Graphes et tests pour l'analyse de données expérimentales

M2 Physiologie et Pathologies Musculaires

M. Bailly-Bechet
(et une importante contribution de M-L Delignette-Muller)

Université Claude Bernard Lyon 1
Laboratoire de Biométrie et Biologie Evolutive
Bât. Mendel 2ème étage, côté rouge

Automne 2012

Table des matières

Objectifs du cours

Graphes : montrer l'importance des graphes en tant que vecteur d'information ; donner quelques techniques simples pour améliorer ses graphiques.

Tests : savoir ce que peut faire un test statistique et quelles sont ses limites ; connaître les problèmes auxquels vous serez confronté dans des labos de recherche.

Planification expérimentale : éviter de commettre des bévues de design expérimental qui peuvent couter cher en temps, en argent et en résultats.

Table des matières

Graphiques statistiques

- Ils ne sont pas obligatoires, mais fortement conseillés.
- Ils donnent un maximum d'informations dans un minimum de place.
- Ils doivent être cités dans le texte, mais. . . .
- . . . ils doivent être informatifs en eux-mêmes grâce à leurs légendes, sous-titres, notes.

Dans un article scientifique, le texte s'organise autour des figures – et donc des graphiques – pas l'inverse.

Quels graphiques pour quelles données ?

Un sujet trop long pour être abordé ici en détails. . . vous trouverez de nombreux exemples là : `http://pbil.univ-lyon1.fr/members/mbailly/Biologie_Modelisation/R_graphiques.pdf`

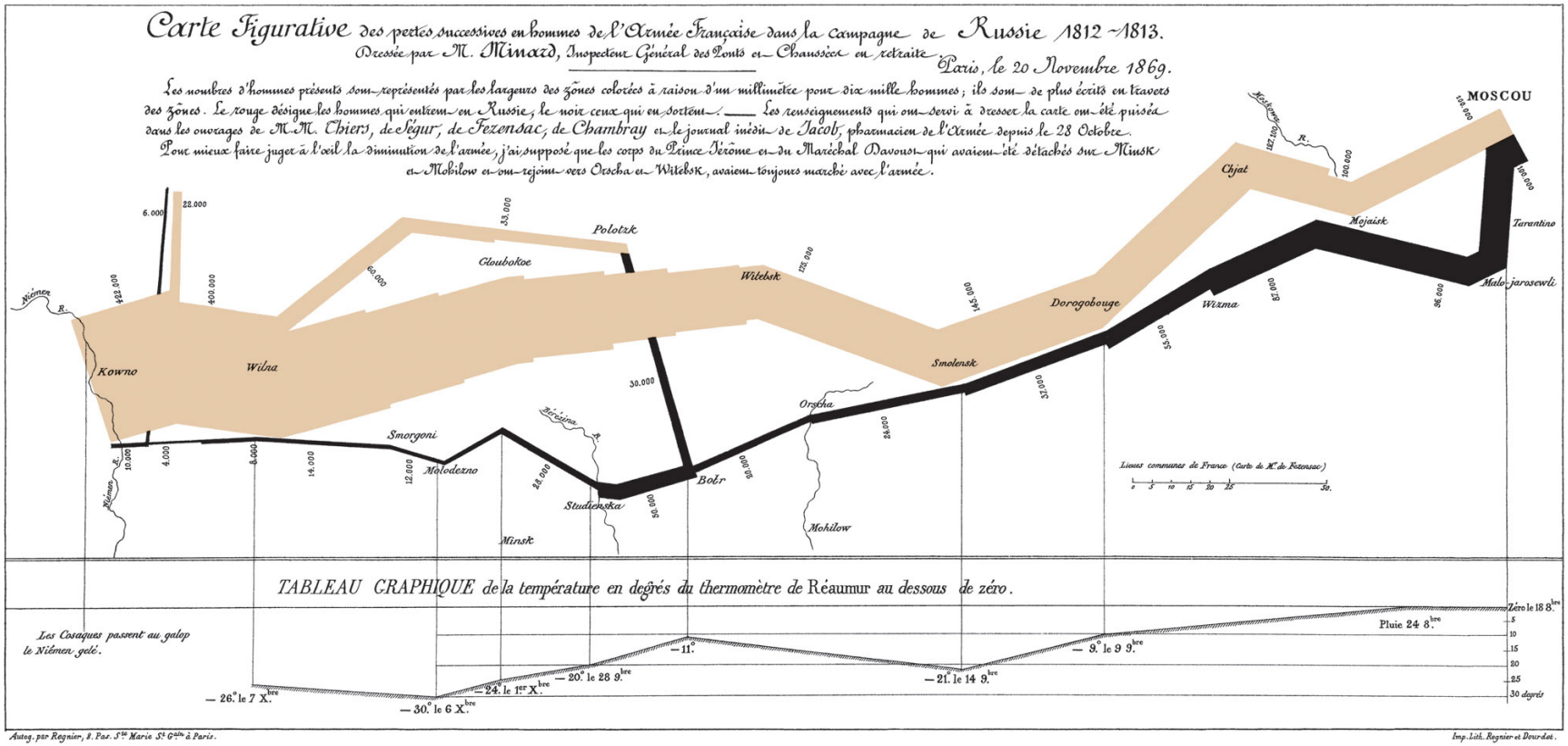
Si vous vous intéressez à des aspects plus poussés, lire la bibliographie de Edward Tufte (voir son site web) est une excellente idée. Aller sur `http://www.datavis.ca/gallery/index.php` en est une autre.

Avec quel logiciel ?

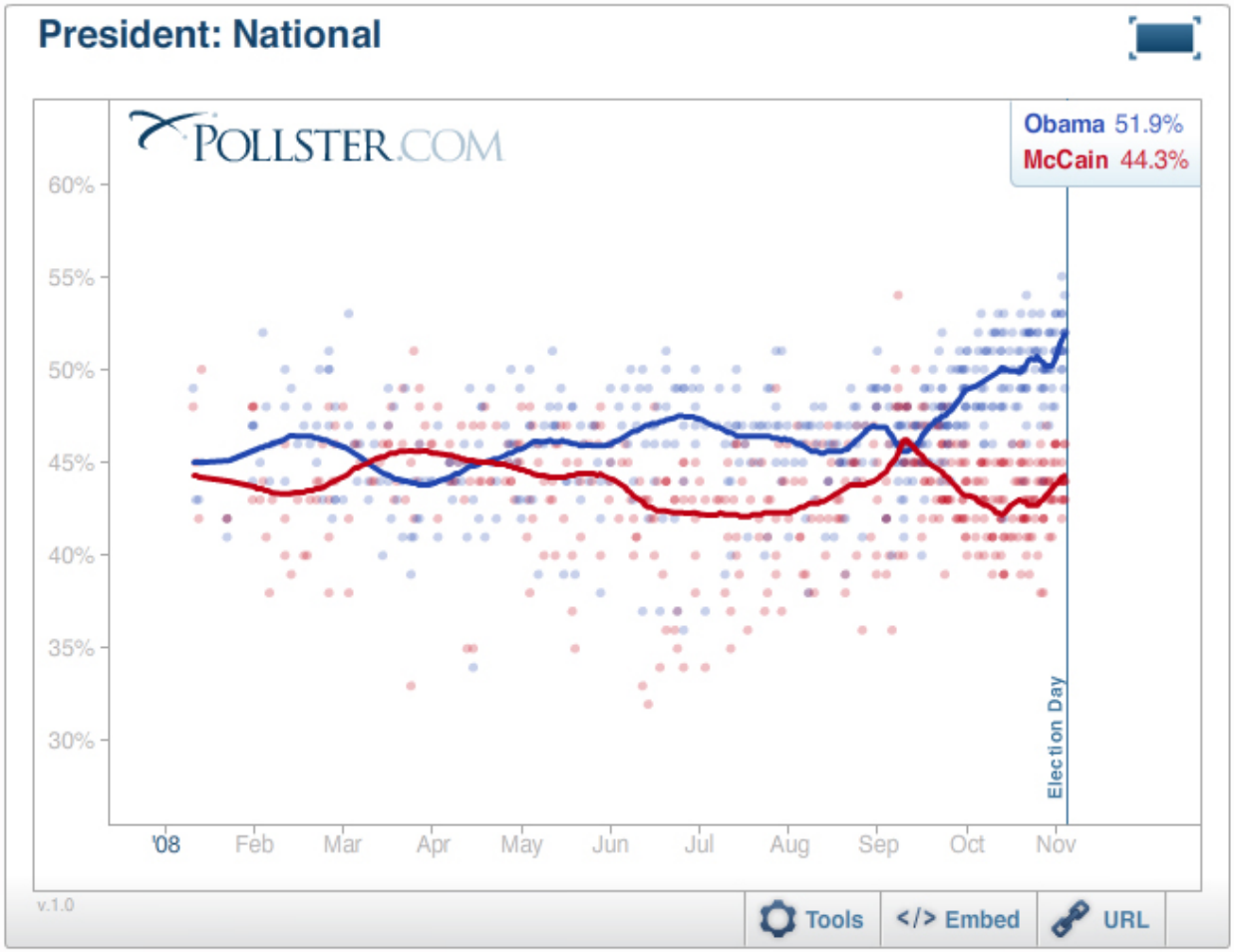
 est très bien, mais vous serez parfois contraints (ou heureux) d'utiliser autre chose.

De toute la finesse et l'expertise que vous pourrez déployer au laboratoire, ce qui en sera perçu à l'extérieur passera essentiellement par les graphes avec lesquels vous communiquerez.

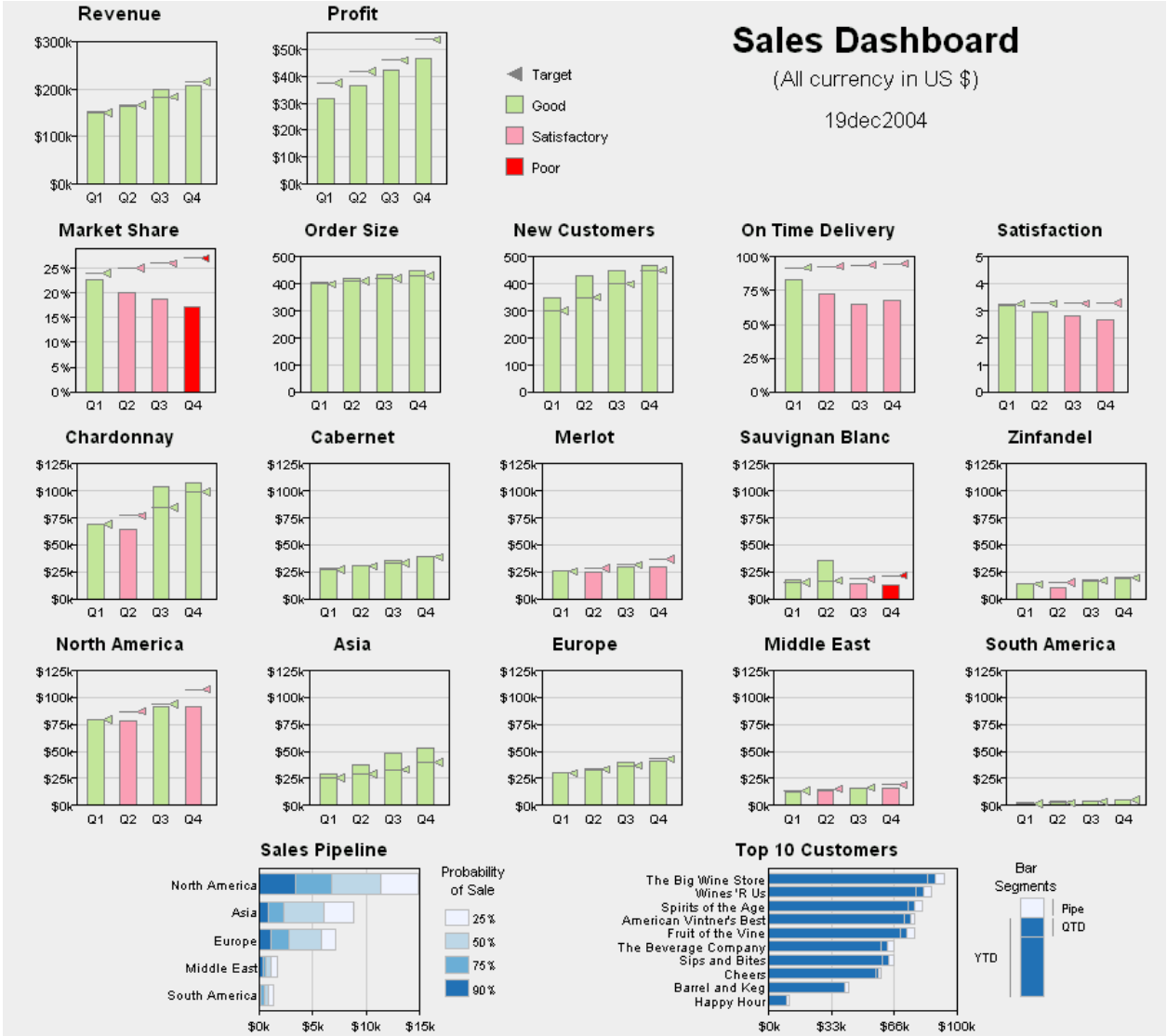
Un bon graphique : la campagne de Russie de Napoléon (1812) vue par Charles Minard



Plus récemment : élections américaines de 2008



Et quand tout ne peut pas tenir sur un graphe



Préservez la nature et le contexte des données

- Toutes les données doivent apparaître.
- La nature et les unités des données doivent apparaître.
- L'échelle doit être constante sur toute la figure.
- Les niveaux de base (comme une mesure de 0, la moyenne. . .) doivent être clairement identifiables.
- Le contexte des données doit être indiqué – par rapport à quoi sont-elles comparables ?

Contre-exemple 1 : niveau de base

Avant	Après
89	75

Table: Poids de l'individu 37, avant et après le traitement révolutionnaire *MegaSlim*

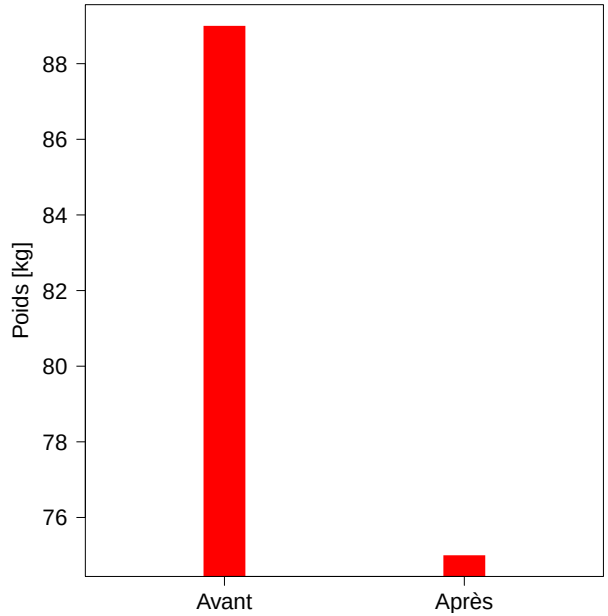


Figure: Poids de l'individu 37, avant et après le traitement révolutionnaire *MegaSlim*

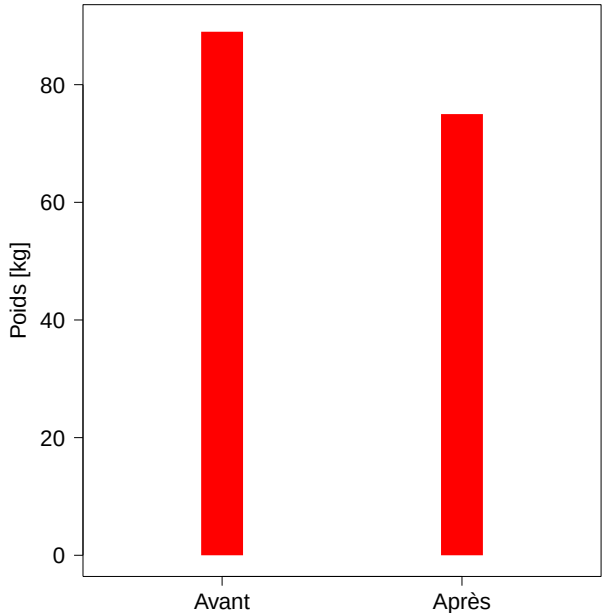
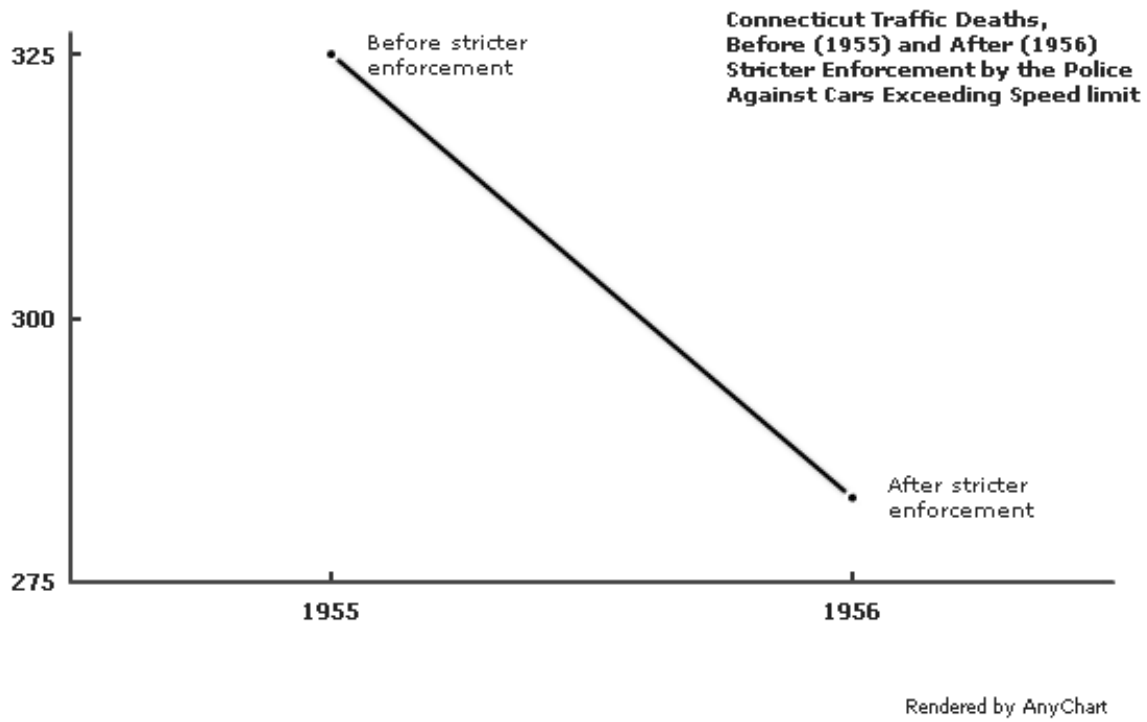


Figure: Poids de l'individu 37, avant et après le traitement révolutionnaire *MegaSlim*

Contre-exemple 2 : contexte

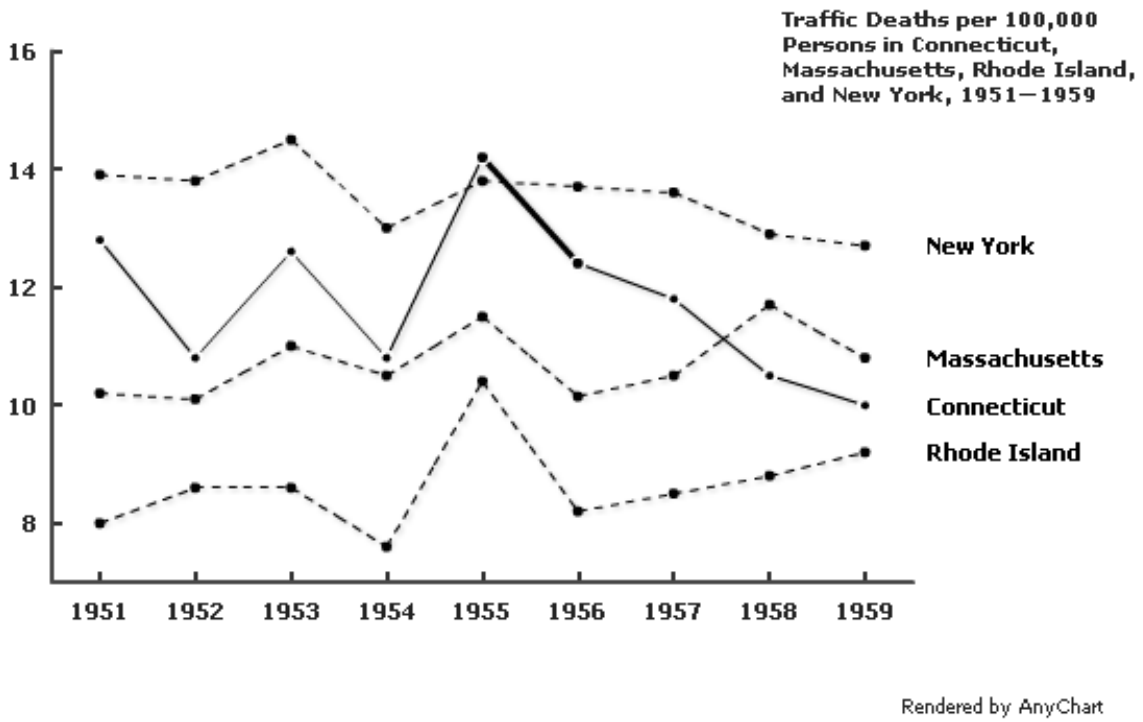


E. Tufte, The visual display of quantitative information

Contre-exemple 2 : contexte

E. Tufte, The visual display of quantitative information

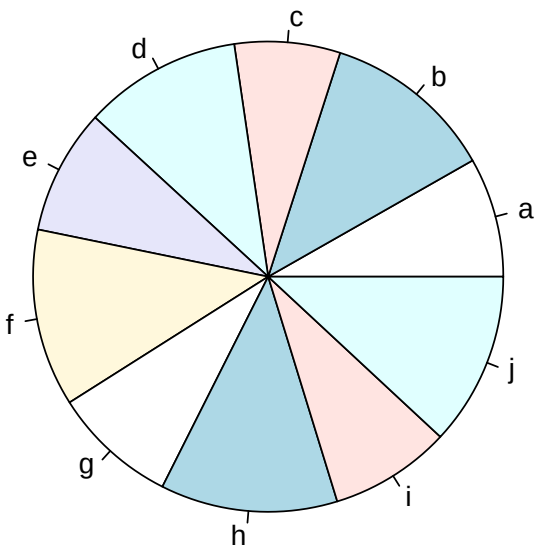
Contre-exemple 2 : contexte



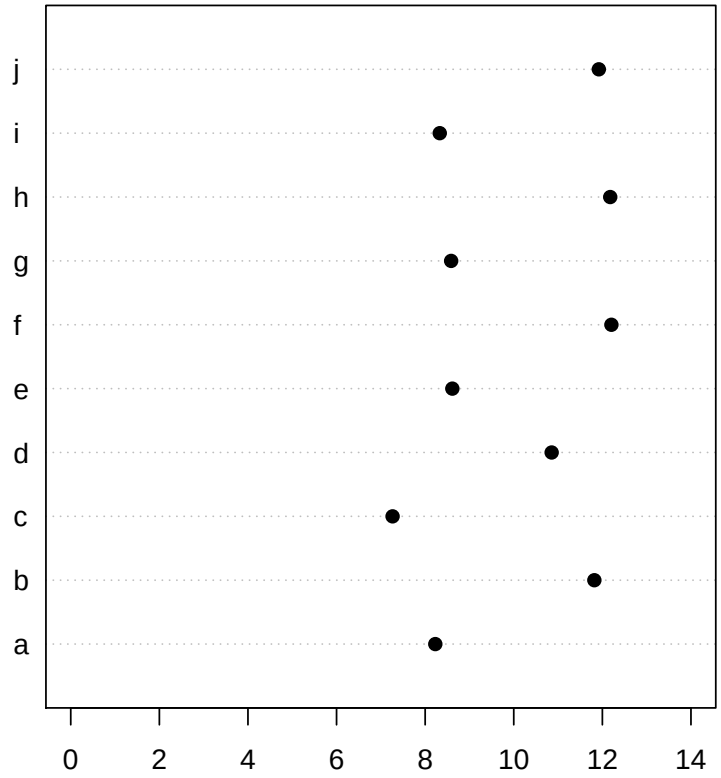
E. Tufte, The visual display of quantitative information

Contre-exemple 3 : angles et mesures

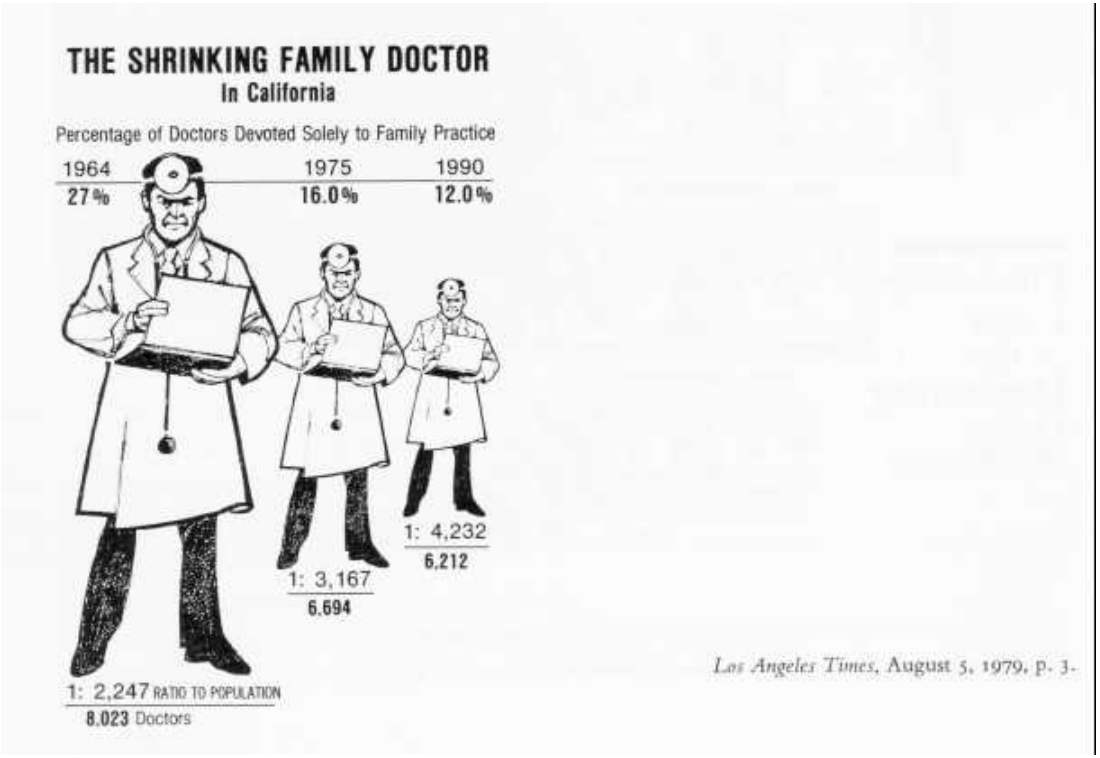
Diagramme en secteur



Graphe de Cleveland

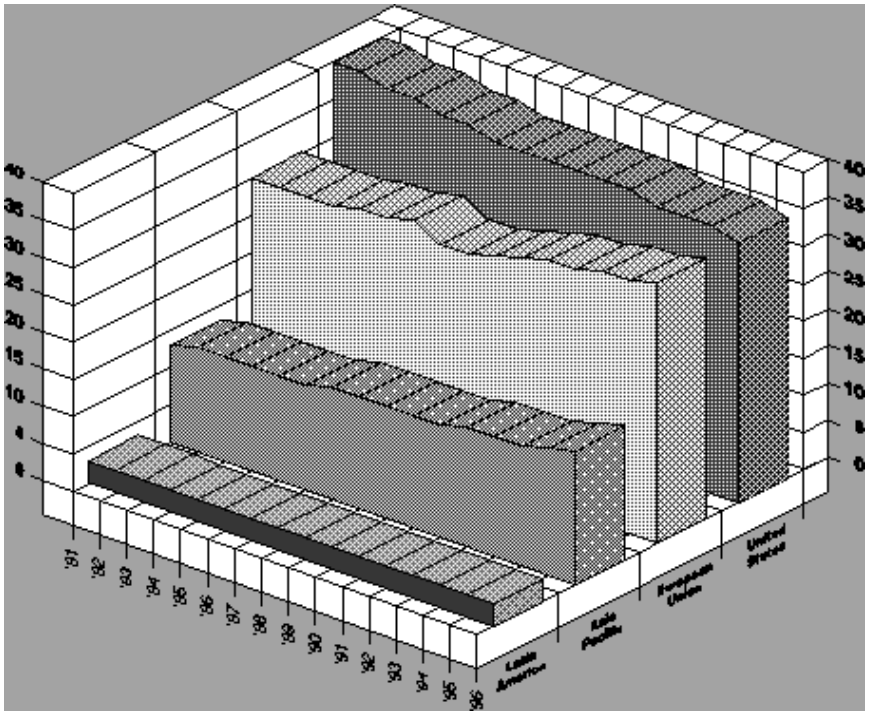


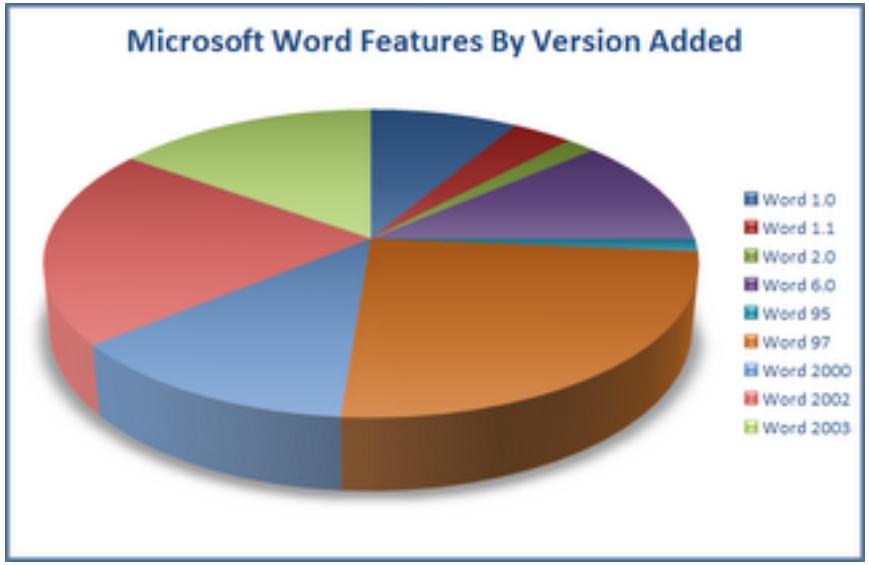
Contre-exemple 3^2 : aires et mesures



Mesure linéaire représentée par une "surface" : on augmente dramatiquement la différence perçue par rapport à ce qui serait observé sur une courbe.

Contre-exemple 3^3 : volumes et mesures



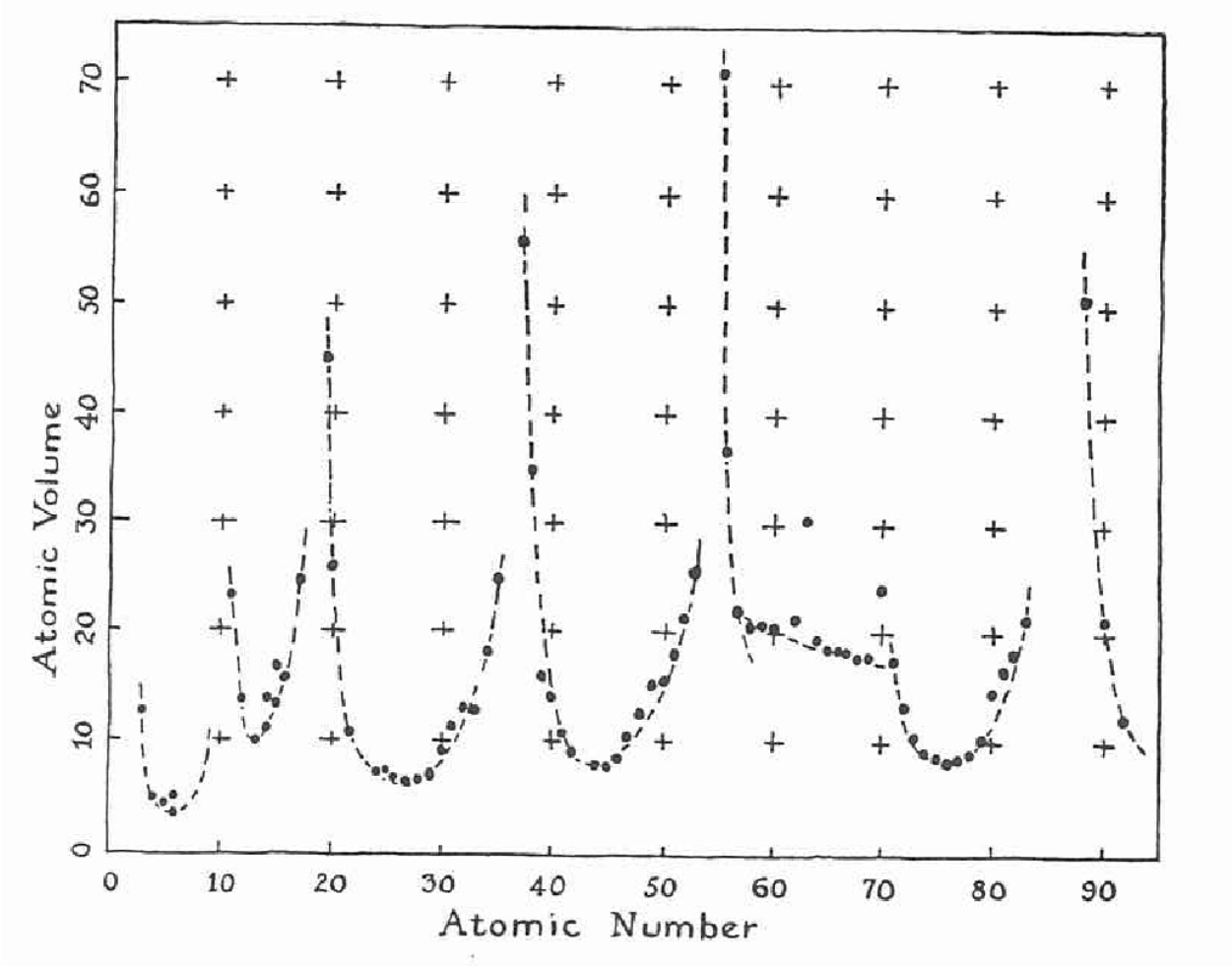


Attention aux choix des couleurs !

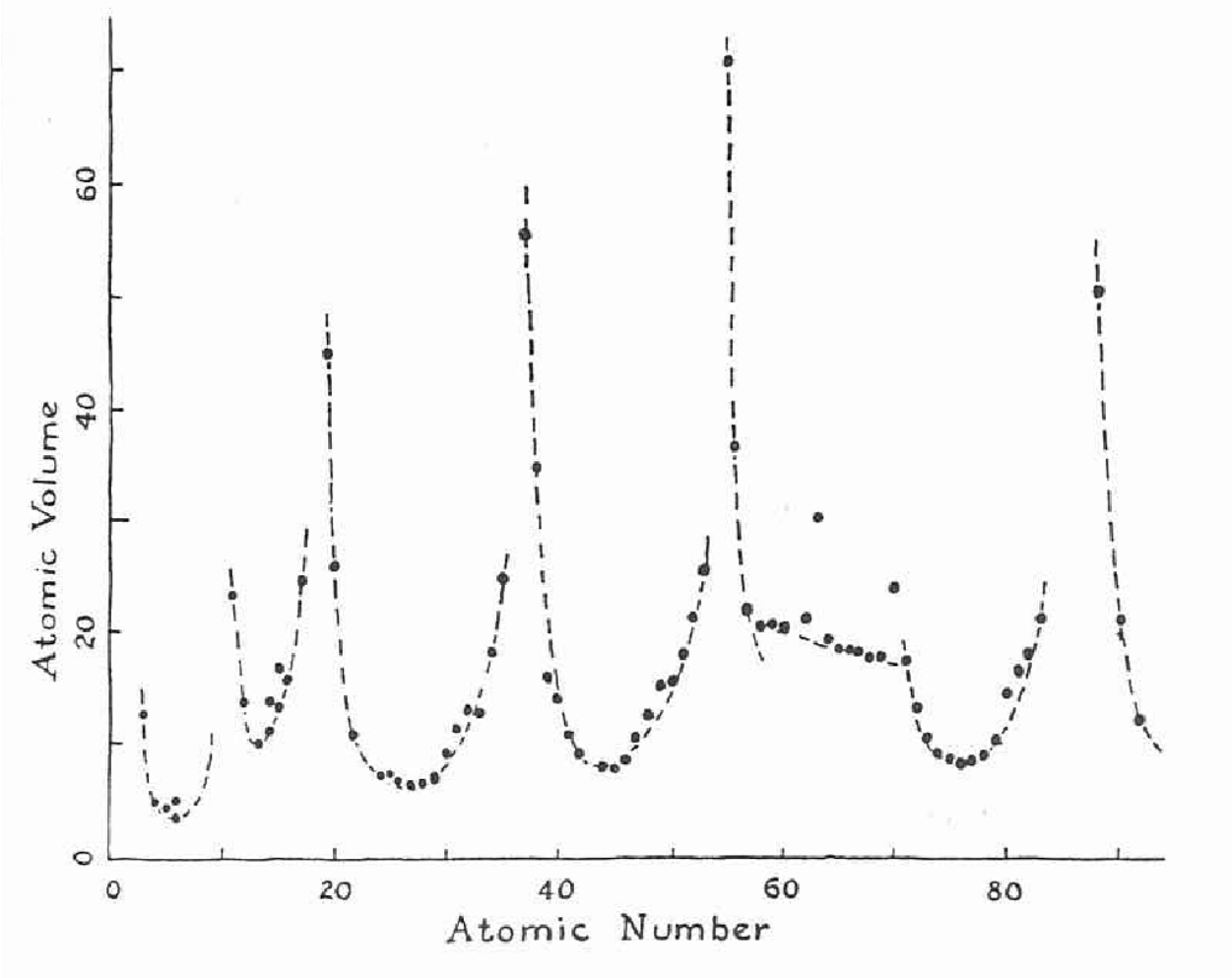
Notions de daltonisme

	Normal Vision	L-cone defect	M-cone defect	S-cone defect
Men	91.4%	2.45%	6.1%	0.011%
Women	99.6%	0.04%	0.36%	0.04%
Overall	95.5%	1.25%	3.24%	0.025%

Rationalisation d'une figure : exemple



Eliminez tout ce qui n'est pas utile . . .



. . . et ajoutez tout ce qui peut l'être

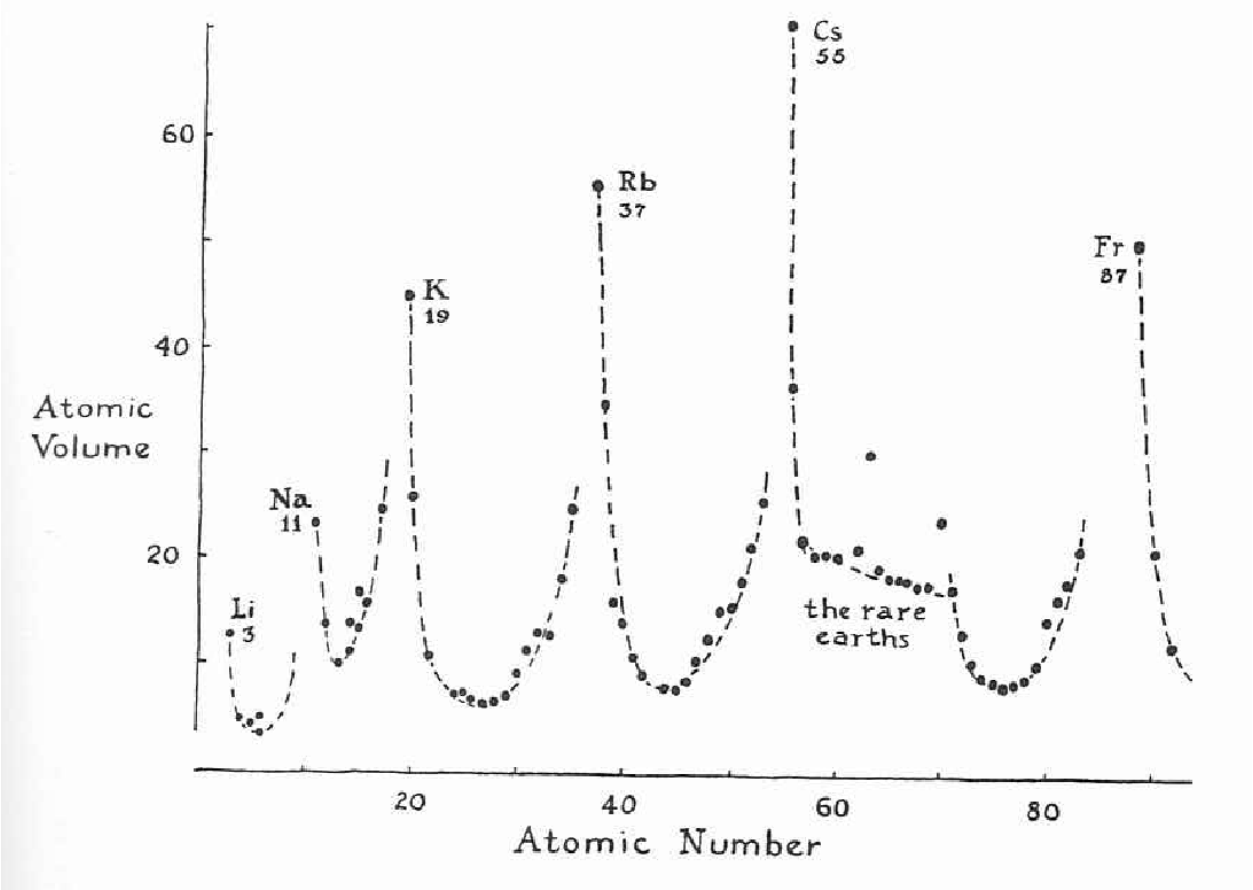
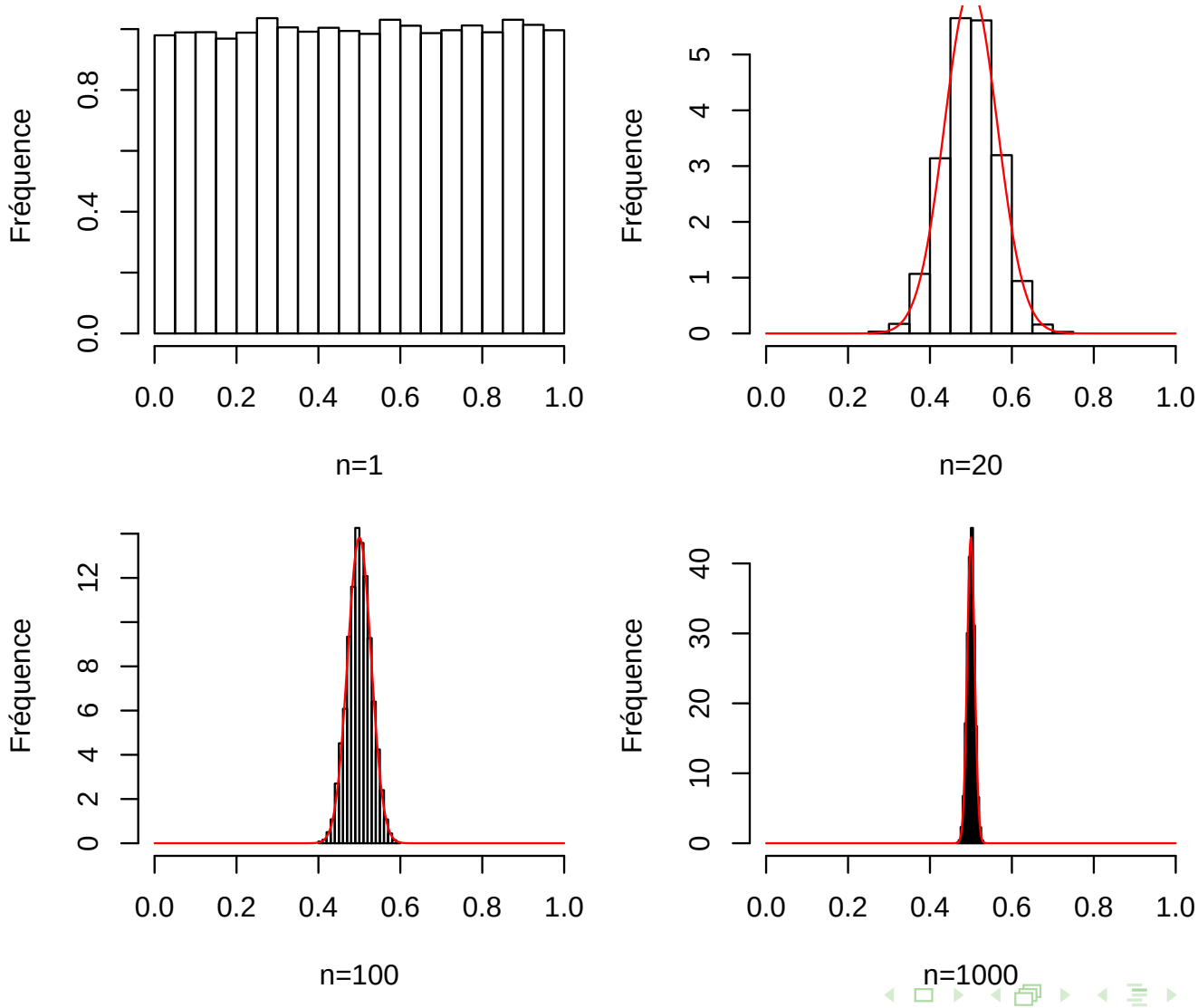


Table des matières

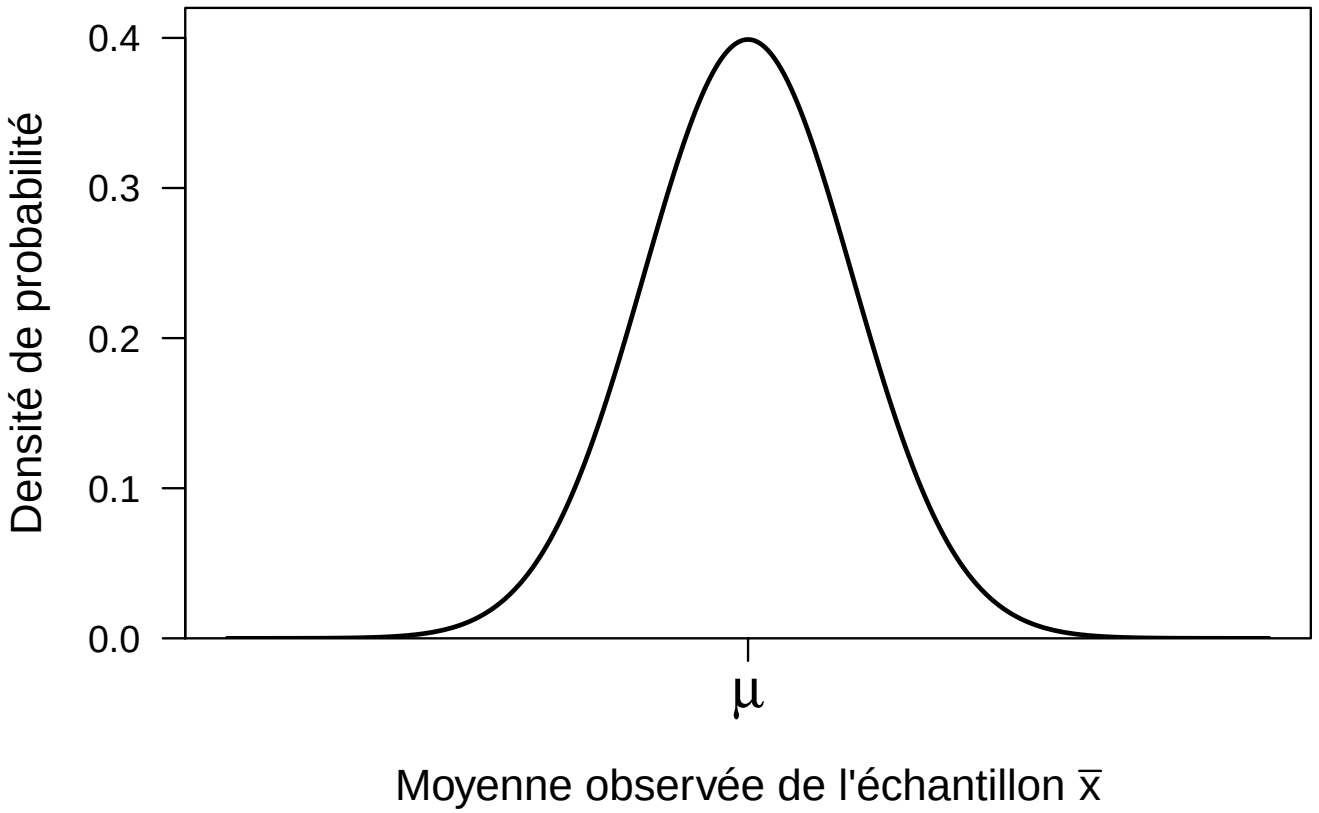
Pourquoi des tests ?

- Les différences observées dans une manip entre l'attendu et l'observé peuvent être dues au hasard,
- . . . qui est particulièrement important en biologie de par la variabilité intrinsèque du vivant.
- On ne peut donc pas conclure sur la base seule de la différence observée.
- On doit appliquer une procédure statistique pour évaluer les chances que le hasard explique la différence observée.

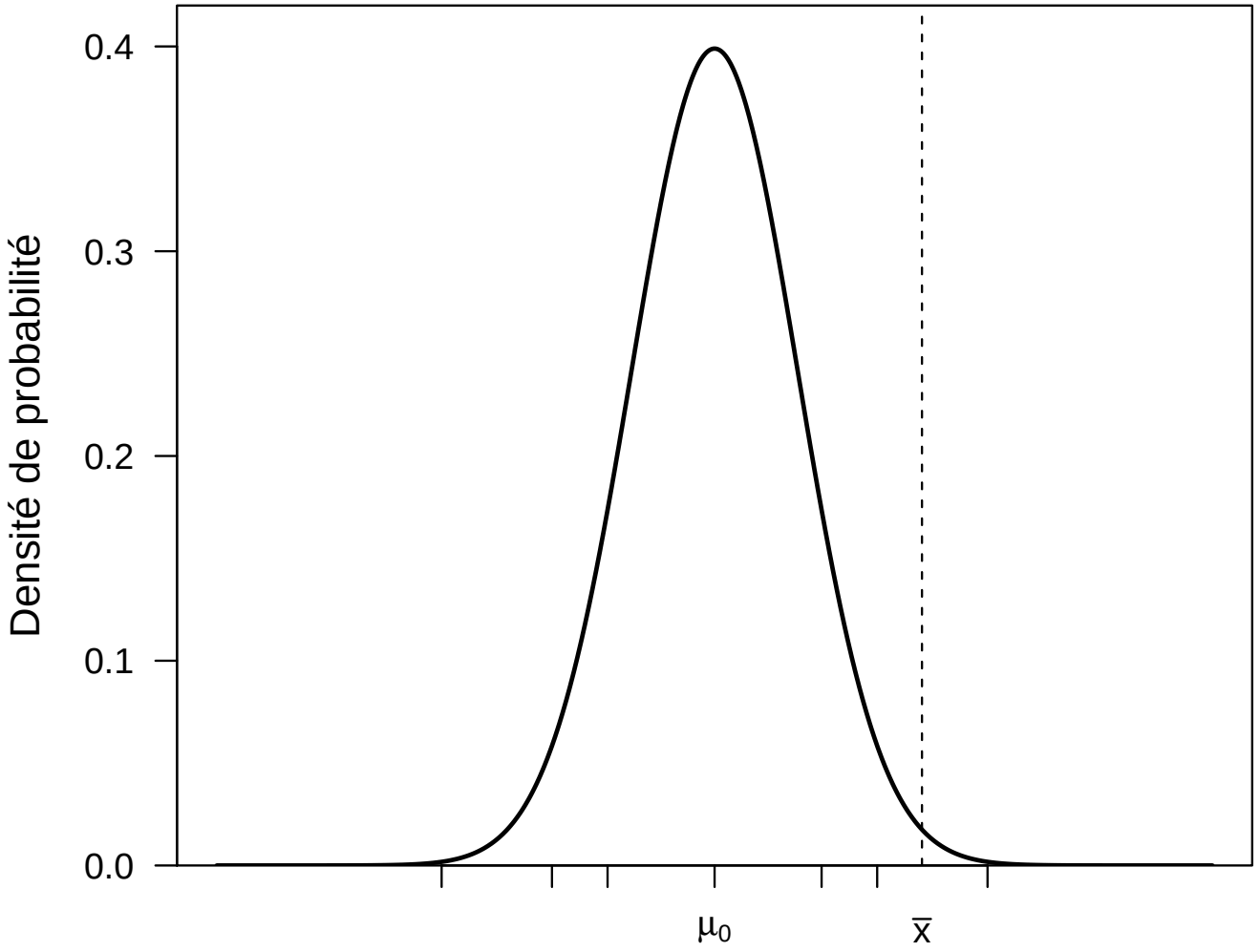
Loi de la moyenne de n v.a. et TCL



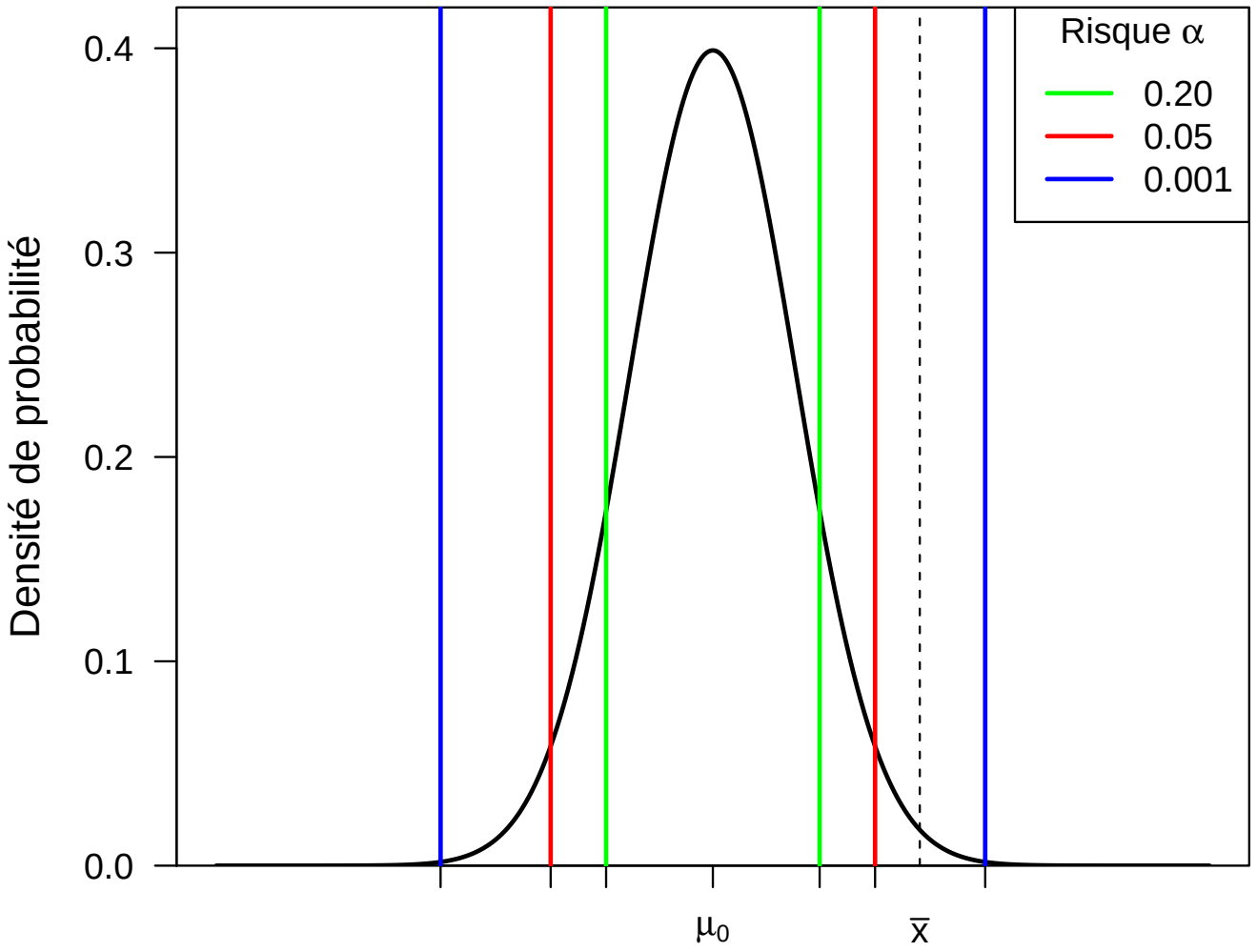
Distribution d'échantillonnage



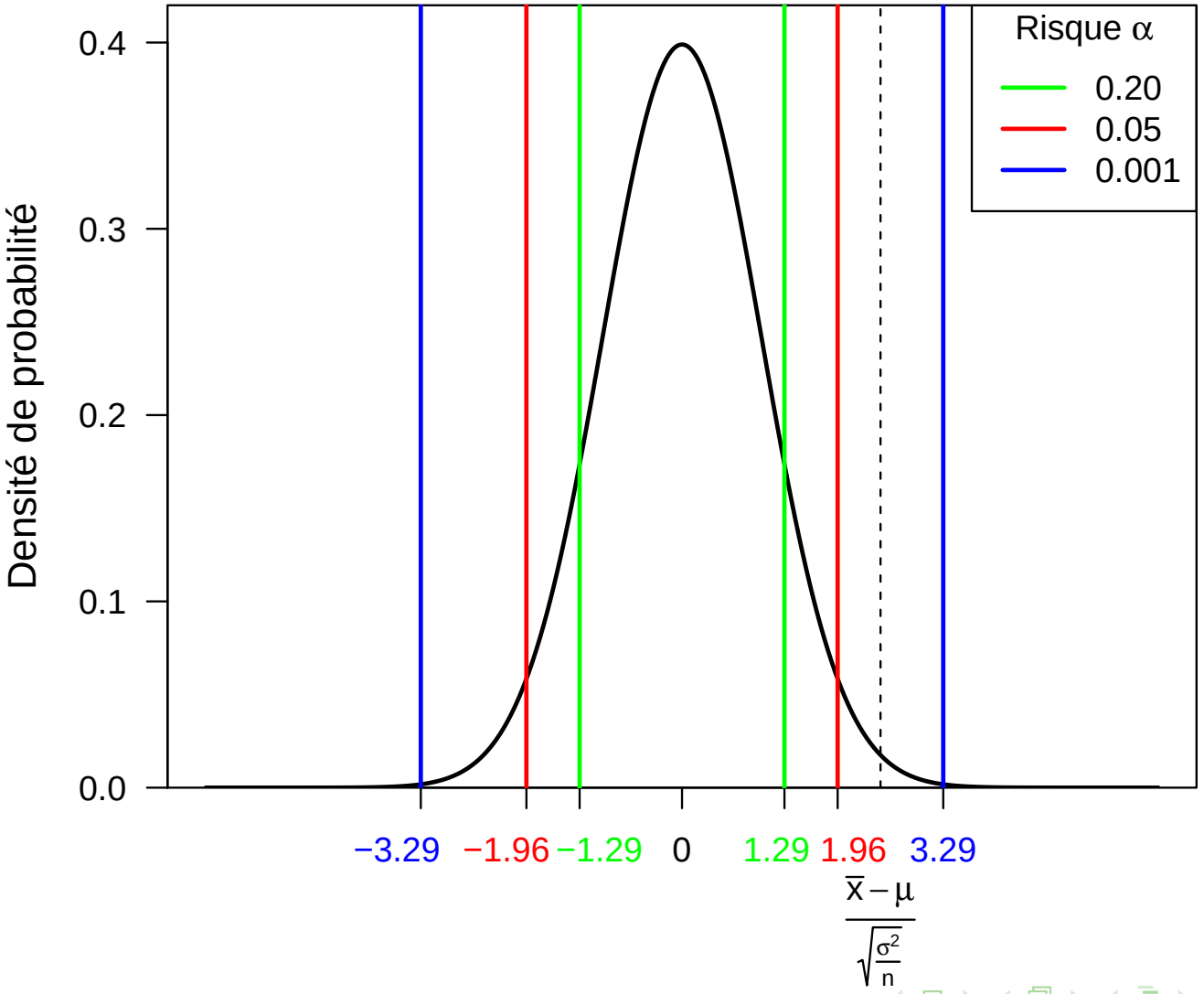
Distribution d'échantillonnage et moyenne observée



Distribution d'échantillonnage et moyenne observée



Distribution d'échantillonnage centrée réduite



Un test, deux hypothèses

On note H_0 l'hypothèse nulle et H_1 l'hypothèse alternative :

H_0 La variable d'intérêt est égale à une valeur théorique précise : $\mu = \mu_0$
Sous-entendu : la différence entre valeur observée et valeur théorique est due aux fluctuations d'échantillonnage.

H_1 La variable d'intérêt ne respecte pas H_0 : $\mu \neq \mu_0$

Un test d'hypothèse c'est une règle de décision qui permet, au vu des résultats d'une expérience, de trancher entre H_0 et H_1.

Logique des tests

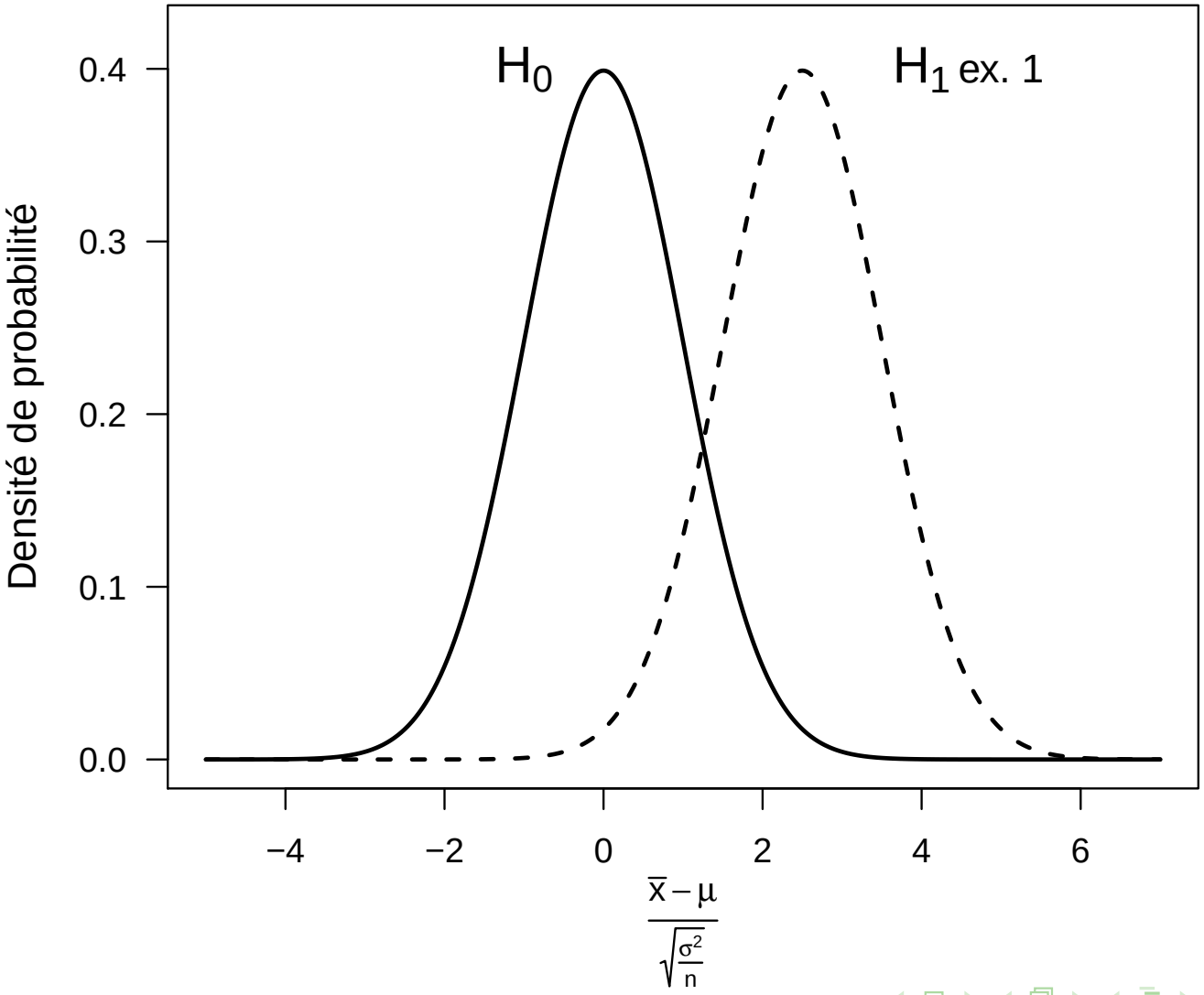
- Si H_0 est vraie, alors on peut calculer les chances que l'on observe $\bar{x}$ (ou toute autre statistique).
- Si la valeur observée est trop peu probable, avec un seuil α donné, on va conclure que notre hypothèse de raisonnement (H_0 vraie) est fausse, et rejeter H_0 pour accepter H_1,
- . . . ce qui sera par définition une erreur dans un pourcentage α des expériences.
- Si la valeur observée est "assez probable", on ne peut pas rejeter H_0. Ce qui ne veut pas dire que H_0 est vraie !

Conclusions possibles d'un test

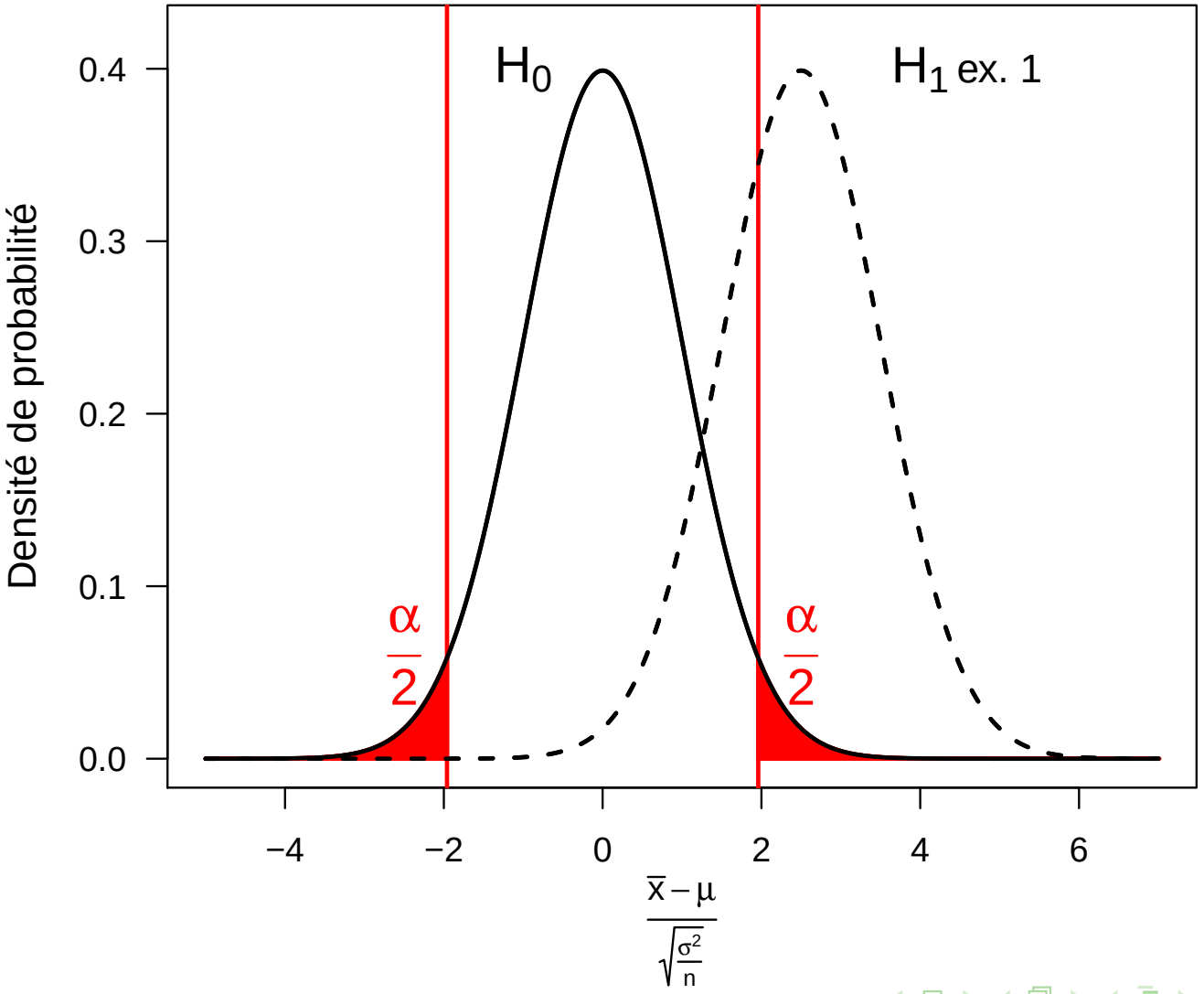
- On accepte H_1 avec un risque α, connu et choisi, de se tromper,
- ou bien on ne peut pas rejeter H_0 au vu des données.

Dans le deuxième cas, on conserve H_0 par défaut, par ce que c'est l'hypothèse que l'on favorise et que c'est ce que l'on concluerait sans avoir de données (rasoir d'Occam).

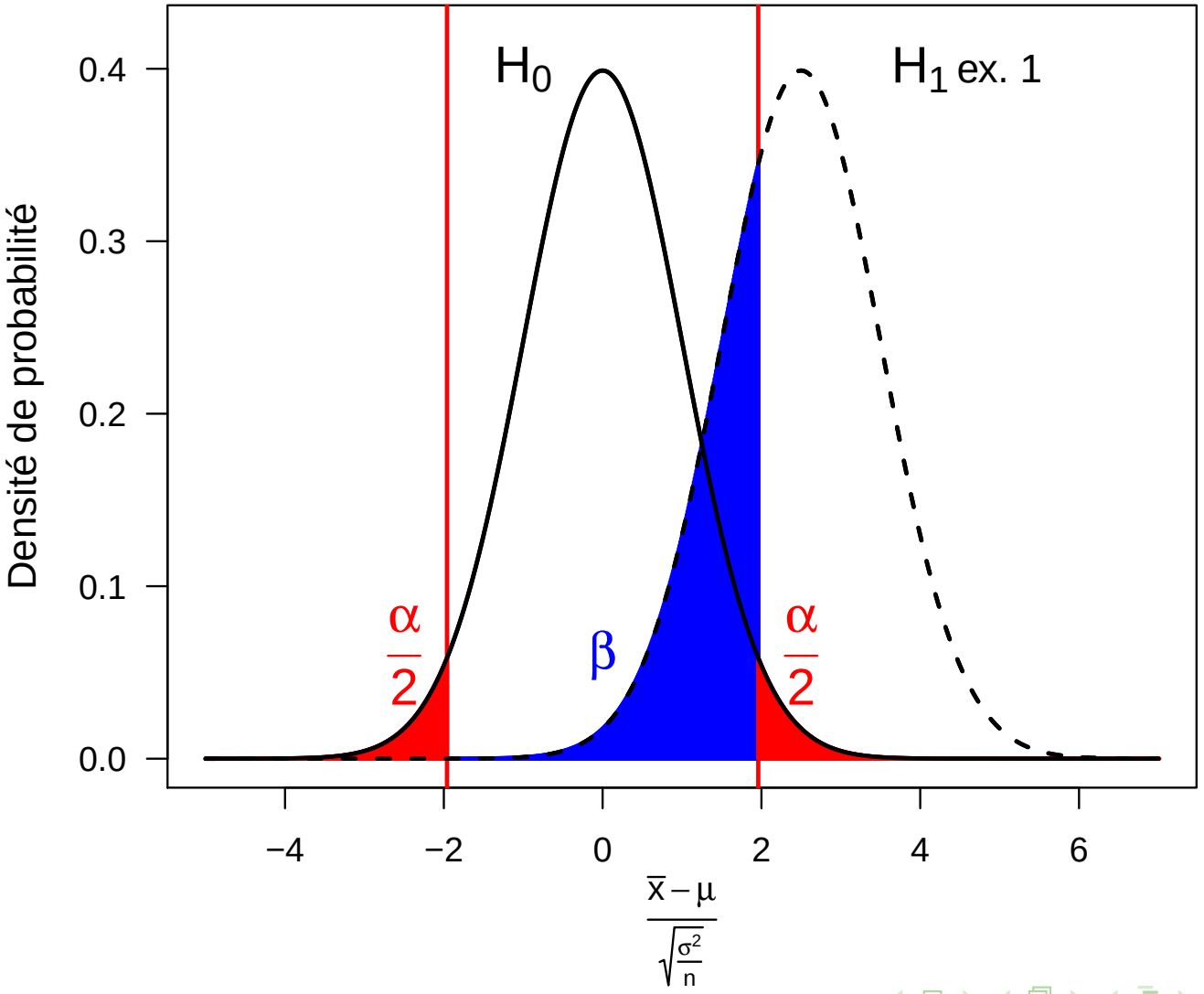
Risque de deuxième espèce



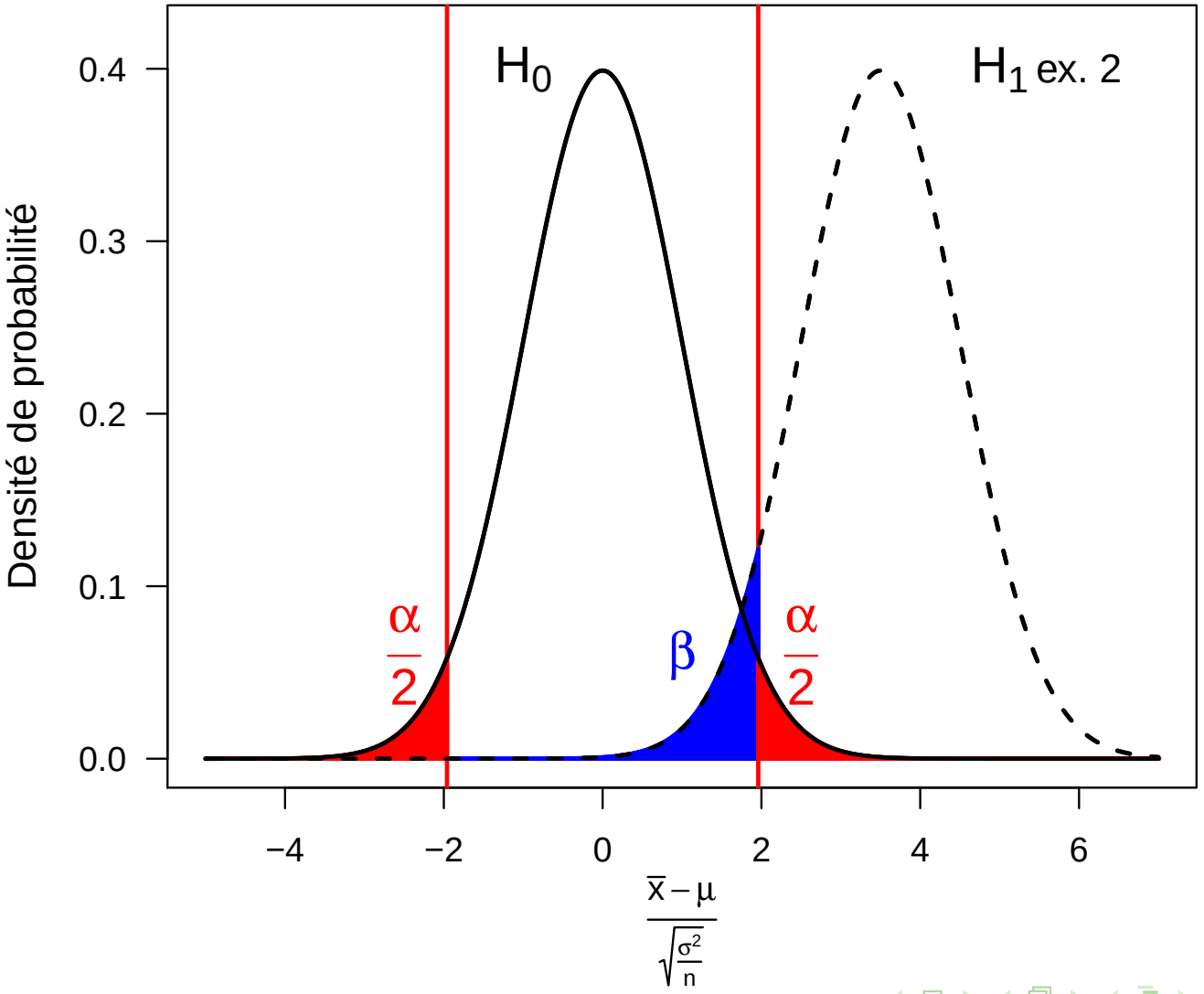
Risque de deuxième espèce



Risque de deuxième espèce



Risque de deuxième espèce : autre cas

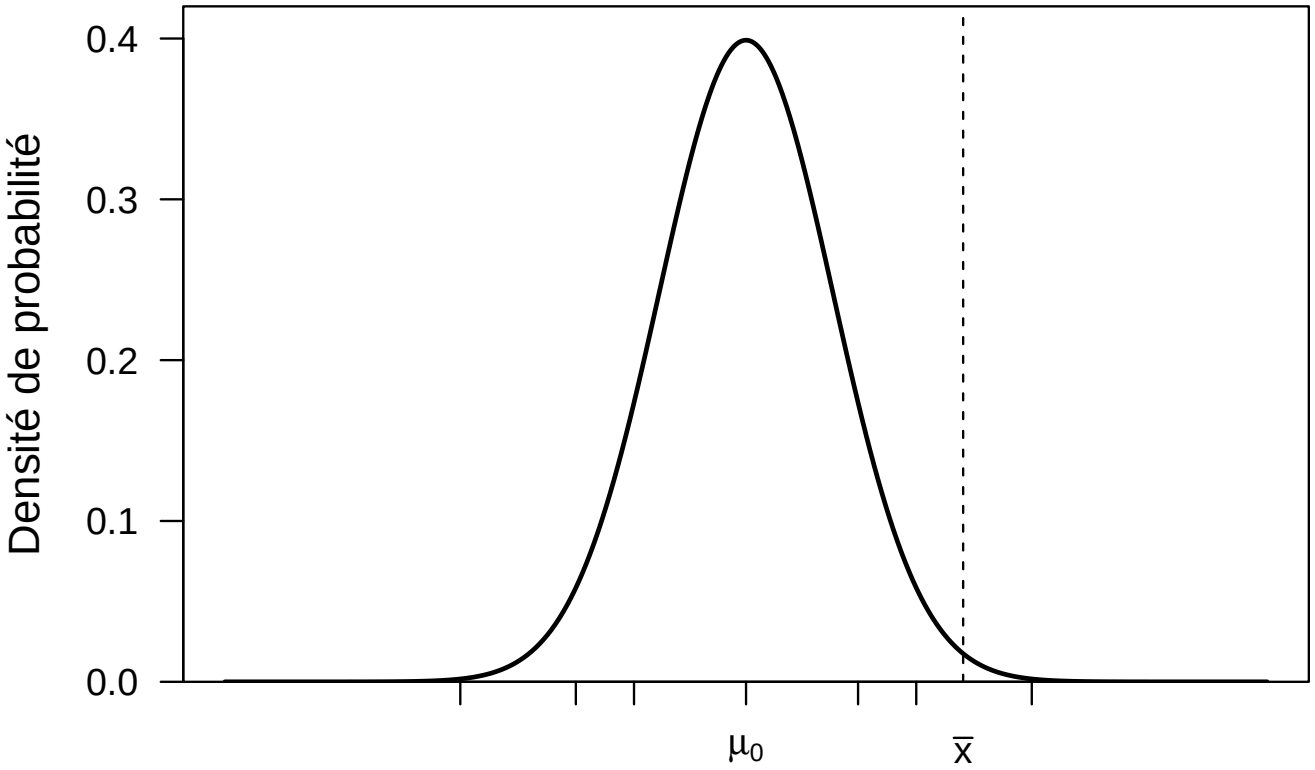


En résumé

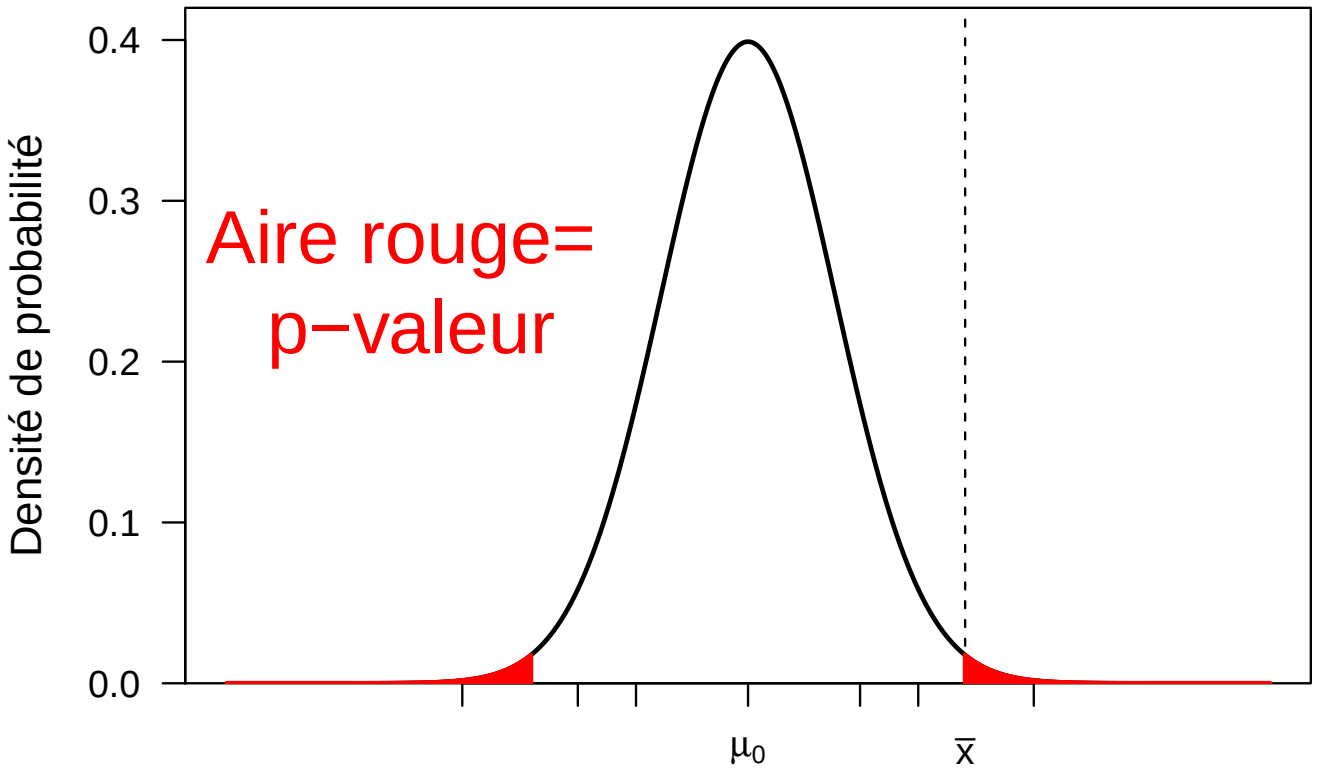
		Réalité inconnue	
		H_0	H_1
Choix	H_0	OK $1-\alpha$	Erreur de type II β
	H_1	Erreur de type I α	OK $1-\beta$

On appelle $1-\beta$ la puissance du test, i.e sa sensibilité. Un test puissant sera moins spécifique, et aura un plus grand α.
On a (presque) le même problème que pour une démarche expérimentale, avec choix entre sensibilité (peu de faux négatifs) et spécificité (peu de faux positifs). En stats on prend un petit α.

Les tests à l'anglaise : *p*-valeur



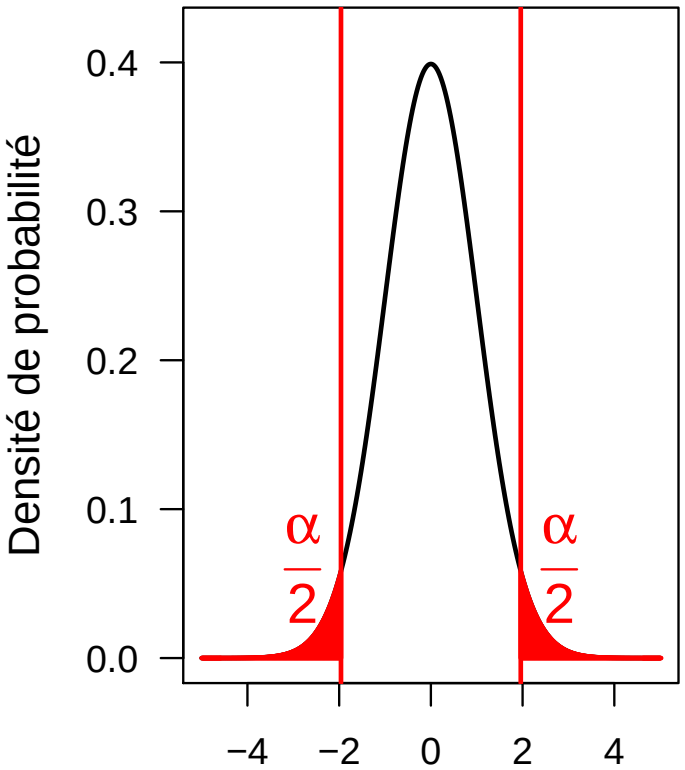
Les tests à l'anglaise : *p*-valeur



La $p-$valeur est la probabilité d'observer un résultat au moins aussi éloigné de la référence si celle-ci est vraie (si H_0 est vraie).

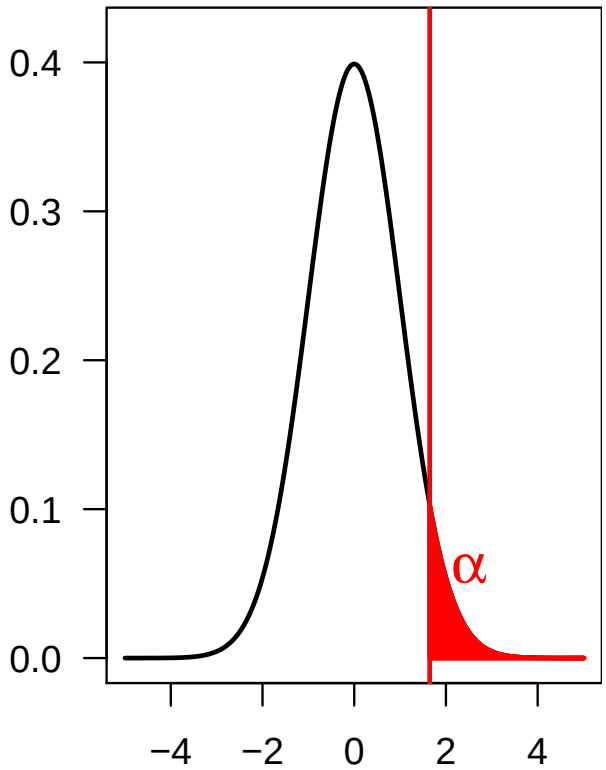
Test unilatéral

$H_1 : \mu \neq \mu_0$



$\varepsilon_{0.05} = 1.96$

$H_1 : \mu > \mu_0$



$\varepsilon_{0.1} = 1.645$

Attention à la latéralité

- Un test unilatéral ne vérifie pas une différence, mais une supériorité/infériorité.
- Pour les mêmes données, un test unilatéral aura une p-valeur deux fois plus faible que le test bilatéral correspondant.
- Un test unilatéral ne peut être pratiqué que si, avant de voir les données, on pouvait prédire l'effet que l'on chercherait à tester et sa direction (exple : médicament).
- Pas de test unilatéral a posteriori !

Un exemple de démarche statistique

Supposons que l'on veuille montrer que les garçons et les filles n'ont pas la même consommation de cigarettes.

- On pose alors l'hypothèse nulle H_0 : il n'y a pas de différence entre les deux échantillons
- Et l'hypothèse alternative H_1 : les deux sexes ne fument pas la même quantité de cigarettes.
- On fait le test.

Les effets de taille

On réalise le test mentionné précedemment, sur de fausses données, pour comprendre un phénomène intéressant.
Commençons à $n = 1000$

```
fum_g <- rnorm(mean=3.35,sd=3.12,n=1000)
fum_f <- rnorm(mean=2.54,sd=2.34,n=1000)
t.test(fum_g,fum_f)
```

```
Welch Two Sample t-test

data:  fum_g and fum_f
t = 6.839, df = 1873.664, p-value = 1.076e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.590692 1.065702
sample estimates:
mean of x mean of y
 3.316485  2.488288
```

Les effets de taille

$n = 500$

```
 fum_g <- rnorm(mean=3.35,sd=3.12,n=500)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=500)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 4.0052, df = 927.72, p-value = 6.692e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3603244 1.0526973
sample estimates:
mean of x mean of y
 3.329484  2.622973
```

Les effets de taille

$n = 100$

```
 fum_g <- rnorm(mean=3.35,sd=3.12,n=100)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=100)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 1.6769, df = 171.073, p-value = 0.09539
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1165286  1.4321784
sample estimates:
mean of x mean of y
 3.399404  2.741579
```

Les effets de taille

Et si la différence est plus grande ?

```
 fum_g <- rnorm(mean=4.35,sd=3.12,n=100)
 fum_f <- rnorm(mean=2.54,sd=2.34,n=100)
 t.test(fum_g,fum_f)

Welch Two Sample t-test

data:  fum_g and fum_f
t = 5.5283, df = 178.401, p-value = 1.133e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.465076 3.091593
sample estimates:
mean of x mean of y
 4.489523  2.211188
```

Deux effets principaux jouent sur le résultat du test

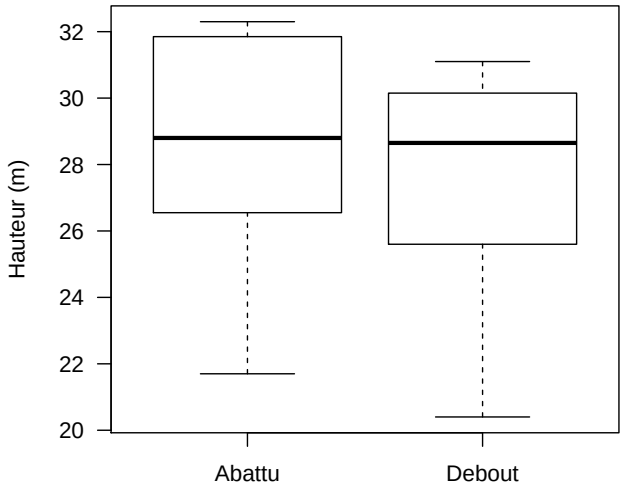
- La taille de l'échantillon : plus celui-ci est grand, mieux on pourra détecter un écart à H_0 (aussi faible soit-il). Attention, avec trop de données on peut conclure que H_0 est fausse, mais il est possible qu'elle soit fausse de très peu de choses...
- La taille de l'écart à H_0 (taille d'effet) : plus il est grand, plus il est détectable.

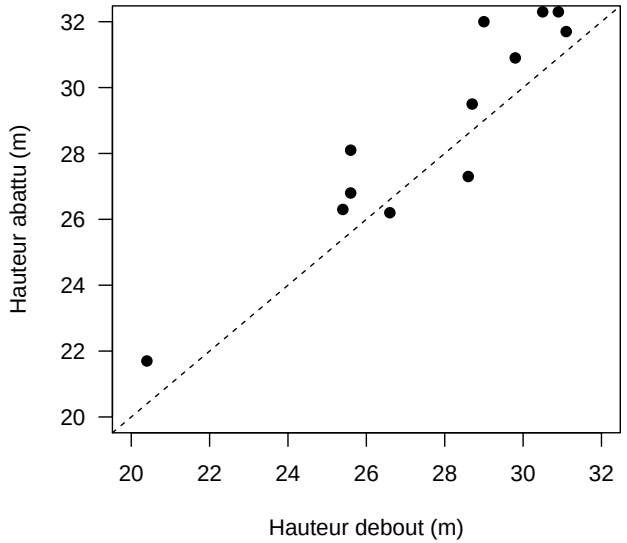
Réfléchir avant le design expérimental aux variables qui seront mesurées et aux variations que l'on cherche à détecter !

Un cas particulier : les données appariées

Ce sont des mesures de type avant/après, où il ne faut pas confondre la variabilité entre individus et la variabilité au sein d'un individu. Un exemple :

```
> debout = c(20.4,25.4,25.6,25.6,26.6,28.6,28.7,29.0,29.8,30.5,30.9,31.1)
> abattu = c(21.7,26.3,26.8,28.1,26.2,27.3,29.5,32.0,30.9,32.3,32.3,31.7)
```





Un cas particulier : les données appariées

Les tests "appariés" et "non appariés" ne vérifient pas la même chose, et donnent des résultats très différents :

```
> t.test(debout,abattu)

Welch Two Sample t-test

data:  debout and abattu
t = -0.8246, df = 21.937, p-value = 0.4185
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.779124  1.629124
sample estimates:
mean of x mean of y
 27.68333  28.75833

> t.test(debout, abattu,paired=T)

Paired t-test

data:  debout and abattu
t = -3.2343, df = 11, p-value = 0.007954
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.8065536 -0.3434464
sample estimates:
mean of the differences
                 -1.075
```

Quel test faire ?

Cas simple : on a deux échantillons à comparer.

- On a souvent le choix entre des tests paramétriques et des tests non paramétriques.
- Les tests paramétriques ont des conditions d'application restrictives (normalité des données) mais sont plus puissants, au sens vu plus haut.
- Les tests non paramétriques peuvent s'appliquer plus souvent mais sont moins puissants.

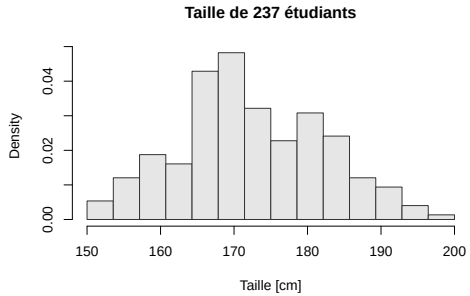
Paramétrique ou non-paramétrique ?

- Si les données sont distribuées normalement ou sont en grande quantité (pour chaque test), on peut souvent appliquer un test paramétrique. Attention, chaque test a ses conditions d'applications.
- Si les données sont visiblement anormales, on peut les transformer (log, racine) ou employer un test non-paramétrique.
- Si les données sont en très faible quantité, il est recommandé d'employer des tests non-paramétriques – car les conditions d'application du test paramétrique sont difficiles à vérifier.

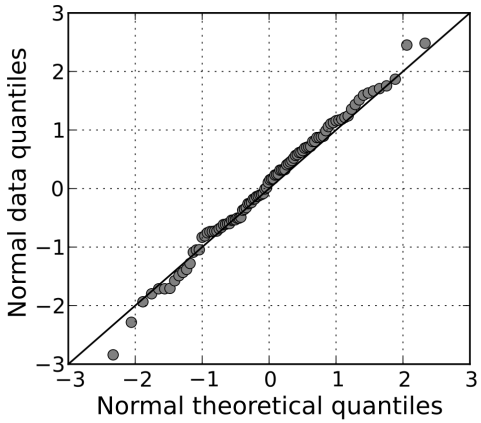
Une condition particulière : la normalité

- Regarder la distribution des données !
- Graphe quantile-quantile (qqplot)
- Si on a assez de données pour que le test soit puissant, faire un test (Shapiro).

Ex1 :



Ex 2 :



Une courte liste de cas classiques

Paramétrique	Non-Paramétrique
t de Student	Wilcoxon Mann-Whitney
t de Student apparié	Wilcoxon apparié
χ^2	Test exact de Fisher
Corrélation r de Pearson	ρ de Spearman

Voir une liste plus complète par ex. sur Wikipédia.

Et si on a plus de deux échantillons ?

- On commence toujours par un test global, pour déterminer si tous les échantillons pourraient venir de la même population.
- S'il y a des différences, on passe à des méthodes locales (contrastes, etc. . .) ou à des tests 2 à 2 (Attention !)
- En général, les procédures sont plus complexes et il est nécessaire de passer par la case "statisticien" avant de conclure.

Une très courte liste de cas classiques

Paramétrique	Non-Paramétrique
ANOVA	Kruskal-Wallis

Au-delà de ce cas simple, il faut consulter. Un statisticien. Ou prendre le temps de se documenter en détails.

Trop de tests tuent les tests...

- Si on fait de nombreux tests indépendants (effet de n gènes sur un phénotype, ou k substances...), on prend un risque (la p-valeur) à chaque fois que l'on conclut H_1.
- Quand on prend n fois un risque p de se tromper, on se trompe en moyenne np fois...
- Il faut donc toujours appliquer dans ces situations des corrections de tests multiples.

Quelques corrections de tests multiples

Bonferroni : on multiplie toutes les p-valeurs par n avant de conclure. Facile mais peu puissant.

Benjamini and co, Dunn-Sidak : on corrige les p-valeurs en les augmentant avant de conclure. Mieux, mais nécessite un peu plus de travail.

False Discovery Rate : on calcule directement le taux de faux positifs.

En pratique : Bonferroni pour faire (trop simple), du temps, Internet et ces mots clés pour faire plus précis.

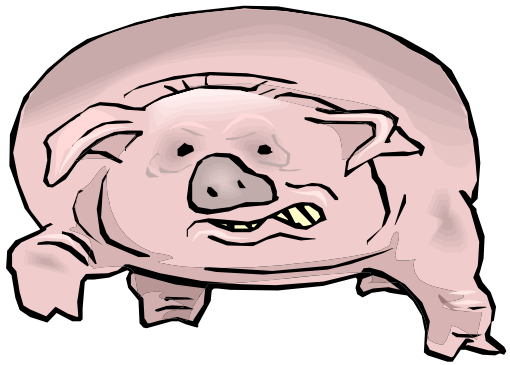
Table des matières

Extraits du cours de M-L. Delignette-Muller

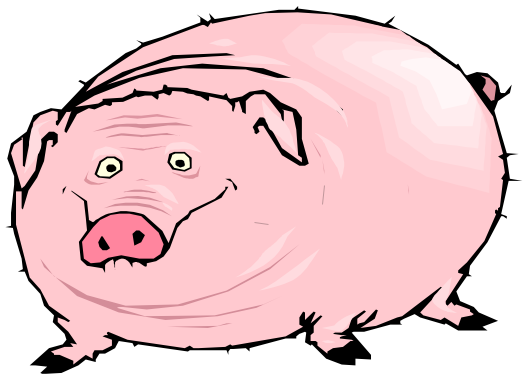
- Cours donné à VetSAgroSup
- Disponible à l'adresse
  `http://www3.vet-lyon.fr/ens/biostat/accueil.html`
- Très complet, avec de nombreux exemples d'analyse en  .

Exemple 1

avant traitement



après traitement



- baisse significative de la température rectale sur un échantillon de 40 porcs malades après 2 jours d'un traitement donné

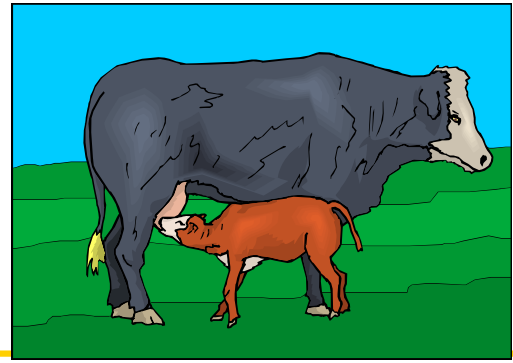
➔ le traitement induit une diminution significative de la température rectale ($p = 0.02$)

La même baisse de température aurait peut-être été observée sans traitement

Erreur : absence de groupe témoin

- On ne peut rien conclure d'une expérience sans groupe témoin
- **Nécessité de comparer le groupe traité à un groupe témoin** (traité avec un placebo ou traitement de référence suivant les objectifs)

Exemple 2



- comparaison de 2 protocoles opératoires sur des césariennes de vaches

en 1998 : 55 vaches opérées selon le protocole P1

en 1999 : 57 vaches opérées selon le protocole P2

- Fréquence de complications significativement plus faible avec le protocole P2 ($p < 0.001$)
- ➔ meilleure efficacité du protocole P2

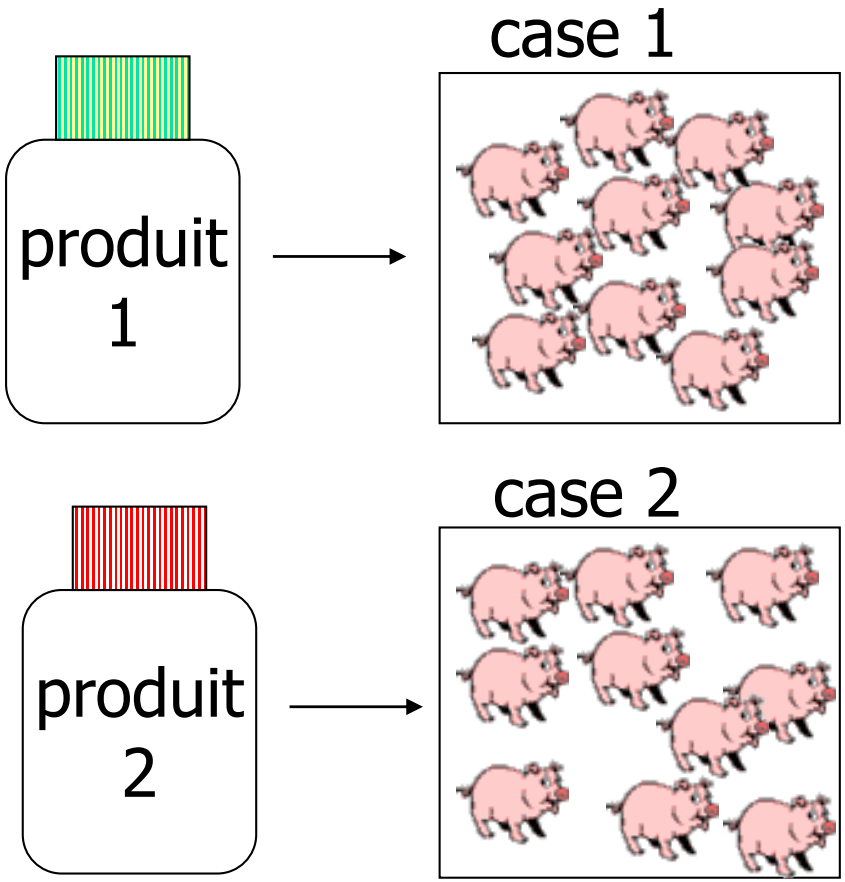
Les deux groupes ne sont pas comparables (effet « année » possible)

Erreur : absence de randomisation

➔Biais de sélection possible

- **Afin d'assurer la comparabilité initiale des groupes, leur constitution doit faire l'objet d'une randomisation**

Exemple 3



- Comparaison des gains de poids moyens sur 2 échantillons randomisés de 10 porcs : p = 0.003
- ➔ effets significativement différents des 2 produits

L'effet « case » est confondu avec l'effet « produit »

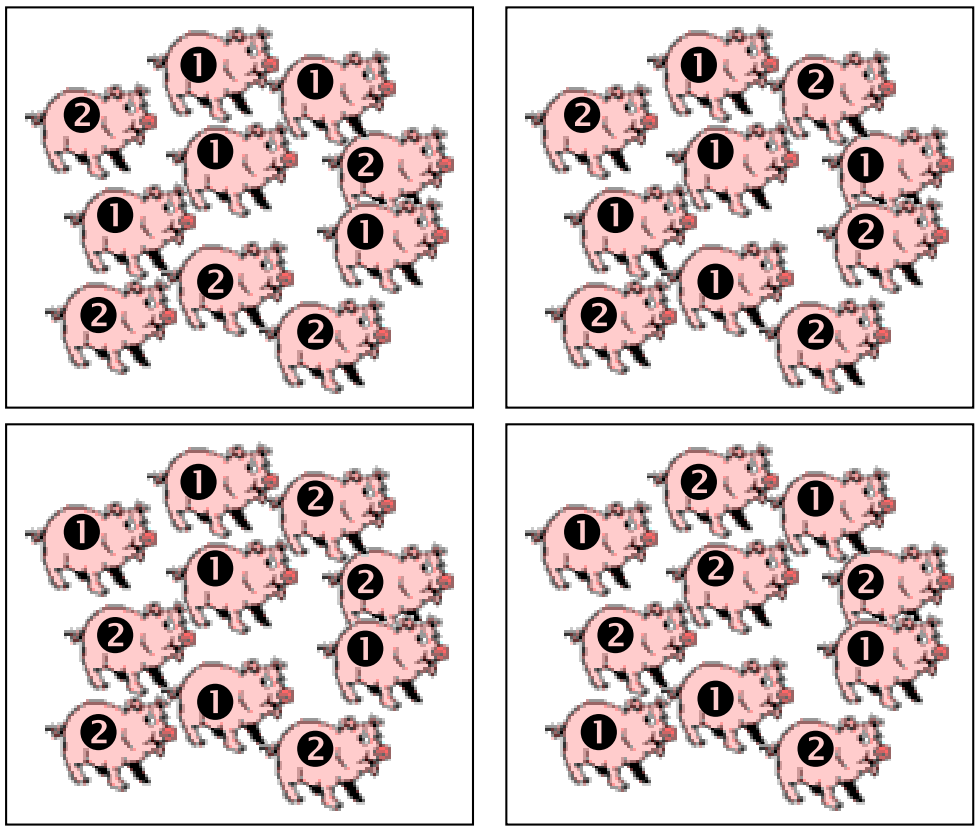
VetAgro Sup

Erreur : non comparabilité des lots en court d'étude

- Non prise en compte de possibles facteurs de confusion
- **Il est nécessaire de mettre en place un plan d'expérience contrôlant les facteurs concomitants** susceptibles d'avoir un effet sur la variable étudiée

VetAgro Sup

Plan d'expérience adapté à l'exemple 3



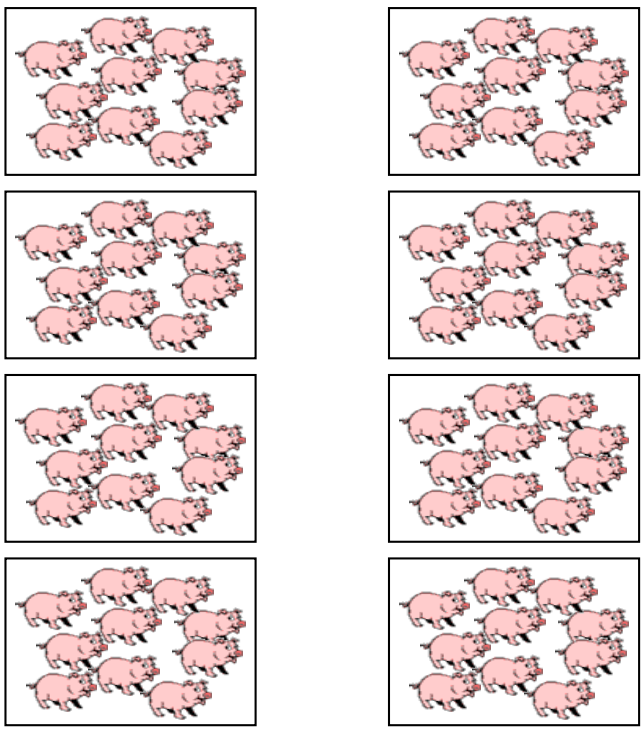
produit 1: ❶

produit 2: ❷

Administration des 2 produits randomisée au sein de chaque case

Exemple 3 bis

produit 1 produit 2



- Comparaison de 2 échantillons de 40 observations
- Attention, ici l'unité expérimentale est la case

➔ comparaison de 2 échantillons de 4 observations

VetAgro Sup

Erreur : mauvaise définition de l'unité expérimentale

- Quand le même traitement est donné à tous les animaux d'une même case (ou cage), l'unité expérimentale est la case (ou cage)
- Autre ex.: lorsque des mesures sont répétées dans le temps sur un même animal, l'unité expérimentale reste l'animal

Exemple 4

- comparaison de 2 traitements anti-parasitaires sur la croissance de bovins
- randomisation des 2 traitements sur 30 jeunes bovins

- Gains de poids à J30:

  lot1 (n=15): m1=25 kg

  σ1=19 kg

  lot2 (n=15): m2=18 kg

  σ2=17 kg
- Différence entre les 2 moyennes non significative ($p > 0.05$)

Les données ne mettent pas en évidence de différence entre les 2 traitements

VetAgro Sup

Erreur : manque de puissance

- Effectifs faibles, variabilité importante et différence attendue peu importante
- ➔ L'essai avait très peu de chance de conclure à une différence significative
- **Calcul de puissance *a priori* nécessaire**

  « quels effectifs sont nécessaires pour avoir une probabilité assez forte de mettre en évidence une différence d'intérêt clinique ou biologique par un test d'hypothèse au risque α=5%? »

VetAgro Sup

A retenir

- Un essai doit être **comparatif** avec des **groupes comparables** au départ et tout au long de l'essai
- La définition de **l'unité expérimentale** dépend du **plan d'expérience**
- Les **effectifs** doivent être **équilibrés** et si possible **calculés *a priori*** en vue d'atteindre la **puissance** souhaitée

Exemple 13

- Une étude comparant un traitement A à un placebo face au traitement d'une pathologie donnée reporte un taux de guérison significativement plus élevé pour le traitement A (p=0.02).
- Une autre étude réalisée indépendamment sur la même pathologie compare les résultats obtenus avec un traitement B et le même placebo et reporte un taux de guérison significativement plus élevé avec le traitement B (p=0.001).

➔ On peut en conclure que le traitement B est plus efficace que le traitement A.

VetAgro Sup

Erreur : comparaison de valeurs de p

- Une valeur de **p plus faible n'implique pas** forcément une **différence plus importante**

  p dépend aussi
  - des effectifs
  - de la variabilité intra-groupe
- **On ne doit jamais comparer des valeurs de p** ni entre essais, ni même au sein d'un même essai

À retenir

- Présentez toutes vos données, de la manière la plus informative et la plus lisible possible
- Prévoyez à l'avance le type d'analyse qui sera nécessaire sur vos données pour en tirer le meilleur parti (tests, puissance)
- Dès que ca devient un peu compliqué, passez discuter avec un biostatisticien. . .
- . . . avant de commencer !