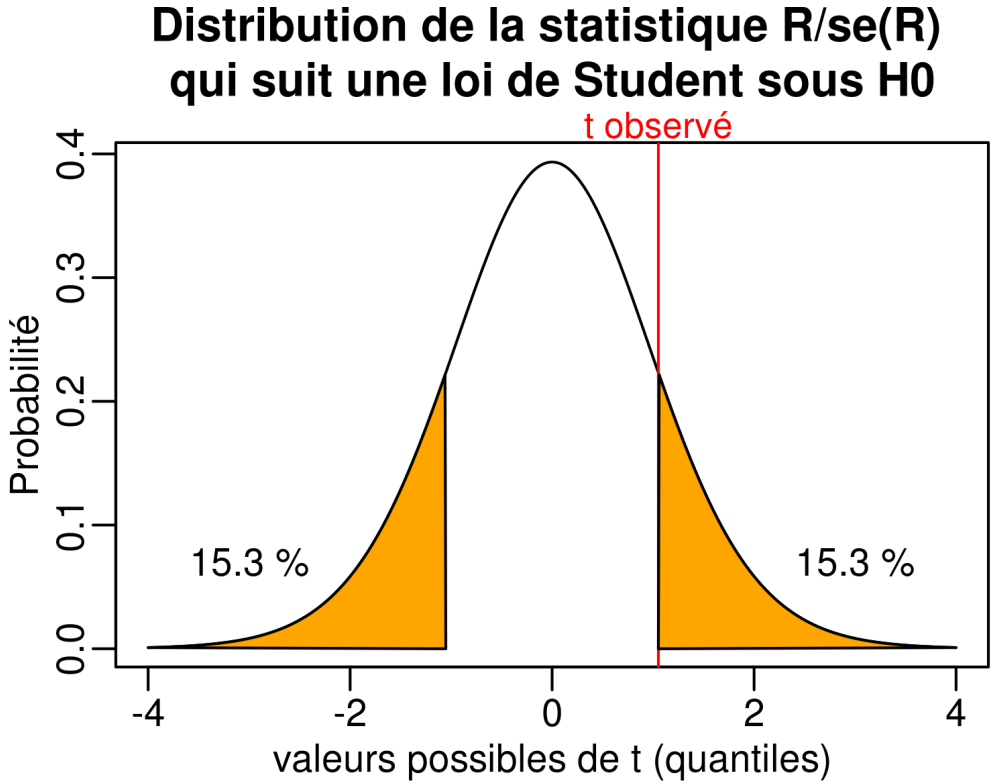
$$t = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}}$$

Inférence : tests d'hypothèse nulle

Lois de Probabilité

On va donc calculer cette statistique t et récupérer la probabilité d'obtenir une telle valeur ou une valeur plus extrême au moyen de la loi de Student

Cela revient à calculer l'aire sous la courbe au delà de la valeur t obtenue.



Inférence : tests d'hypothèse nulle

Lois de Probabilité

Dans R, on utilise la fonction `pt()` (et de manière équivalente pour d'autres distributions : `pnorm`, `pchisq`, ...) pour obtenir cette surface

```
> cor.test(x,y) # two sided test

t = 1.0524, df = 18, p-value = 0.3065
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2258146  0.6174958

> cor.test(x,y, alternative = "greater") # one sided test
t = 1.0524, df = 18, p-value = 0.1533
alternative hypothesis: true correlation is greater than 0

> (Robs <- cor(x,y))
[1] 0.2407653
> n <- length(x)
> (Rt <- abs(Robs / sqrt((1-Robs^2)/(n-2))))
[1] 1.05244
> (Rdf <- n-2)
[1] 18
> (p <- 2*pt(Rt, Rdf, lower.tail = FALSE)) # two sided
[1] 0.3065219
> (p <- pt(Rt, Rdf, lower.tail = FALSE)) # one sided
[1] 0.1532
```

$$t = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}}$$

Uniquement avec distrib. symétriques

Inférence : tests d'hypothèse nulle

Lois de Probabilité

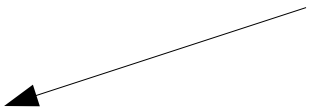
Note : l'intervalle de confiance donné par le logiciel ne se calcule pas au moyen d'une distribution de Student.

On applique d'abord une "z-transformation" au R et on estime son erreur standard. La statistique obtenue suit asymptotiquement (n>50) une distribution normale. Une fois qu'on a les bornes de l'intervalle on les rétro-transforme au moyen de la fonction tangente hyperbolique

```
> cor.test(x,y)
95 percent confidence interval:
 -0.2258146  0.6174958

> z <- 0.5 * log((1+Robs)/(1-Robs))
> sigmaz <- 1/sqrt(n-3)
> IC1 <- tanh(z + qnorm(0.025) * sigmaz)
> IC2 <- tanh(z + qnorm(0.975) * sigmaz)
> c(IC1, IC2)
[1] -0.2258146  0.61749588
```

Qu'est-ce qu'on dit à Sir Fisher pour avoir fait toutes les maths ? Merci !

Ou alors on dit merci à Mr Efron d'avoir développé le bootstrap !!



```
> mat <- cbind(x,y)
> boot <- replicate(10000, cor(mat[sample(1:n, replace = TRUE),])[1,2])
> quantile(boot, probs = c(0.025, 0.975))
      2.5%      97.5%
-0.01775811  0.51435522
```

Tests d'hypothèse nulle : attention à l'interprétation !

Exemple 1 : On teste l'effet d'un nouveau traitement assez cher sur la production de lait chez les vaches

On compare avec un test de Student la production moyenne entre les vaches d'un groupe témoin et d'un groupe traité

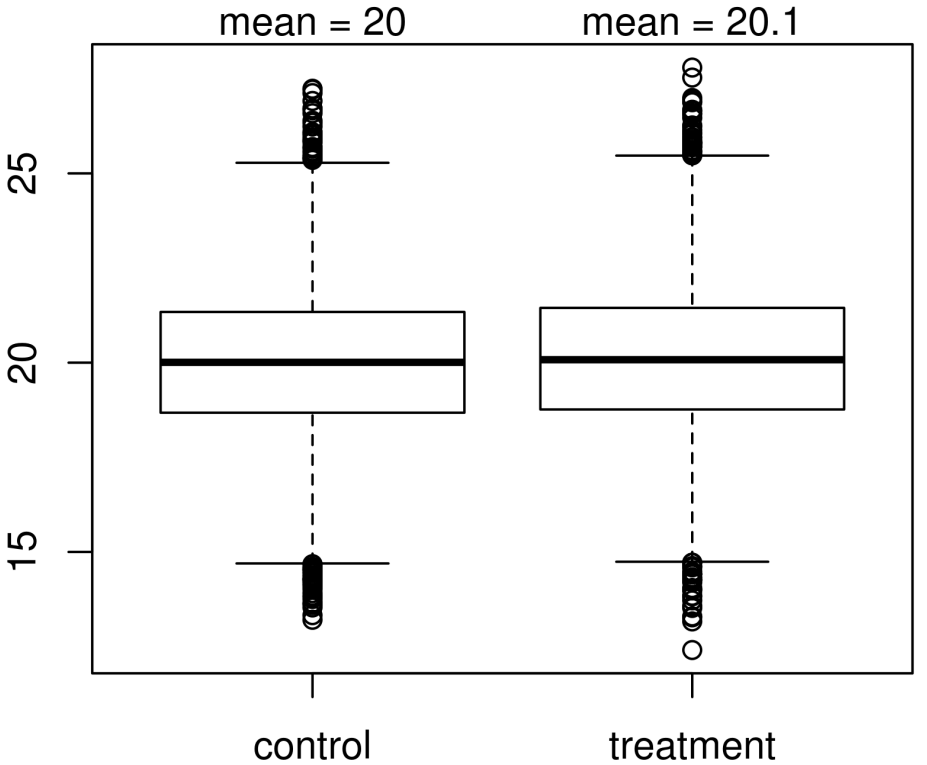
```
> t.test(control, treatment, var.equal = TRUE)

    Welch Two Sample t-test

data:  control and treatment
t = -2.9558, df = 19998, p-value = 0.003123
```

Est-ce que vous investiriez dans ce traitement ?

Tests d'hypothèse nulle : attention à l'interprétation !

Significativité statistique ≠ significativité biologique

```
> t.test(control, treatment, var.equal = TRUE)

data:  control and treatment
t = -2.9558, df = 19998, p-value = 0.003123
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.13808170 -0.02796792
sample estimates:
mean of x mean of y
 20.01223  20.09526
```



Tests d'hypothèse nulle : attention à l'interprétation !

Exemple 2 : autre expérience avec un autre traitement

```
> t.test(control, treatment, var.equal = TRUE)

    Welch Two Sample t-test

data:  control and treatment
t = -1.3331, df = 28, p-value = 0.1933
```

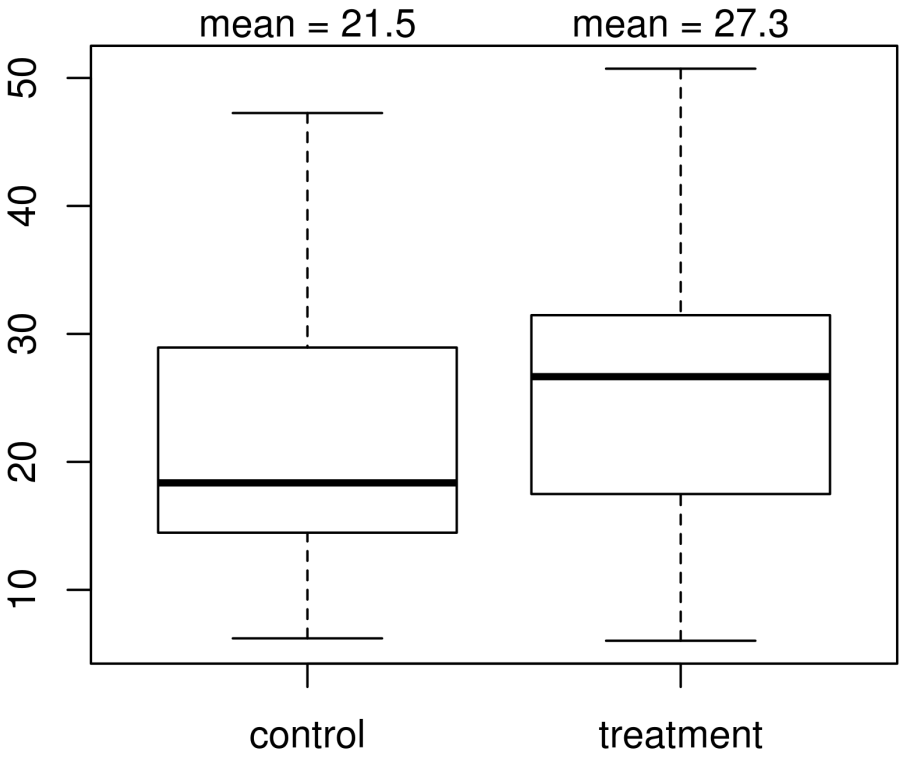
Est-ce que vous investiriez dans ce traitement ?

Tests d'hypothèse nulle : attention à l'interprétation !

Non significatif ne veut pas dire qu'il n'y a pas d'effet !

```
> n <- 15
> set.seed(12345)
> control <- abs(rnorm(n, 20, 15))
> set.seed(123)
> treatment <- abs(rnorm(n, 25,15))

> t.test(control, treatment, var.equal = TRUE)

data:  control and treatment
t = -1.3331, df = 28, p-value = 0.1933
alternative hypothesis: true difference i
95 percent confidence interval:
 -14.723968   3.114799
sample estimates:
mean of x mean of y
 21.48118  27.28577
```

Il y a peut-être un effet mais les données sont insuffisantes
D'où viens cette variabilité ?
Qu'est-ce que çà donne avec un échantillonnage plus grand ?



Tests d'hypothèse nulle : attention à l'interprétation !

la p-valeur dépend de :

- la taille de l'échantillon
- la variabilité de la population
- la taille de l'effet

-->

"*Si on augmente suffisamment la taille de l'échantillon, les p-valeurs finiront toujours par devenir significatives, à moins que l'on teste des hypothèses stupides, ce qui est rarement le cas*"

(d'après Burnham & Anderson 2000)

p valeur : outil d'aide à la décision à utiliser à bon escient !
Certains préfèrent les intervalles de confiances qui sont moins sujets à une mauvaise interprétation.

Intervalles de confiance vs p valeurs

Est-ce qu'un paramètre est significativement différent de 0 au seuil alpha = 5 % ?

Oui si son intervalle de confiance à 95 % ne comprend pas le 0

Le raisonnement fonctionne pour n'importe quelle valeur fixe autre que 0.

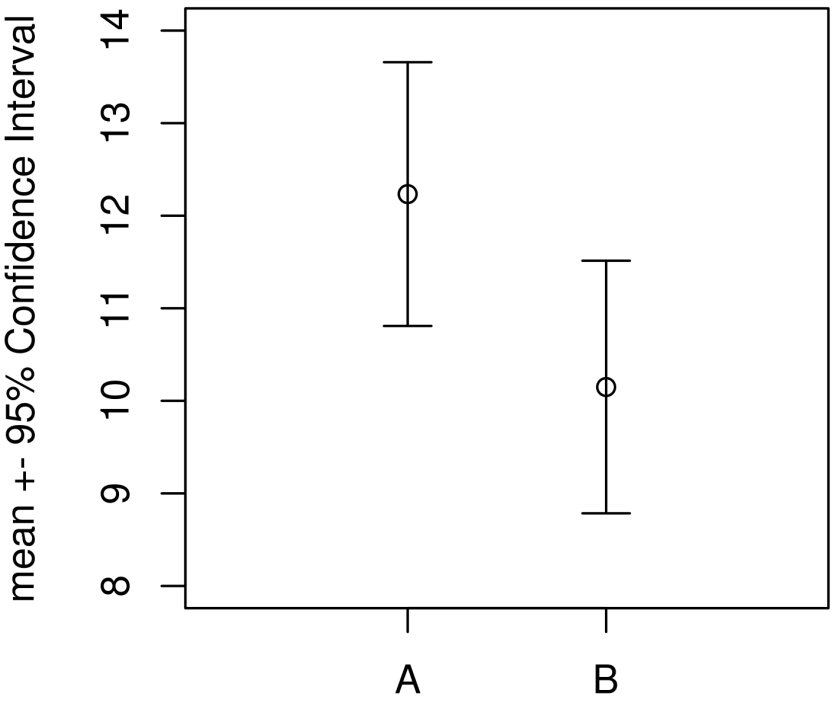
Est-ce que 2 paramètres (par exemple 2 moyennes) sont significativement différents au seuil alpha = 5 % ?

Si leurs intervalles de confiance à 95 % ne se recouvrent pas : oui.
Attention : si les intervalles se recouvrent (jusqu'à ~25% pour une moyenne), çà ne veut pas dire automatiquement que les paramètres ne sont pas significativement différents !!!

Intervalles de confiance vs p valeurs

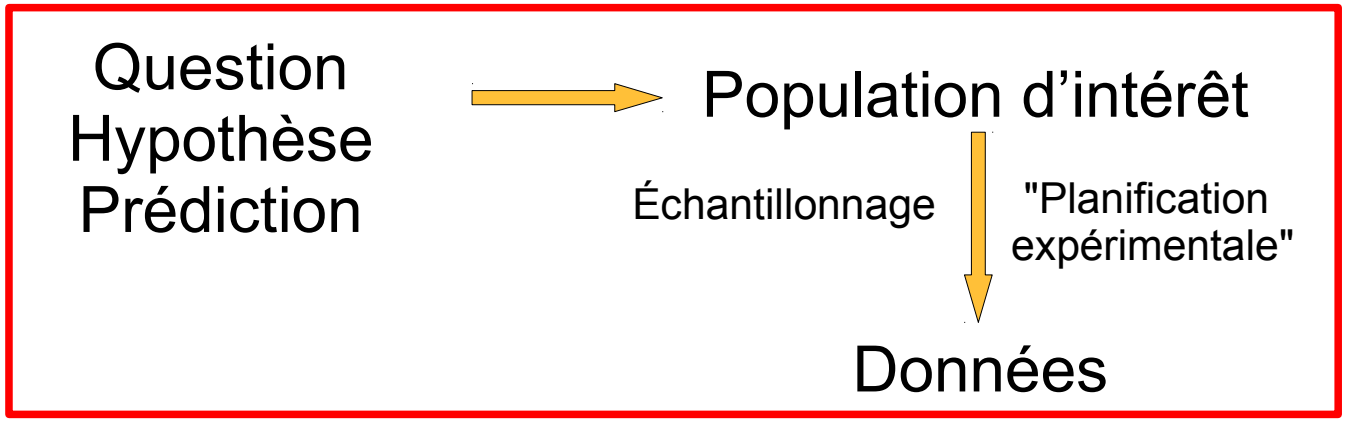
Deux intervalles de confiances peuvent se chevaucher alors que la différence de leur paramètres est significative

```
> set.seed(1234)
> A <- rnorm(10, 13, sd=2)
> set.seed(123)
> B <- rnorm(10, 10, sd=2)
> t.test(A,B)

t = 2.3902, df = 17.967, p-value = 0.02801

> t.test(A)$conf.int
[1] 10.80900 13.65837
attr(,"conf.level")
[1] 0.95
> t.test(B)$conf.int
[1]  8.784659 11.513843
attr(,"conf.level")
[1] 0.95
```



Partie 2 :
Planification expérimentale

Les Statistiques : Pourquoi ?



Généralisation de l'échantillon à la population

Estimation

Résumé numérique des données

position, variabilité, association, droite de régression, ...

Inférence

Précision des estimations

Probabilité d'obtenir un tel résultat par hasard

Graphiques

résumé graphique :

des données

des paramètres estimés

des inférences

Etudes observatives vs expérimentales

Types d'études :

Expérimentale ("controled experiments") :
L'expérimentateur contrôle tous les paramètres et en fait varier certains
--> met en évidence des liens de cause à effet
Mais : les conditions sont-elles réalistes ?

Observative ("natural experiments") :
On échantillonne dans une variété de conditions existantes sans pouvoir les contrôler
--> met en évidence des corrélations mais difficilement des lien de cause à effet
Mais : conditions souvent plus réalistes
Souvent la seule approche possible

Semi-expérimentale
On manipule le facteur d'intérêt (traitement) mais on ne contrôle pas les conditions
exemple typique : essais agronomiques en champs

Les règles d'or de la planification expérimentale concernent aussi les études observatives !

Planification expérimentale

Classiquement planification expérimentale =
1 design expérimental optimisant la récolte de données
+ une analyse pour chaque cas

Ici : on se contentera de voir les grands principes

Les GLM (+ extensions) sont ensuite suffisamment flexibles pour s'adapter à chaque cas d'analyse

Planification expérimentale

La planification expérimentale peut servir à :

1) **éviter/limiter certains problèmes** :

- biais dans la récolte de donnés
(pex : échantillonnage non aléatoire)

- interprétation incorrecte des résultats
(pex : confusion de facteurs, absence de contrôles)

- invalidité des inférences
(pex : non indépendance des échantillons)

2) **Augmenter la puissance d'un test** (diminuer les p-valeurs*)
et/ou la précision des estimations, idéalement à coût égal

NB : définition fausse mais çà donne une idée de l'objectif...

Planification expérimentale

Il est évidemment primordial de penser à tout çà avant de commencer l'étude.

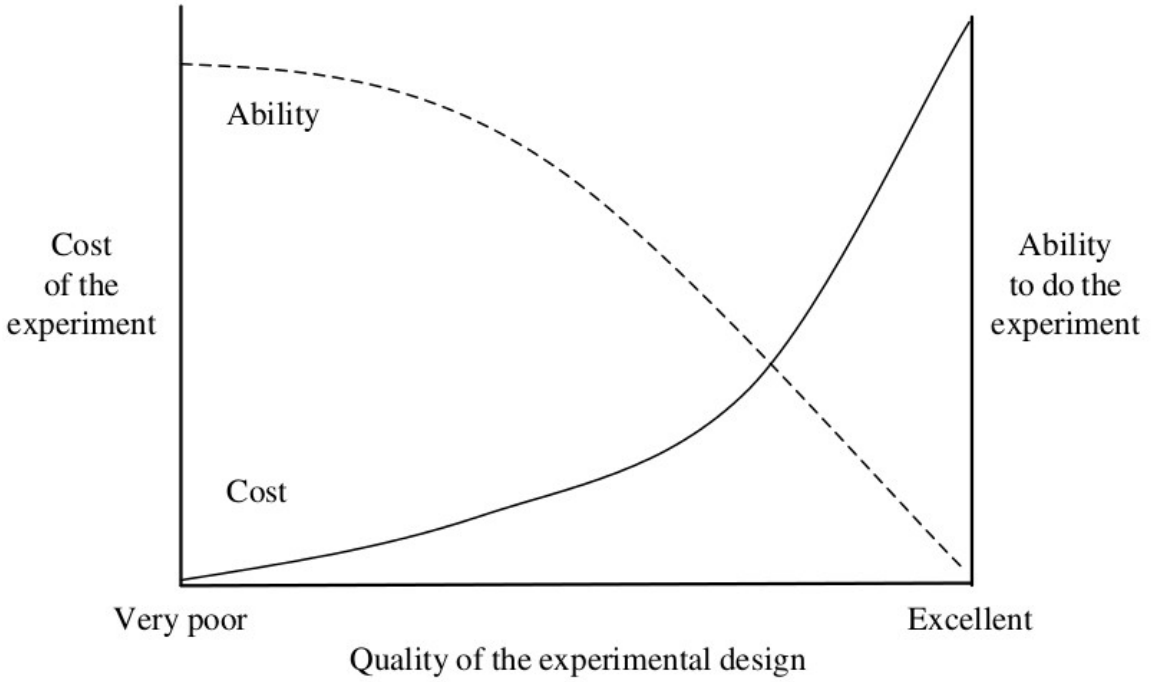
Encore plus pour les études longitudinales où on aura pas la possibilité d'adapter un protocole mal conçu au départ

"To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination: he may be able to say what the experiment died of." R. Fisher

Planification expérimentale

Compromis entre :

design expérimental optimal
et
contraintes pratiques - moyens disponibles



Planification expérimentale

Mais bien avoir en tête ces "règles d'or" permet :

1) d'ajuster quand même son design expérimental au mieux.
Souvent on aurait pu faire mieux avec les mêmes moyens

2) si il y a vraiment certaines règles qu'on doit laisser tomber, il faut le garder en tête pour l'interprétation des résultats
(présence de biais, confusion d'effets, impossibilité de généraliser ses résultats à d'autres cas,...)

3) dans certains cas il faut se rendre à l'évidence avant de commencer l'étude qu'on ne pourra pas répondre à la question

...

Planification expérimentale

Conflit entre :

L'approche du scientifique :
idéalement expérimentale, qui veut comprendre et trancher, en émettant des hypothèses et en les mettant à l'épreuve des faits
Mais approche lente et à coût élevé

L'approche du praticien / naturaliste, plus empirique:
qui accumule des informations descriptives, des observations, des anecdotes, etc... pour se forger une intime conviction ou une expertise sur un sujet.

La récolte d'informations a tendance a être moins "canonique", elle peut-être du coup plus rapide/abondante mais on peut encore moins souvent trancher une question qu'avec une approche scientifique plus classique.
"*The plural of anecdote is no data*"
Attention à la tentation du "bon sens"

Les deux approches peuvent clairement se nourrir l'une de l'autre, en particulier le praticien a l'intuition et la connaissance qui lui permet de poser les bonnes questions à soumettre à la méthode scientifique et d'interpréter finement les résultats

Planification expérimentale

On ne peut pas se contenter du "bon sens" si on aspire a une approche scientifique,

Exemple 1: la rotondité de la terre

Exemple 2 : les jeux vidéos/images violents comme catharsis

En théorie la tradition de la citation est un "garde fou" qui permet de limiter (très imparfaitement...) le problème

"The plural of anecdote is no data"

Exemple : l'influence de la lune sur les accouchements

Confusion de facteurs

Exemple : Agriculture biodynamique

Planification expérimentale

Une seule étude, un seul article ne permet presque jamais de "prouver scientifiquement" un fait, en particulier dans les sciences de la vie et en sciences humaines.

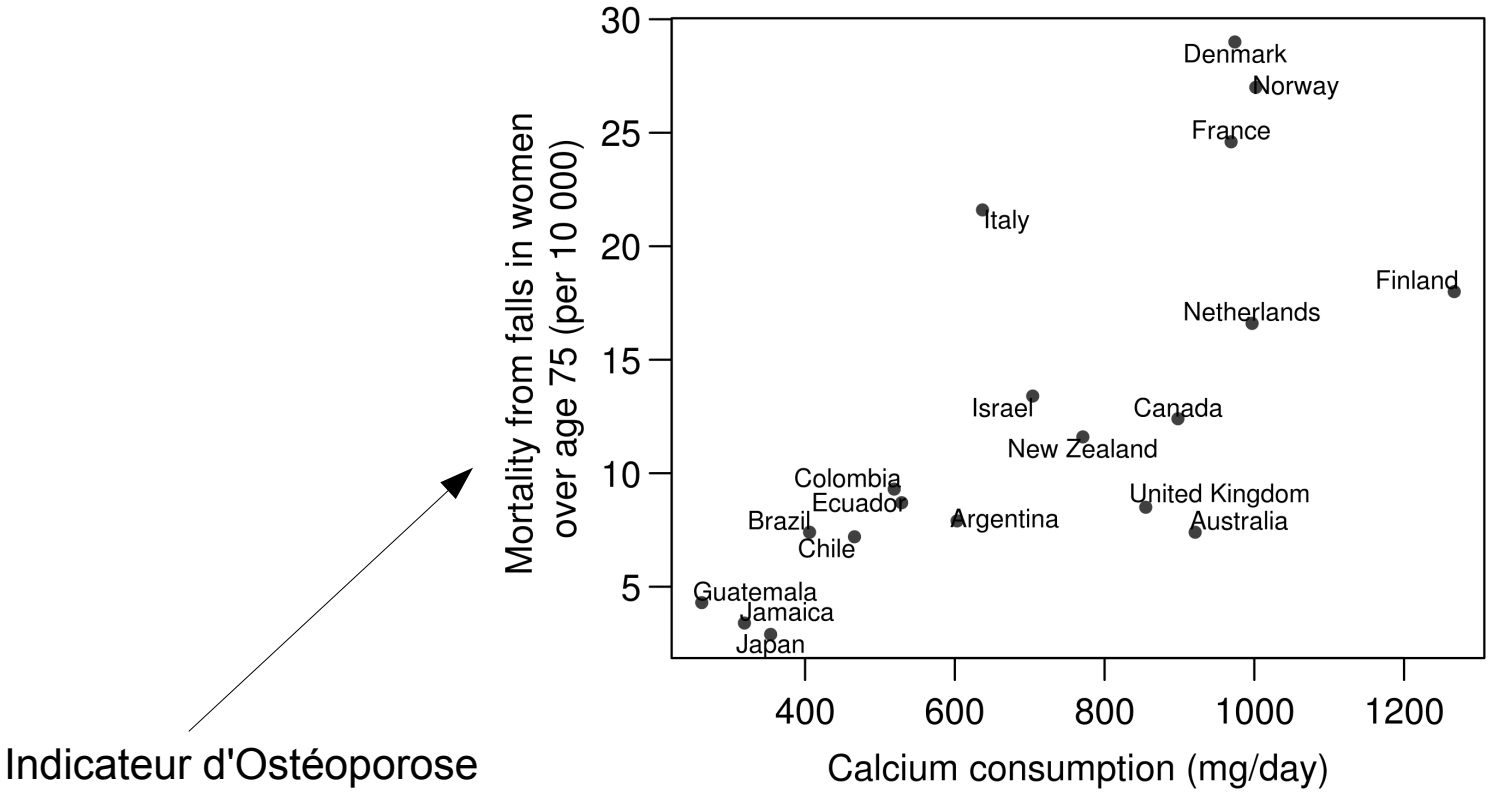
Ce n'est que l'accumulation considérables d'études observatives et/ou expérimentales bien menées (et ce n'est pas toujours le cas !!!) qui peuvent permettre à une communauté scientifique d'arriver à un consensus voire une certitude.
La répétabilité des résultats est un élément essentiel !

Planification expérimentale

Paradoxe du calcium

Exemple de l'intérêt des études observatives pour se poser les bonnes questions.

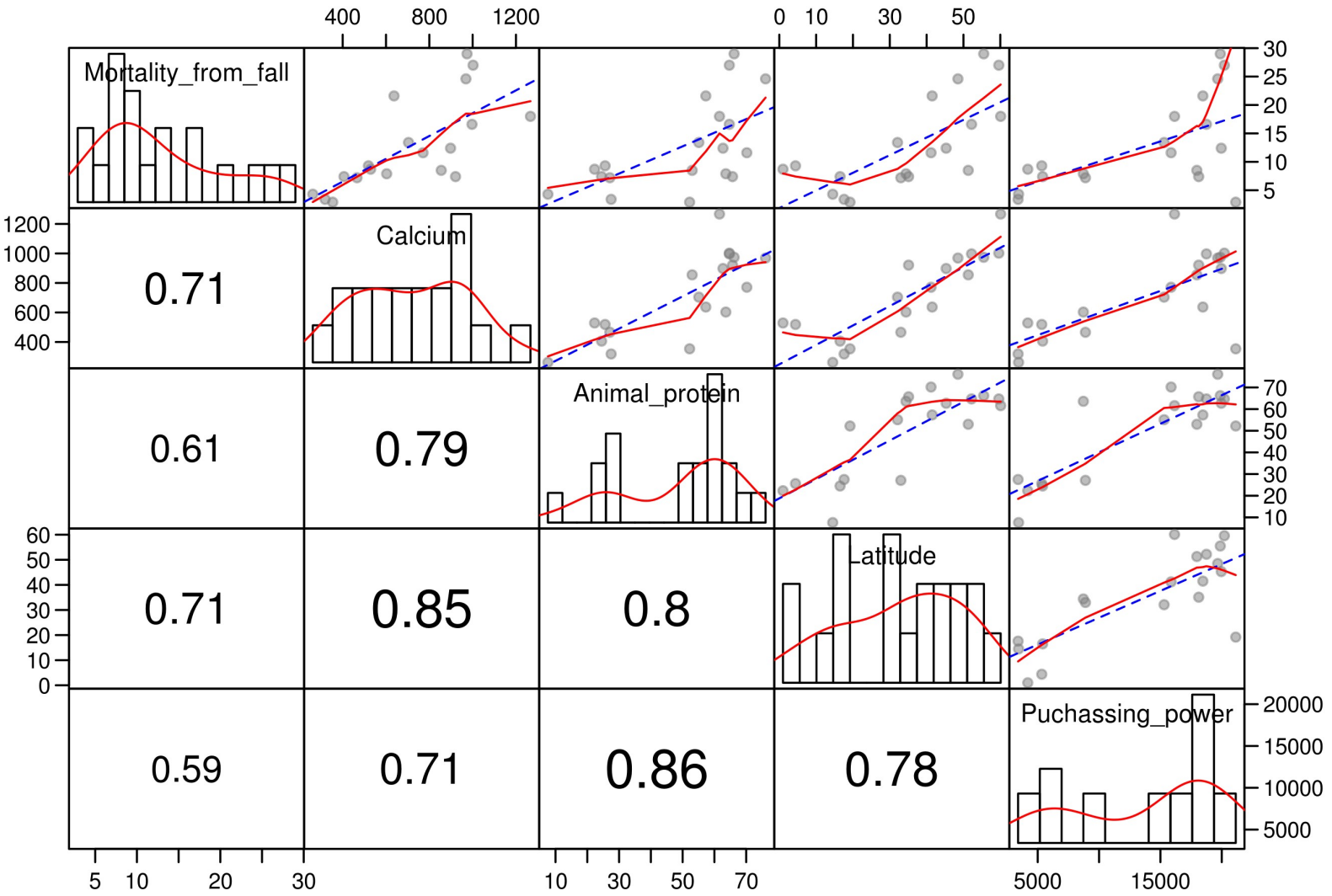
Le "bon sens" voudrait que dans les pays où on consomme le plus de calcium on ait moins de problèmes d'ostéoporose...



Données FAO : http://www.fao.org/docrep/w7336t/w7336t04.htm

Planification expérimentale

Paradoxe du calcium



Données FAO : http://www.fao.org/docrep/w7336t/w7336t04.htm

Planification expérimentale

Paradoxe du calcium

La calcémie dépend de plusieurs facteurs :

- la quantité de calcium dans la nourriture

- la capacité à absorber ce calcium
qui diminue quand la vitamine D diminue et donc quand la latitude augmente

- la perte de calcium via l'urine
qui augmente par exemple avec la consommation de protéines animales et de sel

- etc...

Ce n'est qu'en combinant des études observatives et expérimentales qu'on a pu comprendre les interactions complexes entre ces facteurs corrélés entre eux...

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Etude observative :

colonie d'abeilles prévenants des symptômes de déclin sans cause connue
dosage des pesticides et des virus dans la ruche
caractérisation du paysage agricole autour de la ruche

Résultats :

pas de lien observé avec les virus ni avec les insecticides
une relation inattendue déclin ~ fongicides
pas de lien observé avec la présence de cultures suspectes (ea colza)
une relation positive avec la surface de cultures et négative avec la surface de prairies

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Pas de relation observée ne veut pas dire qu'il n'y a pas de lien !
La charge de la preuve est du côté du chercheur...

On a peut être simplement pas assez de répétitions dans cette étude (pas assez de puissance statistique).
Démontrer une absence d'effet est presque impossible...

Certains critiquent d'ailleurs l'extrême prudence de la communauté scientifique ea le fameux seuil alpha = 0.05

voir pex : Conway & Oreskes (2014)

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Le fait qu'on ait trouvé un lien significatif déclin~fongicides (p = 0.008) n'est pas non plus une preuve définitive

La probabilité de trouver un tel résultat par chance est faible mais pas nulle
Le résultat est peut-être particulier à la population échantillonnée (nord de la Wallonie) et pas extrapolable à d'autres contextes
On ne peut jamais exclure des erreurs, biais, etc... même si l'étude a été faite de manière compétente et honnête

--> ce n'est qu'en trouvant des résultats similaires dans d'autres études, d'autres contextes, que ces observations pourront être confirmées...

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Pourquoi ce lien inattendu déclin ~ fongicides ?
Étude observative : impossible t'établir un lien de causalité

Hypothèses :

1) Lien de causalité direct ou indirect (par quel mécanisme?)
2) Fongicides comme marqueurs d'insecticides non détectés
3) Fongicides marqueurs de milieux agricoles pauvres en nourriture
4) Fongicides marqueurs de conditions climatiques humides
...

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Lien de causalité direct déclin ~ fongicide ?

Test en labo

Larves exposées au fongicide le plus fréquent

1) protocole standard, suivi 10 jours : pas d'effet
2) si on prolonge le suivi au delà de 10 jours :
mortalité très forte à toute les doses (pas dans le contrôle)

Questionne les protocoles standard...
On peut probablement conclure à un lien de causalité
MAIS
Est-ce que conditions de labo sont représentatives de la nature ?
Doses ? Condition de vie des abeilles ?
Conditions de "vie" du produit ?

Planification expérimentale

Etudes Abeilles

(Thèse Noa Simon-Delso - CARI)

Une solution ?

Test semi-expérimentaux au champ

On pulvérise le produit sur une culture et on y place les abeilles pex en tunnel

Très coûteux --> le risque est de ne pas faire assez de répétitions pour mettre en évidence des différences subtiles
(plusieurs dizaines de ruches)
--> importance d'analyser la puissance statistique d'un dispositif expérimental

De plus, le tunnel n'est pas très naturel
Interaction avec les conditions climatiques, la culture, d'autres produits,...

Planification expérimentale

--> Même si toutes les études ne se valent pas,
toutes les études sont critiquables
Le doute fait partie de la méthode scientifique.
Ce n'est pas parce qu'il y a un doute sur les résultats/l'interprétation qu'une étude est sans valeur et doit être oubliée.

Ce n'est qu'en combinant différentes études et approches qu'on peut comprendre un phénomène

Les outils statistiques peuvent aider a évaluer le degré de certitude d'un résultat.

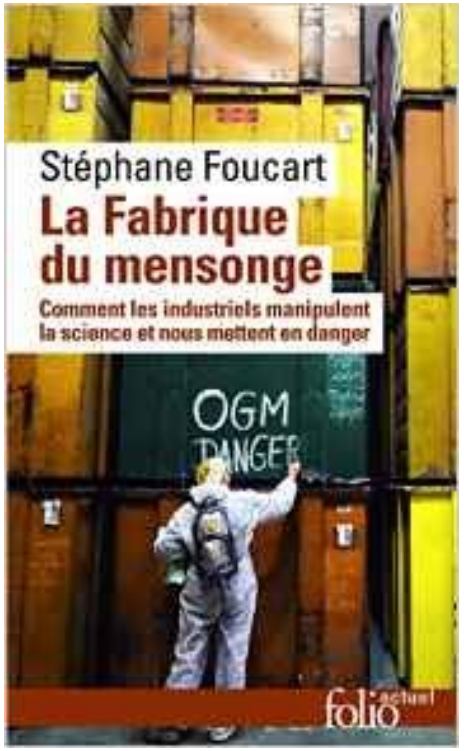
Pour que les résultats soient exploitables, il faut suivre certaines règles garantissant un minimum de qualité

Planification expérimentale

Ces conclusions sont triviales pour beaucoup de scientifiques.
Mais comment autres acteurs de la société (citoyens, politiques, journalistes, industries, organisations gouvernementales,...) voient/utilisent la science ?

2 livres intéressants sur le sujet :

Auteurs : 2 historiens américains



Auteur : 1 journaliste français

Les "règles d'or"

Quelques règles d'or pour **éviter/limiter les problèmes** :

question et population d'intérêt bien définies
adéquation des mesures
réplication
échantillonnage aléatoire
randomisation des mesures et des traitements
indépendance des échantillons
contrôles judicieusement choisis

Quelques règles d'or pour
augmenter la puissance d'un test
et/ou la précision des estimations :

taille d'échantillon maximisée
variabilité résiduelle minimisée
taille de l'effet maximisée

Question et population d'intérêt bien définies

Toujours avoir des questions clairement exprimées
et idéalement des hypothèses et prédictions sur les résultats.

En fonction de la question, définir la population d'intérêt
c'est à dire celle à laquelle on veut généraliser les résultats

Exemple :
On veut étudier l'effet du pâturage sur populations de papillons de jour des pelouses calcaires.

En Europe ? En Wallonie ?
Sur un site bien particulier dont on doit assurer la gestion ?

En fonction de la réponse l'échantillonnage sera différent.

Une erreur classique consiste à extrapoler les résultats au delà de la population échantillonnée.

Adéquation des mesures

Est-ce que les variables mesurées correspondent bien à la question posée ?
Est-ce que ces variables sont mesurées correctement ?

Règle triviale !
Vous pouvez avoir le meilleur design du monde, si le critère d'adéquation n'est pas rempli, en général, votre étude ne sert à rien...

Exemple 1 :
On veut estimer l'effet de l'agriculture biologique sur l'abondance des chauves-souris.
On place des détecteurs d'ultra-sons dans 30 fermes bio et 30 fermes conventionnelles.
On compte le nombre de contacts enregistrés.

--> ce qu'on mesure ici est l'intensité de chasse, pas la taille des populations !

Adéquation des mesures

Exemple 2 :

Un chercheur veut savoir si il existe une relation entre la taille et le poids des œufs d'une espèce d'insecte.

Il commence par peser 1500 œufs ensuite il les remet en vrac dans un pot. Il encode les valeurs et les classe par ordre croissant.

Il mesure ensuite 1482 oeufs (il en a perdu en route) et classe les valeurs encodées par ordre croissant.

Il calcule ensuite la corrélation, qui est très élevée...

On compare ici le poids d'un œuf avec la taille d'un autre...

Adéquation des mesures

Certains cas sont plus subtils...

Exemple 3 :
On veut savoir si la présence d'une espèce de sauterelle est influencée par la hauteur de la végétation.
On parcourt 100 sites où on cherche à vue l'espèce et on note la hauteur de la végétation

Le problème ici est que ce qu'on mesure est à la fois la probabilité de présence mais aussi de détection de l'espèce qui dépend vraisemblablement de la hauteur de la végétation.
Si la détection de l'espèce était indépendante de la variable explicative, çà ne poserait pas de problème (même si la détection varie d'un site à l'autre).
Il s'agit plus ici d'un problème de confusion de facteurs qu'un problème d'adéquation.

Solutions :
Repérer l'espèce au chant
Utiliser des méthodes permettant d'estimer la probabilité de détection (en passant plusieurs fois sur le même site)

Réplication

Les traitements doivent impérativement être répliqués pour pouvoir extrapoler les résultats.

Exemple d'étude sans réplication :
On a fait trois aménagements piscicoles différents sur 3 rivières (une échelle à poisson sur la première rivière, un aménagement naturel des berges sur la seconde, et un aménagement de frayères sur la 3ème).
On fait une pêche électrique avant et après l'aménagement.
On veut savoir quel type d'aménagement est le plus favorable aux poissons.

On peut décrire ce qui s'est passé dans ces 3 cas particuliers mais on ne pourra en tirer aucune conclusion généralisable et on ne pourra établir aucune inférence. Les différences observées pourraient être dues à d'autres facteurs que les aménagements et même simplement au hasard de l'échantillonnage.

Réplication

Mais les traitements ne peuvent pas toujours être répliqués :-(

En particulier dans des études à très large échelle, souvent en milieu aquatique.
Pex : Quel est l'effet d'une pollution aux métaux lourds sur la faune aquatique d'un lac ?

On peut réaliser une étude descriptive (avec pseudo-réplicats) et discuter les résultats de manière prudente en tenant compte de l'absence de réplication.

Ces résultats pourraient peut-être être utilisés plus tard dans une méta-analyse en conjonction avec des études similaires.

Voir "BACI" designs

Échantillonnage aléatoire

Un échantillonnage aléatoire dans l'ensemble de la population est nécessaire pour que **l'échantillon** soit **représentatif**.

Aléatoire signifie que chaque élément de la population a la **même probabilité de se faire échantillonner** que les autres.

On ne prend souvent pas assez de précautions pour rendre l'échantillonnage réellement aléatoire.
L'observateur peut très souvent introduire des biais.

Échantillonnage aléatoire

Exemple 1 :

Des agronomes ont mis au point un modèle prédisant la production de sucre dans les champs de betterave en allant les prélever eux-mêmes dans des champs d'essai.
Le modèle fonctionne très bien.
Pour garantir un échantillonnage aléatoire l'agriculteur doit se placer où il veut en bordure du champ, faire entre 20 et 50 pas vers l'intérieur du champ, jeter par dessus l'épaule un bâton et prélever la betterave qui se trouve la plus proche de la pointe. Il doit ensuite recommencer 15 fois et envoyer les betteraves pour analyse.
Lors de la mise en pratique chez les agriculteurs on se rend compte que le modèle sur-estime systématiquement la production

Que se passe-t-il ?

Les agriculteurs introduisent un biais systématique en éliminant inconsciemment ou non les betteraves les plus chétives et en évitant des parties du champ moins belles.

Échantillonnage aléatoire

Bien souvent on prétend avoir récolté des données aléatoirement ou "au hasard" alors qu'il n'en est rien en réalité

Exemple 2 :

On dispose "aléatoirement" des cadrats sur un site pour faire des relevés botaniques. En fait on se ballade sur le site et on décide à un moment "tiens, je vais mettre mon quadrat ici" ou au mieux, on le jette par dessus l'épaule. Mais en pratique on va souvent (comme dans le cas des betteraves) éviter inconsciemment certaines zones qui nous semblent "pas représentatives"

Exemple 3 :

On sélectionne "aléatoirement" 15 sites avec des pins noirs de plus 5 m de haut pour y réaliser une étude sur les communautés de coccinelles
En pratique, ces sites ne sont pas très faciles à trouver, on les cherche et on prend les 15 premiers qu'on trouve... Souvent on choisit aussi des sites accessibles ou pas trop éloignés de son domicile/lieu de travail, etc...

A chaque fois que c'est possible, on devrait utiliser des générateurs de nombre aléatoires pour sélectionner réellement aléatoirement ses sites, la position des quadrats, etc...

Échantillonnage aléatoire

Dans certains cas l'expérimentateur choisi délibérément de ne pas faire un échantillonnage aléatoire en particulier quand il veut échantillonner des événements rares.
C'est parfois une bonne stratégie mais gare aux conséquences...

Exemple 4 :

Une chercheuse veut caractériser les sites de ponte d'un papillon. Elle veut ensuite produire une carte prédictive de la qualité des sites de pontes afin de savoir quels sont les sites les plus propices et ceux qui peuvent être améliorés par exemple par des mesures de gestion.
Elle sélectionne 50 plantes hôtes avec des œufs et 50 plantes hôtes sans œufs. Elle caractérise ensuite ces plantes (taille, nombre de feuilles, exposition,...) et met en relation ces valeurs avec la présence/absence des oeufs.

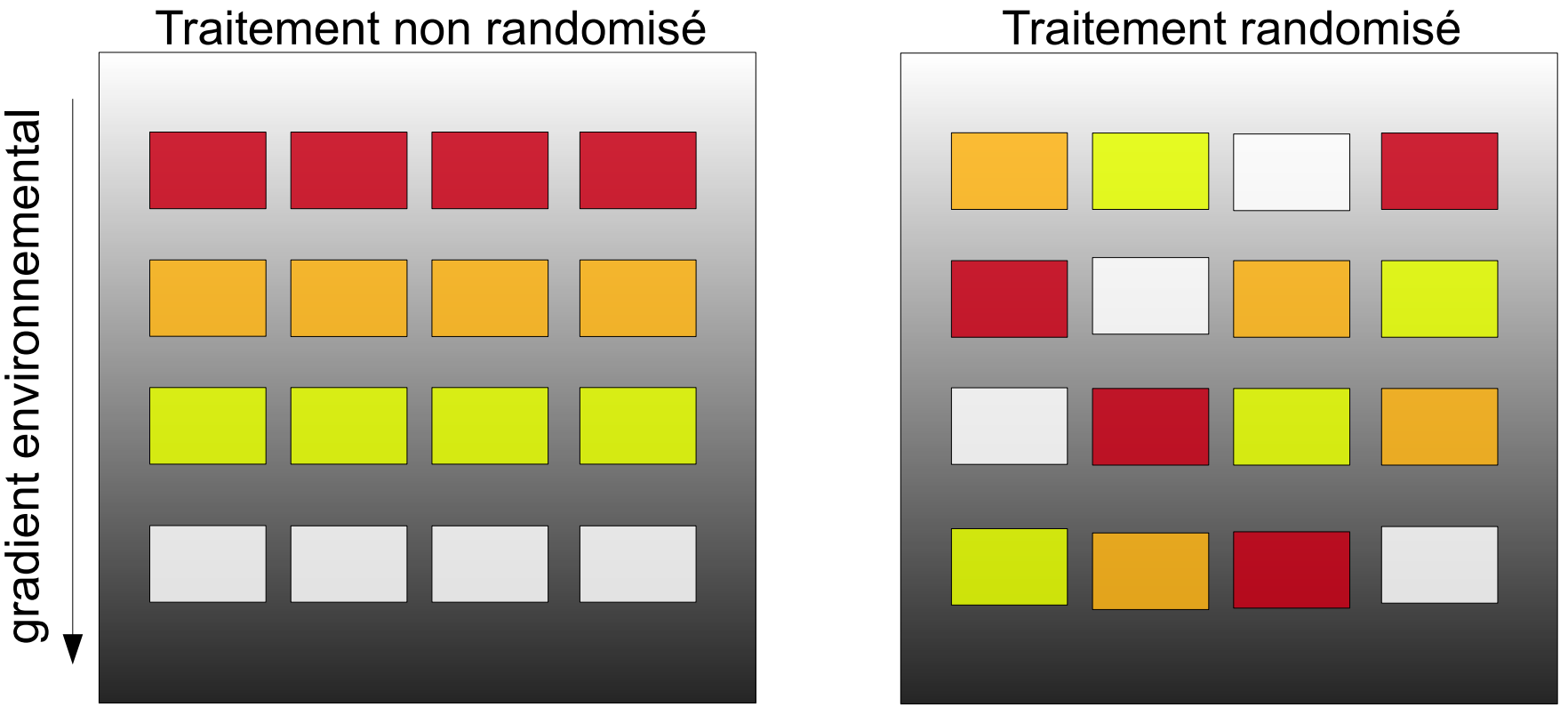
Avec cette approche on pourra mettre en évidence les variables les plus importantes pour caractériser les sites de pontes mais on ne pourra pas faire de carte prédictive parce que les plantes avec des œufs sont vraisemblablement surreprésentées dans notre échantillon par rapport à un échantillon aléatoire.
On peut estimer un effet relatif mais pas absolu...

Randomisation des mesures et traitements

La randomisation permet de limiter la confusion de facteurs càd les cas ou les effets observés peuvent être dus à plusieurs facteurs en même temps sans qu'on puisse les distinguer

Exemple 1:

Confusion entre le traitement et un gradient environnemental inconnu



Cas particulier : "carré latin"

Randomisation des mesures et traitements

La randomisation des mesures

Exemple 2:

Riri, Fifi et Loulou font une étude sur l'effet de pesticides sur des mouches cécidogènes du froment. Ils ont pulvérisé des parcelles avec 8 produits différents (et un témoin). Ils fauchent les parcelles et prélèvent aléatoirement 90 plants sur lesquels il faut chercher des galles pas toujours faciles à trouver...

Cas1 :

Par souci d'efficacité Riri s'occupe des 3 premiers traitements, Fifi des 3 suivants et Loulou des 3 derniers. Mais en fait Riri est beaucoup plus expérimenté et trouve beaucoup plus de galles, et Loulou en a marre de tout le temps chercher ces galles et en trouve beaucoup moins... Il y a confusion entre l'effet du traitement et l'effet observateur.

Cas 2 :

Pour éviter la confusion traitement-observateur, ils prennent d'abord le premier traitement et chacun s'occupe de 30 plans, ensuite ils passent au traitement suivant. Le problème est que au cours du temps leur capacité à trouver des galles s'améliore. il y a donc une confusion entre l'effet du traitement et l'ordre dans lesquels ils sont mesurés.

Conclusion : utiliser un générateur de nombres aléatoires pour distribuer les lots de 30 plantes...

Indépendance des échantillons

Toutes les inférences sont basées sur la quantité d'information indépendante disponible.
Si les échantillons ne sont pas réellement indépendants on "ment" sur la quantité d'information, et les p-valeurs, erreurs standards, intervalles de confiance sont faux.

On parle souvent de "**Pseudoréplication**"

Indépendance des échantillons

Très souvent c'est le design d'échantillonnage lui même qui implique la pseudoréplication

Exemple extrême :
On veut savoir si les hommes ont en moyenne des cheveux plus longs ou plus courts que les femmes. On sélectionne un homme et une femme et on mesure 50 de leurs cheveux. On compare les moyennes avec un test de student.
Ici : l'échantillonnage est répliqué mais pas le "traitement" (homme/femme).
La population échantillonnée ici sont les cheveux de ces deux personnes en particulier et pas des hommes/femmes en général.

Il n'est pas rare de voir des études comme celle-ci :
On veut caractériser l'effet de l'arboriculture biologique sur les communautés de carabes. on sélectionne un verger "bio" et un verger "conventionnel" et on place 30 pièges pitfall dans chacun. On compare ensuite par exemple le nombre moyen d'espèces par piège ou l'abondance des différentes espèces.
Ce qu'on mesure ce sont les différences entre ces deux sites particuliers qui diffèrent par le mode d'agriculture mais aussi sans doute par de nombreuses autres caractéristiques

Indépendance des échantillons

"Nested designs"

= mesures répétées, sous-échantillons, au sein d'un même réplicat
(pas spécialement au cours du temps)

Que vaut-il mieux ? :

2 vergers avec 30 pièges par verger ?

12 vergers avec 5 pièges par verger ?

60 vergers avec 1 piège par verger ?

Dans la grande majorité des cas, il faut maximiser les réplicats indépendants càd ici les vergers quitte à n'avoir aucun pseudoréplicat (ici les pièges).
C'est particulièrement vrai quand la variation entre pièges d'un même verger est faible. Dans ce cas on ne fait que mesurer plusieurs fois la même chose.

Indépendance des échantillons

"Nested designs"

Problème :
Même si au total on a le même nombre de pièges, échantillonner 60 vergers avec 1 piège sera sans doute plus coûteux que 12 vergers avec 5 pièges.
Ajouter des pseudoréplicas est souvent peu coûteux par rapport aux vrais réplicats et apporte malgré tout une information potentiellement utile.
--> les designs "hiérarchisés" (nested designs) sont très fréquents

2 problèmes classiques cependant avec les "nested designs" :

1) On a trop peu de vrais réplicats
(pex 4 vergers avec 15 pièges par verger)

2) On compare les 30 pièges "bio" aux 30 pièges conventionnels comme si ils étaient indépendants

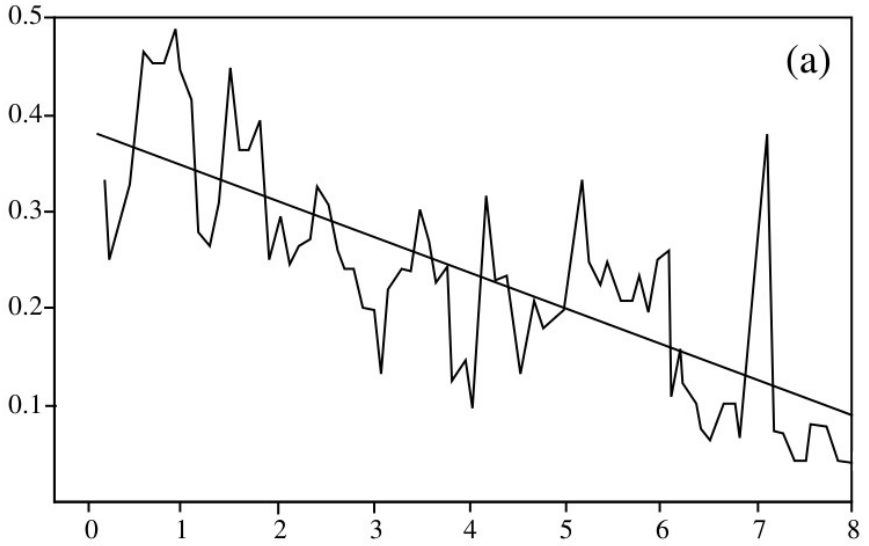
2 solutions principales :
a) prendre la moyenne par verger (mais on perd de l'info)
b) inclure la variable "verger" comme facteur aléatoire (bloc) dans l'analyse

Indépendance des échantillons

Autres cas fréquents de non indépendance

Corrélation temporelle

exemple : on veut comparer l'évolution des populations de chevreuils au cours du temps entre plusieurs modes de gestion cynégétique.
On nous finance grassement et on peut donc mesurer précisément les populations de chevreuils toutes les secondes pendant 10 ans...
On obtient donc plus de 315 milliards de points par population mais ces points sont-ils réellement indépendants ?



Indépendance des échantillons

Autres cas fréquents de non indépendance

Corrélation spatiale

Pour différentes raisons qui seront détaillées plus tard, deux points d'échantillonnage très proches dans l'espace sont *a priori* susceptibles d'avoir des valeurs similaires.

Idéalement il faut éviter les points d'échantillonnage trop proches

Contrôles judicieusement choisis

Les **"contrôles" ou "témoins"** sont des traitements qui n'ont généralement pas d'intérêt en eux-mêmes mais qui **permettent d'éliminer certains facteurs de confusion**, certaines explications autres que le traitement d'intérêt qui permettraient d'expliquer le résultat.

Un ou plusieurs contrôles judicieusement choisis peuvent faire toute la différence au niveau de l'interprétation.

Certaines études observatives peuvent mettre en évidence des lien de cause à effet si les contrôles et l'échantillonnage sont bien choisis ("quasi-experimental designs")

Les témoins doivent en général être les plus proches possibles des traitements à l'exception de l'hypothèse que l'on veut éliminer.
Les témoins sont rarement des échantillons où on a "rien fait".

Contrôles judicieusement choisis

Exemple ecotox :

On teste la toxicité d'un produit sur des insectes et on observe 100 % de mortalité. On peut être tenté de dire que le produit est toxique. Mais qu'est-ce qui nous prouve que notre pulvérisateur n'était pas contaminé par un autre produit ou que notre souche d'insectes était malade ?

-> on ajoute un contrôle négatif : on pulvérise quelques réplicats avec de l'eau

On teste un autre produit et on observe 0 % de mortalité.
Est-ce que le produit n'est pas toxique ? Peut-être.
Ou alors notre pulvérisateur était défectueux, ou il y avait trop de vent et le produit n'est pas arrivé sur les insectes ou bien on a une souche d'insectes particulièrement résistants.

--> on ajoute un contrôle positif : on pulvérise un produit dont la toxicité est bien connue.

--> plusieurs contrôles différents peuvent être utiles pour éliminer des facteurs de confusion différents

L'expérimentateur a souvent tendance à être trop bienveillant avec ses propres études !
--> se mettre dans la peau de qqn de mal veillant pour choisir ses contrôles !

Contrôles judicieusement choisis

Exemple "BACI" = "Before-After Control-Impact" designs

On veut évaluer l'effet de la fauche dans des prairies humides sur les population d'une espèce de papillons.

On choisi 30 sites que l'on fauche chaque année pendant 10 ans et on compte chaque année les papillons.
On constate une augmentation des effectifs. Peut-on en conclure que ce mode de gestion est favorable ? Peut-être, à moins que les populations n'étaient déjà en augmentation avant le début du traitement.
--> idéalement on devrait récolter aussi 10 années de données avant de commencer à faucher. C'est rarement possible...

On recommence avec une autre espèce de papillon et des données pendant 10 ans avant et après le début du fauchage. On ne constate aucune évolution du nombre d'individus (la population stagne). Est-ce qu'on peut conclure que ce mode de gestion n'a aucun effet ? Peut-être, sauf si dans les sites où il n'y a aucun fauchage les populations sont en augmentation. Dans ce cas l'effet est probablement négatif.
Il faut donc idéalement ajouter des parcelles contrôles (sur les mêmes sites ? dans des sites différentes les plus proches et similaires possibles?). Surtout quand on a pas de données avant...

Puissance d'un test

La puissance d'un test se mesure comme la probabilité de rejeter l'hypothèse nulle quand elle est effectivement fausse.

Autrement dit plus un test est puissant plus il arrivera à "détecter" les relations/différences réelles entre échantillons.
On veut en général maximiser la puissance des tests à coût égal et en tous cas avoir une puissance suffisante pour détecter des effets biologiquement significatifs

La puissance dépend :

du type de test
du seuil de significativité alpha
de la taille de l'effet
de la variabilité
de la taille de l'échantillon

NB : La précision des estimateurs dépend principalement de ces 2 derniers points qui sont globalement ceux sur lesquels on peut le plus jouer

Puissance d'un test

La puissance dépend : du type de test

Pour un échantillon parfaitement identique certains tests seront parfois plus puissants.
Par exemple si on a bien une relation linéaire entre deux variables, le test de corrélation de Pearson est plus puissant que le test de corrélation de rang de Spearman.

Mais en général on choisi le test le plus adapté à son cas.
--> on joue rarement sur ce point volontairement.

Puissance d'un test

La puissance dépend : du seuil de significativité alpha

Un moyen trivial d'augmenter la puissance est simplement de considérer par exemple que tout $p<0.1$ est significatif (au lieu du traditionnel $p<0.05$).

On augmente cependant dans ce cas la probabilité de dire qu'il y a un effet alors qu'il n'y en a pas (erreur de type I, on rejette l'hypothèse nulle alors qu'elle est vraie).

Si on diminue alpha par contre, on augmente les erreurs de type II : c'est à dire qu'on a plus de chance de "rater" un effet réel (on ne rejette pas l'hypothèse nulle alors qu'elle est fausse).

Puissance d'un test

La puissance dépend : du seuil de significativité alpha

C'est rare mais il arrive que l'on joue sur ce paramètre.

Par exemple on peut considérer qu'il est moins grave de dire qu'une espèce est en déclin alors qu'elle ne l'est pas en réalité que de de manquer une espèce réellement en déclin en prétendant qu'elle ne l'est pas.

Ou qu'il est plus grave de dire qu'une substance n'est pas nocive pour la santé humaine alors qu'elle l'est en réalité que d'affirmer qu'une substance est nocive alors qu'en réalité elle ne l'est pas.

--> dans ces deux cas on choisirait plutôt un seuil à 0.1

Puissance d'un test

La puissance dépend : de la taille de l'effet

Toutes choses étant égales par ailleurs, on "détectera" plus vite une corrélation réelle de 0.8 que de 0.1.

Dans certains cas on peut jouer sur ce point en choisissant des traitements plus extrêmes

Par exemple si on veut voir si la température a un effet sur la ponte d'un insecte on choisira des 2 températures les plus extrêmes possibles (soit en contrôlant la température, soit en choisissant des sites extrêmes).

Si on veut vérifier si la relation est linéaire on rajoutera quelques échantillons à une température intermédiaire.

C'est aussi aux extrémités qu'il faut mettre le plus de réplicats car c'est là qu'on a le moins de précision.

Puissance d'un test

La puissance dépend : de la variabilité

Il y a trois moyens principaux de jouer sur ce point :

1) On peut essayer de contrôler au mieux les conditions expérimentales, choisir des sites les plus similaires possibles, ...

2) on peut mesurer une covariable qui ne nous intéresse pas en tant que telle mais que l'on sait avoir un effet sur la variable d'intérêt
On pourra alors enlever au moyen d'outils statistiques la variabilité due à cette covariable avant d'examiner l'effet de notre traitement.

Exemple :
On veut mesurer en champ l'effet de divers traitements azotés sur la production céréalière en Wallonie. Les tests sont répartis dans plusieurs champs en Wallonie. On sait que le rendement est aussi influencé par le nombre d'heures d'ensoleillement que l'on mesure donc sur chaque champ. On peut alors enlever la variabilité due à l'ensoleillement avant d'examiner l'effet des traitements azotés.

Puissance d'un test

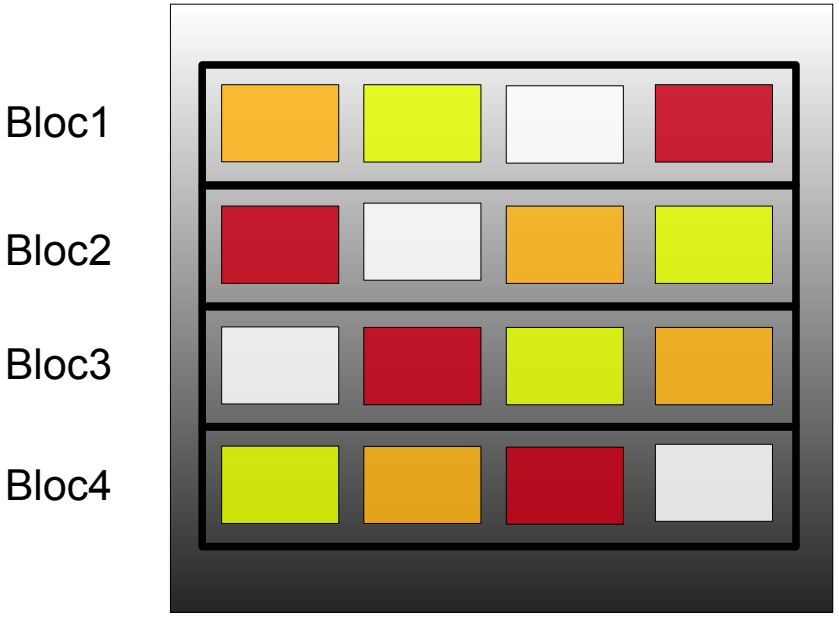
La puissance dépend : de la variabilité

3) Dans certains cas, on suspecte une hétérogénéité environnementale mais on ne sait pas exactement quels facteurs entrent en jeu ou on ne sait pas les mesurer.
Dans ce cas une technique très utile consiste à créer ce qu'on appelle des "blocs".

Puissance d'un test

La puissance dépend : de la variabilité

Un bloc est une surface ou période de temps ou toute autre unité que l'on considère comme relativement homogène. On assigne ensuite au sein de chaque bloc de manière aléatoire les différents traitements.
Les blocs vont permettre de tenir compte statistiquement la variation environnementale entre blocs.



Cas particulier : "carré latin" : on pourrait ajouter des blocs verticaux qui seraient moins utiles dans ce cas

Puissance d'un test

La puissance dépend : de la taille d'échantillon

Plus on a d'échantillons indépendants plus on aura de puissance mais en général le coût augmente en proportion.
Dans certains cas cependant on peut faire des choix à coût égal.

Par exemple pour les comptages hivernaux de chauves-souris est-ce qu'il vaut mieux suivre 50 grottes chaque année ou 100 grottes une année sur deux ?

Puissance d'un test

La puissance dépend : de la taille d'échantillon

Une question très fréquente est de savoir quelle taille d'échantillon on a besoin. Dans ce cas il faut bien sûr définir la taille d'effet que l'on veut pouvoir détecter (et il faut avoir une idée de la variabilité)

Par exemple : combien de grottes faut-il suivre pour pouvoir détecter un déclin de 30 % en 10 ans ?
Ou bien : si on suit 100 grottes, combien d'années faudra-t-il avant de détecter un déclin significatif de 30 % ?

--> c'est le rôle de l'**analyse de puissance**

Puissance d'un test

L'analyse de puissance

En fixant la taille de l'effet (sur base de seuils biologiquement importants), la variabilité (issue d'expériences précédentes), le niveau alpha et le test, on peut déterminer la puissance pour différentes tailles d'échantillon.

En général on essaye d'atteindre une puissance de 0.8 (pure convention)

Autrefois il fallait dériver des formules mathématiques complexes pour chaque type de test. Aujourd'hui on peut utiliser des méthodes de simulation qui peuvent s'appliquer à des cas beaucoup plus complexes.

Puissance d'un test

L'analyse de puissance

On peut aussi par exemple fixer la taille d'échantillon (au maximum possible) et estimer la puissance pour différentes tailles d'effet pour déterminer si une expérience vaut le coup d'être menée.

NB : certains statisticiens recommandent l'analyse de puissance pour établir un design avant l'expérience pas pour l'interpréter à postériori