Des Statistiques en Kit …

Prof. Catherine Legrand

Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA-IMMAQ), UCL

Catherine.legrand@uclouvain.be

www.uclouvain.be/isba

www.uclouvain.be/lsba

Congrès SBPM – Aout 2017

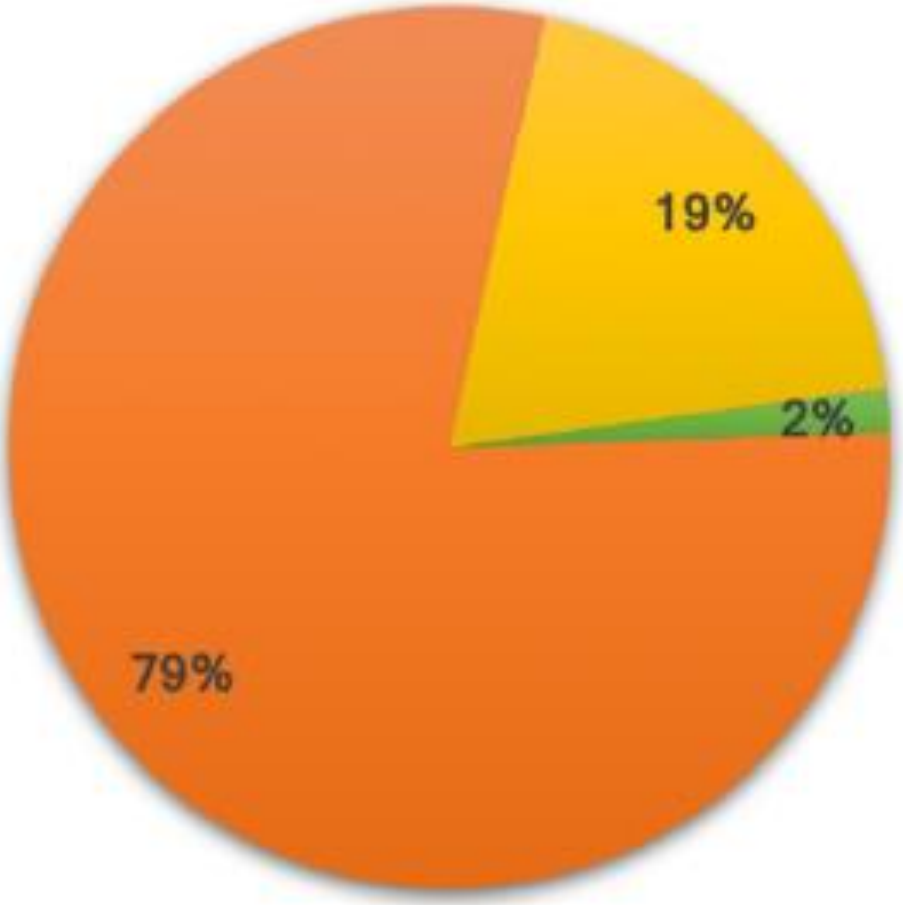
Avant de commencer ….

Thèse de Doctorat en Statistique (UHAsselt)

- Athénée Royale Riva-Bella (BA)
- "Licence" en Sciences Mathématiques, orientation Statistique - ULB
- European Organisation for Research and Treatment of Cancer (EORTC)
- Merck Sharp & Dhome (MSD)
- Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA) - UCL

Ce cher graphique en camembert…

Ce cher graphique en camembert…



Ce cher graphique en camembert…

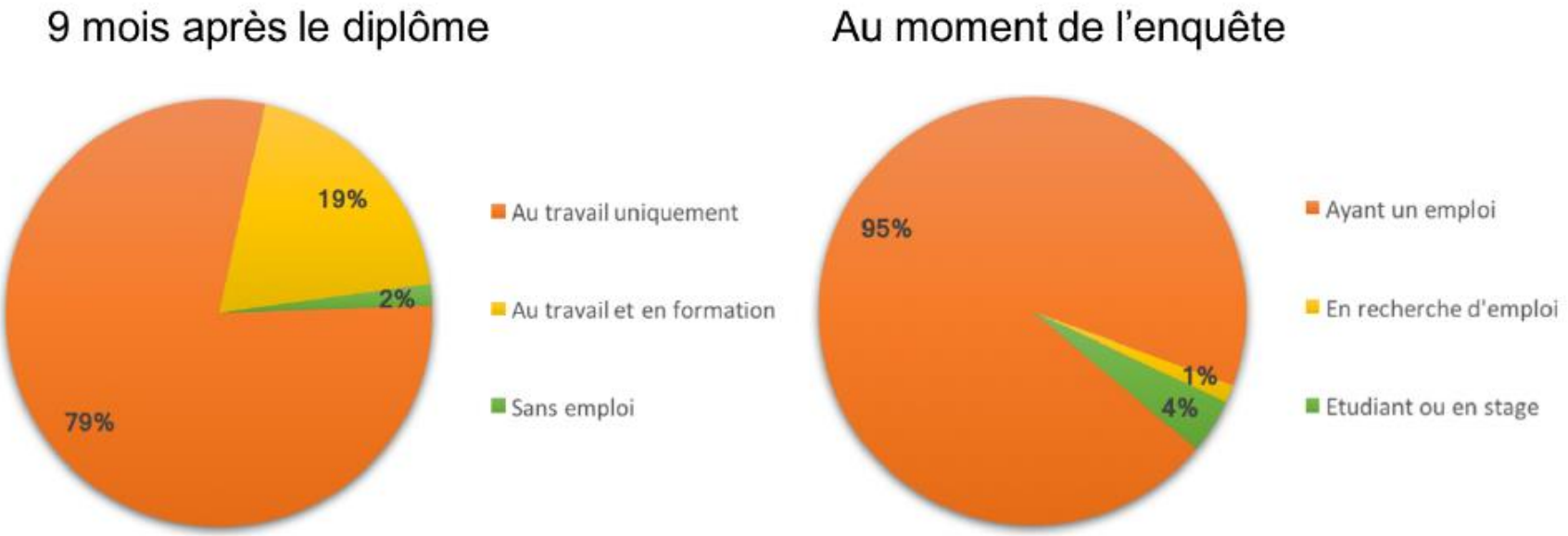
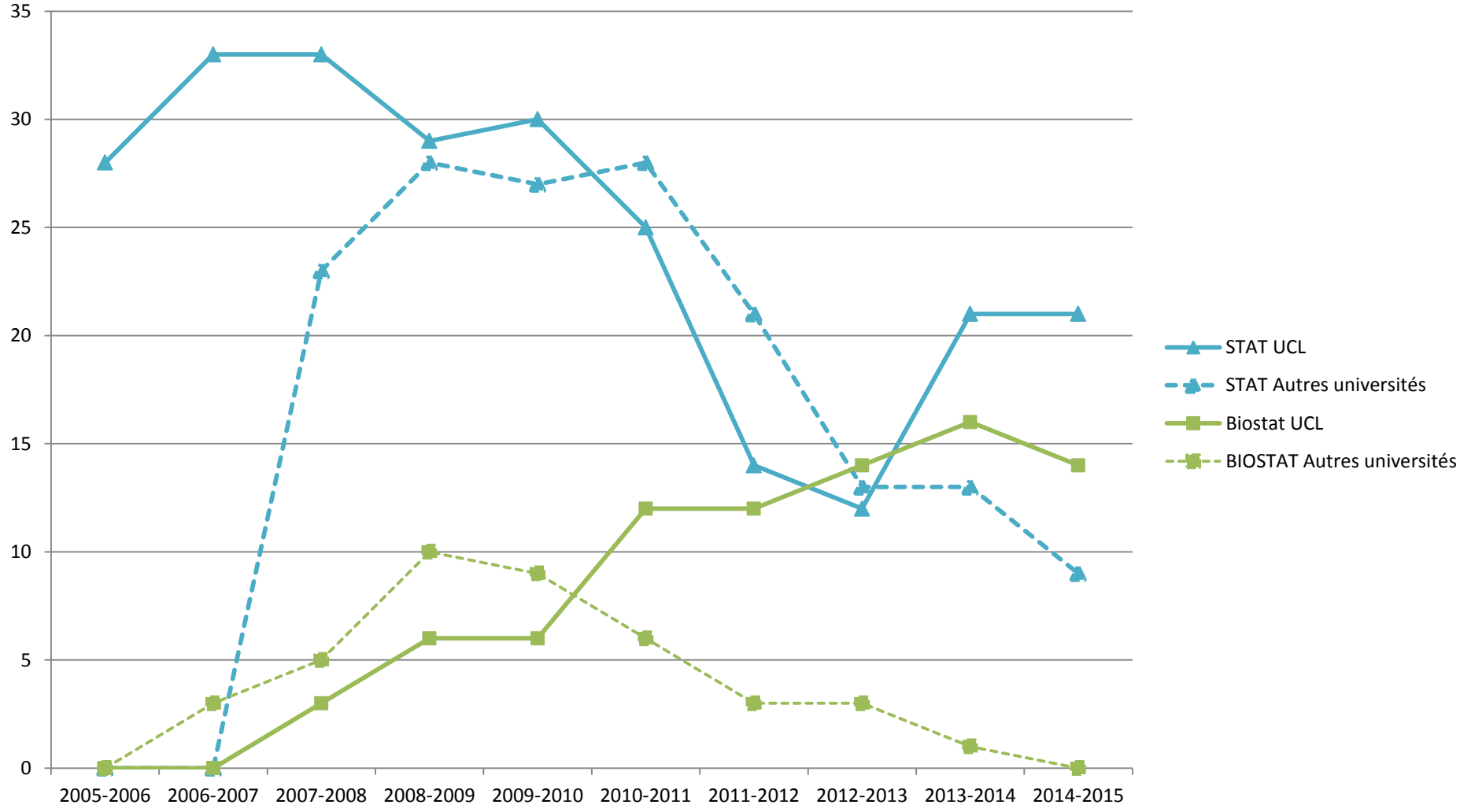


FIGURE 4.5 – Insertion socio-professionnelle des diplômés de Master en sciences actuarielles, en statistiques et biostatistiques entre 2011 et 2016, à gauche, 9 mois après leur diplômation, à droite, au moment de l'enquête

Un autre graphique …



Statisticien ??

- Nature orpheline du master en statistique
- Manque d'attrait pour les études « mathématiques »
- Méconnaissance du métier de statisticien
- Matière en générale non-enseignée en secondaire

Des statistiques en Kit ??

- Kit à destination des enseignants du secondaire
- Fournit tous les outils nécessaires pour enseigner un concept de stat/proba via une application « ludique »
- Leçon à donner en 2-3 heures maximum

UCL Université catholique de Louvain

ISBA

https://e-mediasciences.uclouvain.be/

UCL Université catholique de Louvain

ISBA

scienceinfuse

Home | contact

e-mediasciences

UCL Université catholique de Louvain

Biologie | Chimie | Physique | Maths | Géographie

matériel empruntable

recherche

s'abonner à l'e-news

e-news précédente

Réalisation : www.afd.be | Back office

Matériel empruntable

01 Kits

- Biologie
- Chimie
- Physique
- Maths
  - Fractions et polygones
  - Introduction à la cryptographie
  - Les polyèdres et la relation d'Euler

https://e-mediasciences.uclouvain.be/

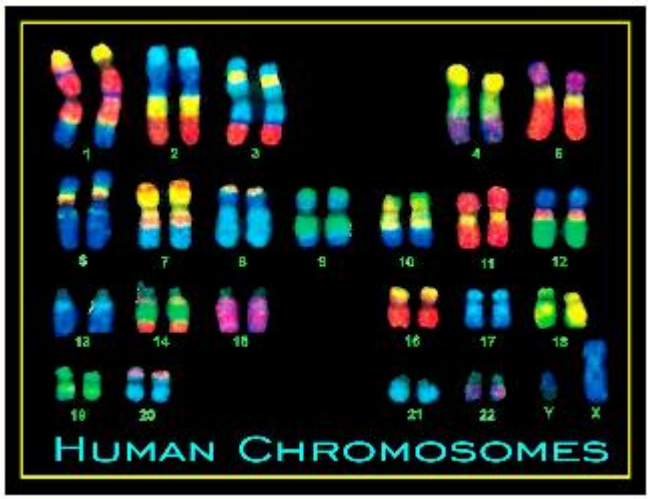

Coming Soon
Kit Proba/Stat
Equilibre de Hardy-Weinberg

Description générale

- **Contexte:**
  - Epidémiologie génomique
  - Equilibre de Hardy-Weinberg
- **Notions abordées:**
  - Probabilité: notion d'aléatoire
  - Statistique: notion de test statistique (et donc d'échantillon et de population)
- **Matériel nécessaire:**
  - Quelques feuilles de papier
  - Une classe d'étudiants suffisamment grande
- **Prérequis**
  - Très peu de notions mathématiques, mais un peu de « logique »

Quelques notions préalables

- **Génétique:** science de la transmission des caractères héréditaires dans une population d'êtres vivants
- **Génome:** ensemble du matériel génétique d'un organisme vivant, présent dans les cellules sous forme de longues molécules d'ADN, nommée **chromosomes**



Etres humains:
2 x 22 chromosomes
+ 2 chromosomes sexuels
(XX pour la femme et XY pour l'homme)

→ La transmission du génome des parents aux enfants se manifeste par le fait que les cellules humaines contiennent 23 chromosomes provenant de la mère et 23 provenant du père.

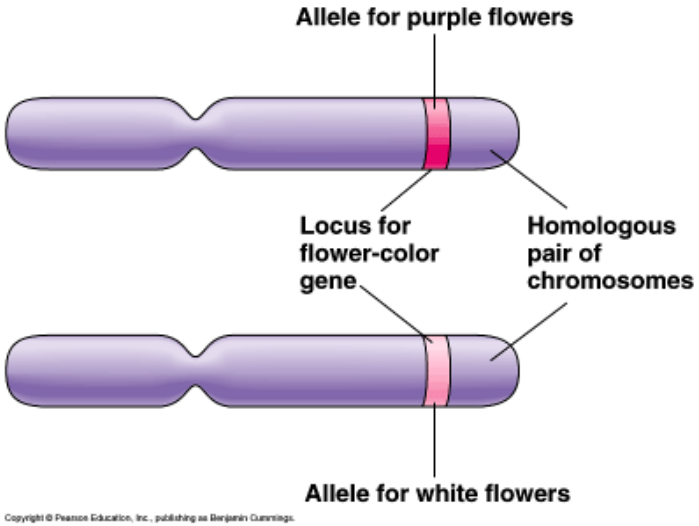
http://igbiologyy.blogspot.be/2014/03/chromosomes-dna-genes-and-alleles.html

Chez les humains:
2 x 22 chromosomes + 2 chromosomes sexuels

- **Chromosomes**: portent les gènes, qui a leur tour représentent, dans un sens, l'unité d'information génétique.

  →Les chromosomes dont nous avons deux copies ainsi que les gènes qui s'y trouvent sont dit **autosomes**.

- A l'exception des gènes se trouvant sur les chromosomes sexuels, nous possédons deux copies de chaque gène.



→ Une copie d'un gène est appelée un **allèle**.
→ Chaque individu possède donc deux allèles de chaque gène et ce couple de gènes détermine son **génotype**.

http://igbiologyy.blogspot.be/2014/03/chromosomes-dna-genes-and-alleles.html

Le contexte

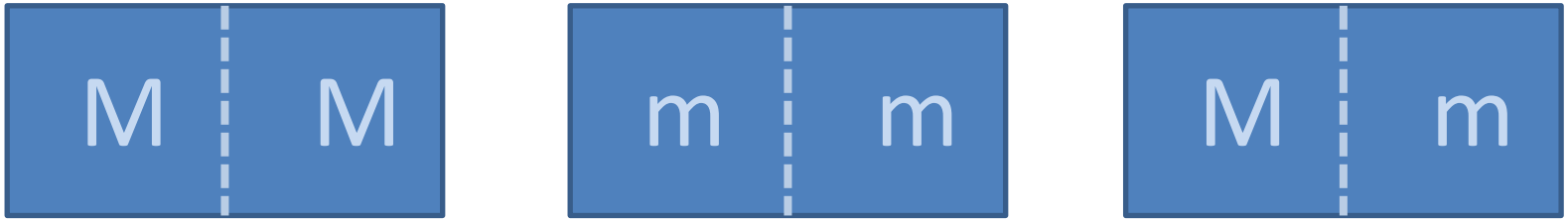
- On suppose que les étudiants de la classe constitue un échantillon de notre population d'intérêt
- On suppose que l'on s'intéresse à un gêne M/m qui code le fait d'être fort en math [M] ou pas [m] (*).
  - On suppose évidemment que l'allèle M (fort en math) est dominant.
  - On a donc 3 génotypes:

MM	Fort en math
Mm	Fort en math
mm	Nul en math

(*) Il s'agit évidemment d'un exemple totalement fictif !!

Première partie

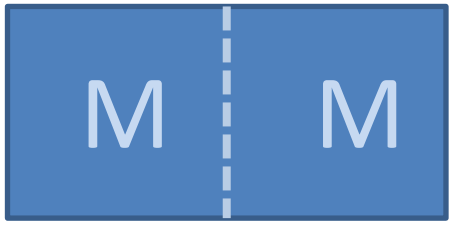
- Chaque étudiant inscrit (et cache) son génotype sur une feuille de papier, et dissocie celui-ci en deux allèles, i.e.,

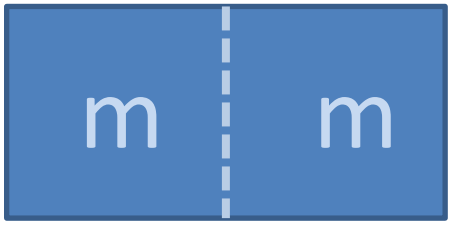


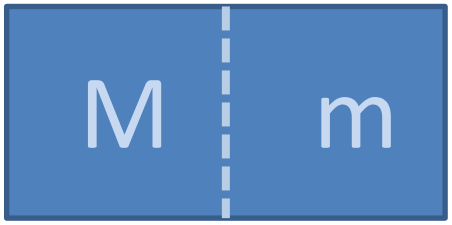
QUESTION 1: Si le professeur connait le nombre d ’élèves ayant chaque génotype, peut-il retrouver le nombre de chaque allèle dans la classe ?

FACILE !!!

QUESTION 1: Si le professeur connait le nombre d ’élève ayant chaque génotype, peut-il retrouver le nombre de chaque allèle dans la classe ?







EXAMPLE (30 élèves)		
8 élèves	10 élèves	12 élèves
→ 16 allèles M → 0 allèles m	→ 0 allèles M → 20 allèles m	→ 12 allèles M → 12 allèles m
→ 16+12 = 28 allèles M → 20 + 12 = 32 allèles m		

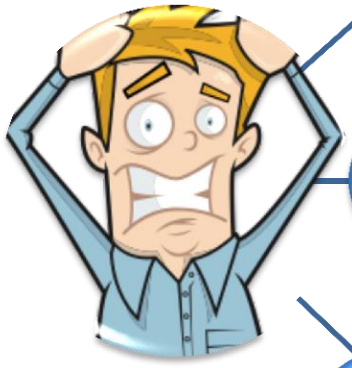
⌦ Vérifier la réponse dans la classe!!

- Le professeur crée des couples au hasard (sans connaitre le génotype) et pour chaque couple, crée un enfant en prenant au hasard un allèle à chaque parent

M	M	m	m	M	m

QUESTION 2: Si le professeur connait le nombre de parents ayant chaque génotype, peut-il retrouver le nombre de chaque allèle chez les enfants ?

En théorie, OUI!!



P_{MM}

P_{Mm}

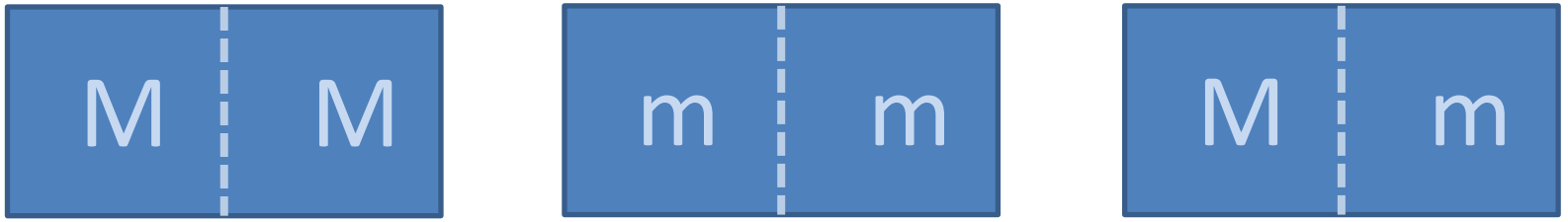
P_{mm}

- M • 100% de chance
- m • 0% de chance
- M • 50% de chance
- m • 50% de chance
- M • 0% de chance
- m • 1000% de chance

$$P_M = P_{MM} + 0.5P_{Mm}$$
$$P_m = P_{mm} + 0.5P_{mm}$$

QUESTION 2: Si le professeur connait le nombre de parents ayant chaque génotype, peut-il retrouver le nombre de chaque allèle chez les enfants ?

$$P_M = P_{MM} + 0.5P_{Mm}$$
$$P_m = P_{mm} + 0.5P_{mm}$$



EXAMPLE (30 élèves)		
8 élèves MM	10 élèves mm	12 élèves Mm
$P_{MM} = 8/30 = 0.27$	$P_{mm} = 10/30 = 0.33$	$P_{Mm} = 12/30 = 0.40$
$\rightarrow P_M = 8/30 + 0.5 \times 12/30 = 0.47$ $\rightarrow P_m = 10/30 + 0.5 \times 12/30 = 0.53$		

EXAMPLE (30 élèves)		
8 élèves MM	10 élèves mm	12 élèves Mm
$P_{MM} = 8/30 = 0.27$	$P_{mm} = 10/30 = 0.33$	$P_{Mm} = 12/30 = 0.40$
$\rightarrow P_M = 8/30 + 0.5 \times 12/30 = 0.47$ $\rightarrow P_m = 10/30 + 0.5 \times 12/30 = 0.53$		

→ Parmi les 15 enfants, on s'attend a retrouver:

- 0.47 x 15 x 2 = 14 allèles M
- 0.53 x 15 x 2 = 16 allèles m

☒ Comparer avec les résultats dans la classe!!

☒ Si on a le temps, refaire un grand nombre de fois l'expérience

Deuxième partie

- Le professeur a un nombre connu d'allèles de chaque type, et en distribue 2 au hasard et sans les regarder à chaque étudiant; l'étudiant connait ainsi son génotype.

QUESTION 3: Si le professeur connait le nombre d 'allèle de chaque type des élèves, peut-il en déduire la proportion de génotype de ceux-ci?

En théorie, OUI!!
… mais avec des hypothèses supplémentaires

UCL Université catholique de Louvain

ISBA

Equilibre de Hardy-Weinberg

= Ensemble d'hypothèses dont les principales sont:

- Unions aléatoires
- Fertilité normale (pas d'avantage d'un génotype)
- …

***Théorème*: Si toutes les hypothèses de l'équilibre de Hardy-Weinberg sont respectées alors on peut montrer que**

$$P_{MM} = (P_M)^2$$

$$P_{Mm} = 2 \times P_M \times P_m$$

$$P_{mm} = (P_m)^2$$

QUESTION 3: Si le professeur connait le nombre d ’allèle de chaque type des élèves, peut-il en déduire la proportion de génotype de ceux-ci?

EX: 28 x M 32 x m

$P_{MM} = (P_M)^2$

$P_{Mm} = 2 \times P_M \times P_m$

$P_{mm} = (P_m)^2$

M M m m M m

EXAMPLE (30 élèves)		
$P_{MM} = (28/60)^2 = 0.22$	$P_{mm} = (32/60)^2 = 0.28$	$P_{Mm}=2x28/60x32/60 = 0.50$
→ 0.22x30 = 6.6 élèves MM	→ 0.28x30 = 8.4 élèves mm	→ 0.50x30 = 15 élèves MM

⌦ Comparer avec les résultats dans la classe!!

Troisième partie

- A nouveau le professeur distribue au hasard deux allèles à chaque élève.
- Comment savoir si les « enfants » créés par le professeur lors de la distribution des allèles constituent un échantillon issus d'une population respectant les hypothèses de l'équilibre de Hardy-Weinberg ?

☑ On sait quelle proportion de chaque génotype on a dans notre échantillon

☑ Grâce au théorème ci-dessus, on sait quelle proportion de chaque génotype on s'attend à avoir dans notre échantillon si les hypothèses de l'équilibre de Hardy-Weinberg sont respectés

Il suffit de comparer les deux résultats !!

- Comment savoir si les « enfants » créés par le professeur lors de la distribution des allèles constituent un échantillon issus d'une population respectant les hypothèses de l'équilibre de Hardy-Weinberg ?

☑ On sait quelle proportion de chaque génotype on a dans notre échantillon

☑ Grâce au théorème ci-dessus, on sait quelle proportion de chaque génotype on s'attend à avoir dans notre échantillon si les hypothèses de l'équilibre de Hardy-Weinberg sont respectés

Si les résultats sont fort différents, c'est qu'on ne respecte pas les hypothèses de l'équilibre de Hardy Weinberg

⌦ Nous devons vérifier si la différence entre nos observations et ce que l'on s'attendait a observer sous une hypothèse donnée (= dans notre cas: équilibre de Hardy-Weinberg) est due au ***hasard*** ou au fait que cette ***hypothèse est fausse***.

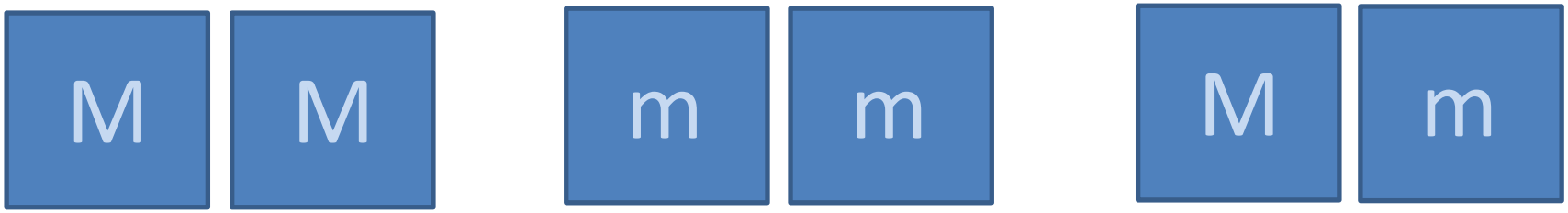
⌦ Nous disposons uniquement des données de l'**échantillon** pour vérifier cette hypothèse dans la **population**

⌦ Permet d'introduire le concept statistique de test d'hypothèse

H_0: L'équilibre d'Hardy-Weinberg est respecté dans la population

H_1: L'équilibre d'Hardy-Weinberg n'est pas respecté dans la population

☑ On sait combien d'enfants de chaque génotype on a dans notre échantillon



EXAMPLE (30 élèves)		
→ 9 élèves MM	→ 7 élèves mm	→ 14 élèves Mm

☑ Grâce au théorème ci-dessus, on sait quelle proportion de chaque génotype on s'attend à avoir dans notre échantillon si les hypothèses de l'équilibre de Hardy-Weinberg sont respectés

EXAMPLE (30 élèves)

On observe dans notre échantillon:

→ 9 élèves MM	→ 7 élèves mm	→ 14 élèves Mm

→ 9 x 2 + 14 = 32 allèles M ⇒ $p_M = 32/60 = 0.53$
→ 7 x 2 + 14 = 28 allèles m ⇒ $p_m = 28/60 = 0.47$

Si H_0 est vraie, on s'attendait à avoir dans notre échantillon:

$P_{MM} = P_M^2$	$P_{mm} = P_m^2$	$P_{Mm} = 2xP_MxP_m$
$P_{MM} = 0.53^2 = 0.28$ → 8.4 élèves MM	$P_{mm} = 0.47^2 = 0.22$ → 6.6 élèves mm	$P_{Mm}=2x0.53x0.47= 0.5$ → 15 élèves Mm

- On compare les deux résultats

	MM	mm	Mm
Observé:	9	7	14
Attendu si H0 est vrai:	8.4	6.6	15

Pour ce faire, on définit une « distance » entre les deux résultats:

$$Distance = \frac{(9-8.4)^2}{8.4} + \frac{(7-6.6)^2}{6.6} + \frac{(14-15)^2}{15} = 0.13$$

- Discuter l'intuition de cette formule avec la classe
- On peut donner l'expression théorique

⌦ Comment définir si cette distance est « trop grande »

Sans rentrer dans les détails:

Il existe des résultats théoriques qui permettent de définir un **seuil** au-delà duquel on considère la distance comme étant trop grande

Au dessus de ce seuil: la différence est trop grande pour être due au hasard, on conclut que H_0 est fausse: **la population ne respecte pas l'équilibre de Hardy-Weinberg**

En dessous de ce seuil: la différence est trop petite, on a pas assez de preuves pour conclure que H_0 est fausse: **on ne rejette pas l'hypothèse que la population respecte l'équilibre de Hardy-Weinberg**

→ Dans ce contexte, on montre que le seuil est égal à +/- 6

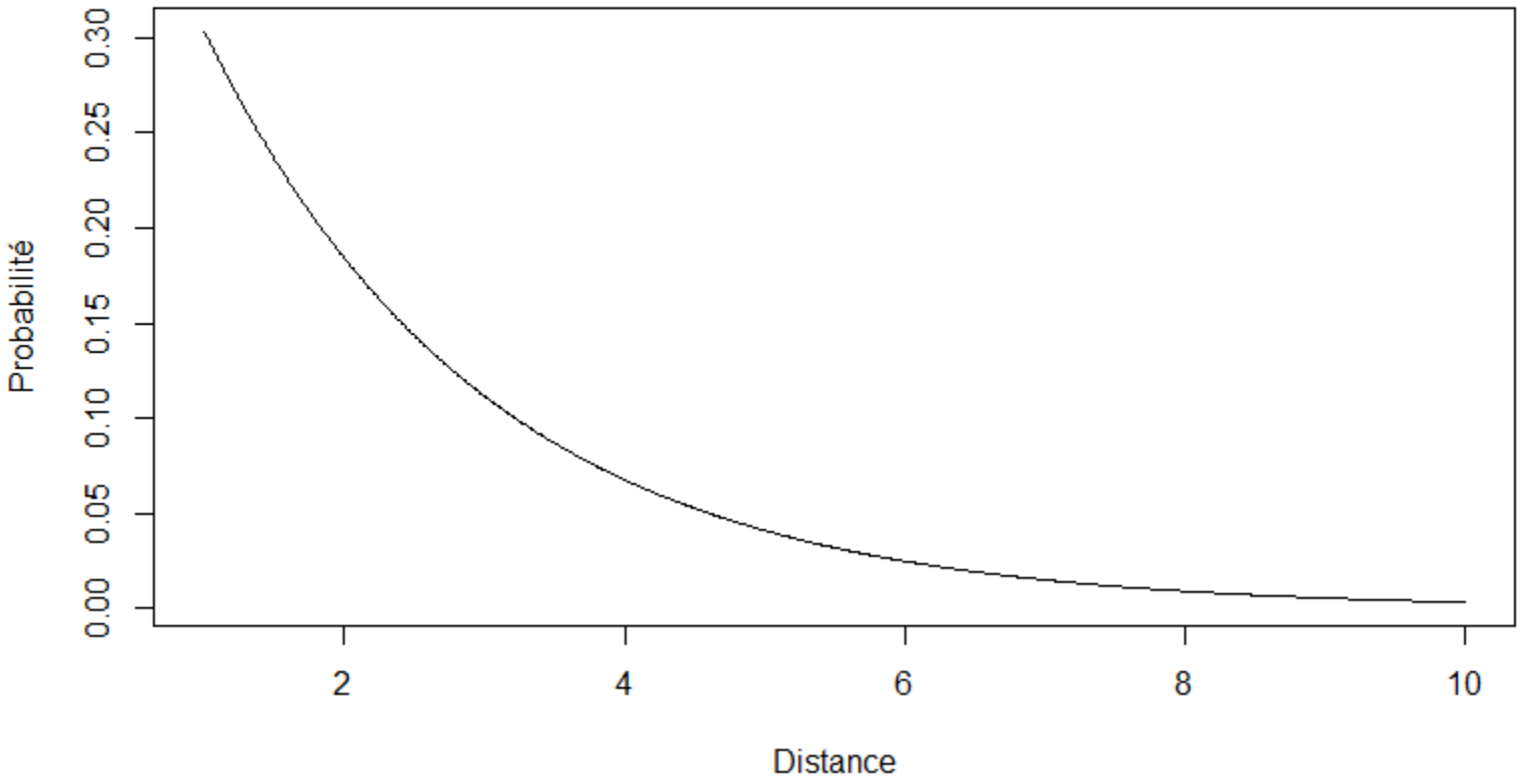
⌦ Comment définir si cette distance est « trop grande »

Si on veut rentrer (un peu plus) dans les détails:

Grâce à des résultats théoriques (notamment le ***Théorème Centrale-Limite***), on peut connaitre:

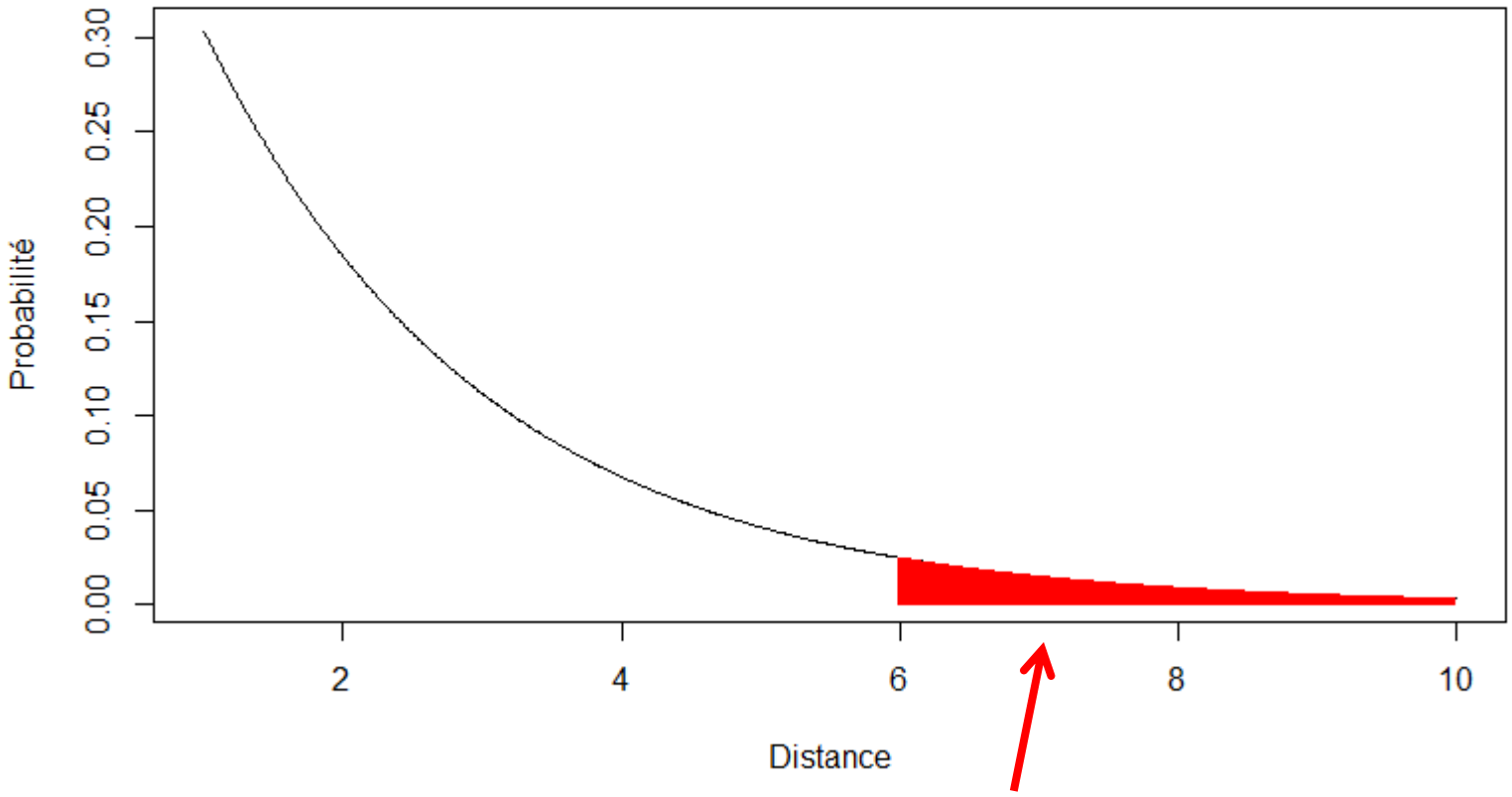
- Toutes les valeurs que l'on peut obtenir pour la Distance quelque soit l'échantillon choisit dans la population
- La probabilité (approximative) avec laquelle notre Distance va effectivement prendre chacune de ces valeurs **si H_0 est vraie**

→ on représente ces résultats par une ***fonction de densité***



→ Fonction de densité (chi-carré à deux degré de liberté)

[Surface sous la courbe entre a et b, représente la probabilité d'obtenir une valeur entre a et b]



Si la valeur de notre Distance fait partie des 5% de valeurs les plus grandes, on conclut (comme on que 5% de chances d'observer une telle valeur si H_0 est vraie) que H_0 n'est probablement pas vraie

$\Rightarrow$ on rejette H_0

Donc …

Grâce à un ***test d 'hypothèse statistique***, on peut savoir au départ des données d'un échantillon, si l'équilibre de Hardy-Weinberg est respecté dans la population dont provient cet échantillon

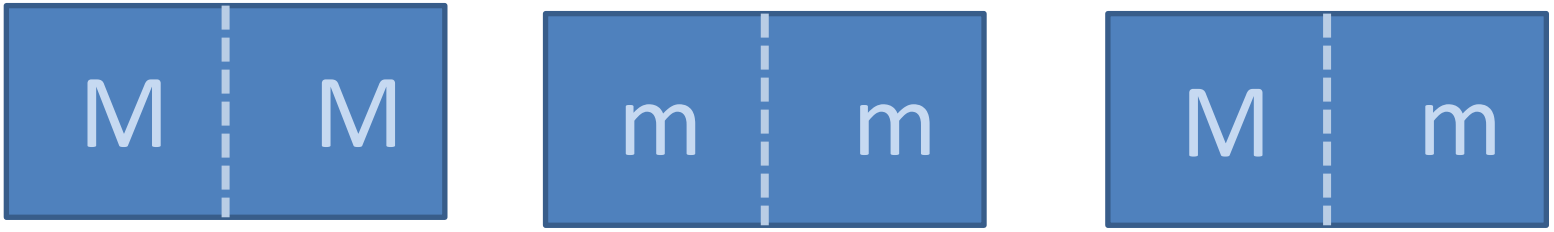
OK, mais pourquoi est-ce intéressant de savoir si une population respecte les hypothèses de l'équilibre de Hardy-Weinberg ?

❶ Peut servir de "contrôle de qualité"

❷ Peut mener les chercheurs à faire des expériences plus poussées pour identifier quelle hypothèse en particulier n'est pas respectée et faire des découvertes importantes (e.g. gène donnant un avantage de survie, …)

Pour illustrer ce dernier point …

- Le professeur distribue a nouveau de façon aléatoire un génotype chaque étudiant, qui le dissocie celui-ci en deux allèles, i.e.,



- Plutôt que de créer des couples aléatoire, on demande aux étudiants de se mettre en couple, MAIS, les étudiants choisissent en priorité un compagnon du même génotype.
- Chaque couple, crée un enfant en mettant en commun un allèle choisi au hasard de chacun des génotypes des parents

QUESTION 4: La génération des enfants est-elle un échantillon d'une population respectant l'équilibre d'Hardy-Weinberg?

QUESTION 4: La génération des enfants est-elle un échantillon d'une population respectant l'équilibre d'Hardy-Weinberg?

EXAMPLE (30 élèves ⇒ 15 enfants)

On observe dans notre échantillon:

→ 9 élèves MM	→ 7 élèves mm	→ 14 élèves Mm
→ 4 enfants MM	→ 6 enfants mm	→ 5 enfants Mm

→ 4 x 2 + 5 = 13 allèles M ⇒ $p_M = 13/30 = 0.43$

→ 6 x 2 + 5 = 17 allèles m ⇒ $p_m = 17/30 = 0.57$

Si H_0 est vraie, on s'attendait à avoir dans notre échantillon:

$P_{MM} = P_M^2$	$P_{mm} = P_m^2$	$P_{Mm} = 2xP_MxP_m$
$P_{MM} = 0.43^2 = 0.18$ → 2.7 enfants MM	$P_{mm} = 0.57^2 = 0.32$ → 4.9 enfants mm	$P_{Mm}=2x0.43x0.57= 0.49$ → 7.4 enfants Mm

⌦ On compare les deux résultats

	MM	mm	Mm
Observé:	4	6	5
Attendu si H0 est vrai:	2.7	4.9	7.4

Pour ce faire, on définit une « distance » entre les deux résultats:

$$Distance = \frac{(4-2.7)^2}{2.7} + \frac{(6-4.9)^2}{4.9} + \frac{(5-7.4)^2}{7.4} = 10.3$$

⌦ Discuter ce résultat avec la classe

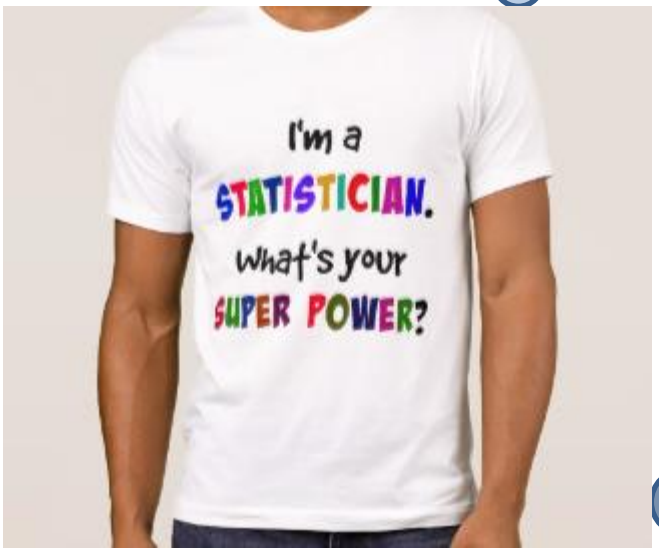
En conclusion

- **Objectif**: donner le « gout des stats » aux étudiants du secondaire
- **Idée**: mettre a disposition des professeurs du secondaire des « kits » contenant tout le matériel nécessaire pour animer un cours de quelques heures abordant la proba/stat de façon ″ludique″ via une application.
  - Dossier pédagogique pour l'enseignant
  - Proposition d'activité à faire en classe, avec documentation
  - Eventuellement code R
- **Pour le futur:**
  - Créer d'autres kits
  - Organiser une (demi-) journée sur la proba/stat pour les enseignants du secondaire à la LSBA

www.uclouvain.be/lsba

Comment devient-on statisticien?

Master en Statistique, orientation générale
Master en Statistique, orientation biostatistique
Master en Sciences des Données
Master en Sciences Actuarielles



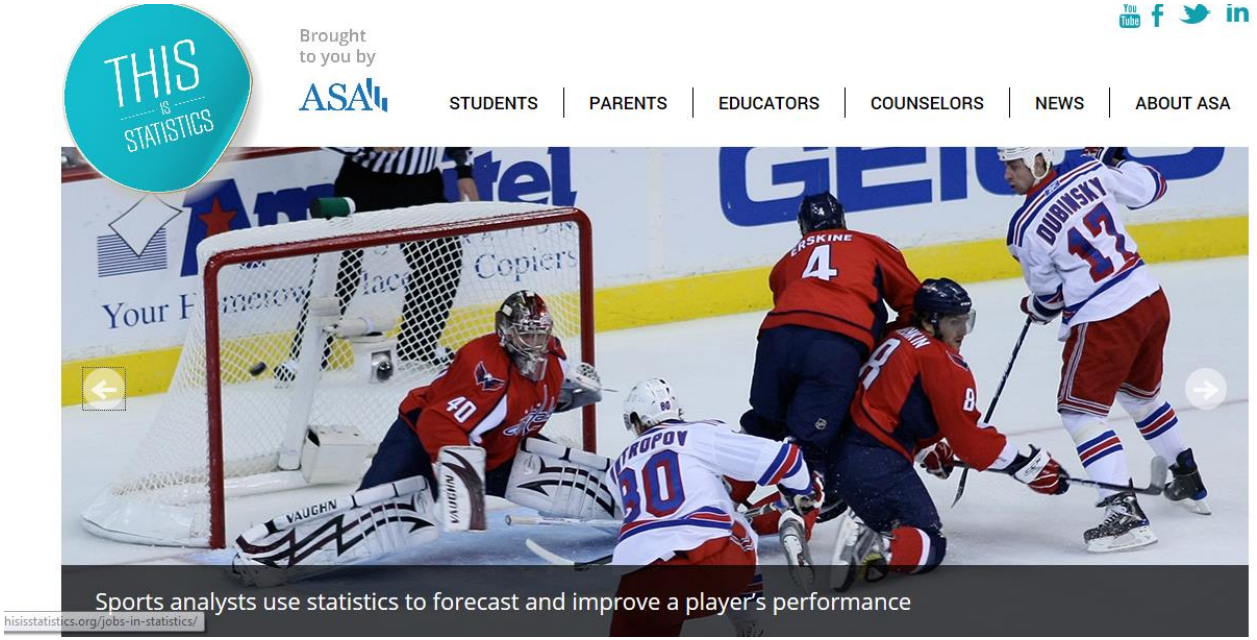
Baccalauréat / Master avec accès direct
Ex: mathématique, (bio-)ingénieur, sc eco, psycho, …

Baccalauréat / Master sans accès direct
→ Via mineure ou module complémentaire ou certificat

UCL Université catholique de Louvain

ISBA

Thisisstatistics.org



YouTube: La statistique expliquée à mon chat