Partie II : points méthodologiques - plan du cours -

1. Adapter les méthodes statistiques aux Big Data
   A. Répétition de tests et inflation du risque alpha
   B. Significativité ou taille de l'effet ?
2. Exemples de problèmes méthodologiques
   A. Problème des définitions utilisées
   B. Surajustement des modèles prédictifs
3. Exemples de problèmes liés à l'observation rétrospective
   A. Problème des maxima locaux
   B. Problème des études écologiques

1- Adapter les méthodes statistiques aux Big Data

A. Répétition de tests et inflation du risque alpha
B. Significativité ou taille de l'effet ?

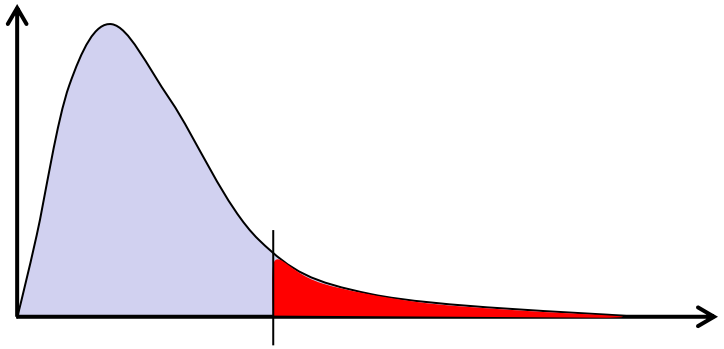
Rappel : principe des tests statistiques (1)

- Exemple : on souhaite tester si deux variables A et B sont indépendantes. Ex :
  - Deux variables qualitatives : Khi²
  - Deux variables quantitatives : test de nullité du coefficient de corrélation
  - Une variable binaire et une variable quantitative : test de Student (non apparié)
- On pose l'hypothèse nulle H_0 : les deux variables sont indépendantes. H_1 : hypothèse alternative
- On calcule une statistique de test :
  - On ne connait pas son comportement sous H_1 !
  - Mais sous H_0, on sait quelle loi la statistique de test suit

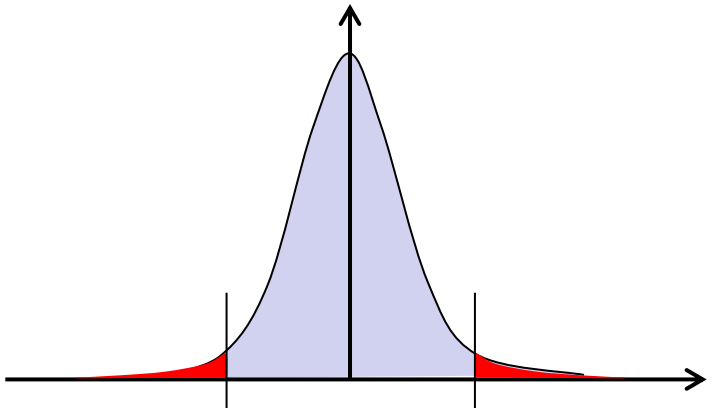
Rappel : principe des tests statistiques (2)

- Sous H_0, on sait quelle loi la statistique de test (ex : χ^2, t) suit.
- Zone de rejet de H_0 : valeurs extrêmes de la statistique de test correspondant à une probabilité choisie α (généralement 5% ; surface rouge sur ces schémas)

Zone de rejet d'un test du Khi²



Zone de rejet d'un test de Student



Rappel : principe des tests statistiques (3)

- Utilisation du test : on suppose H_0 vraie, si la statistique observée est peu probable (probabilité associée *p* inférieure à α), on décide de rejeter H_0.
- Mais si H_0 réellement vraie, le risque de rejeter H_0 à tort existe : c'est α, risque de première espèce (généralement 5%).

1.A- Répétition des tests et inflation du risque alpha

Et quand on réalise plusieurs tests ?

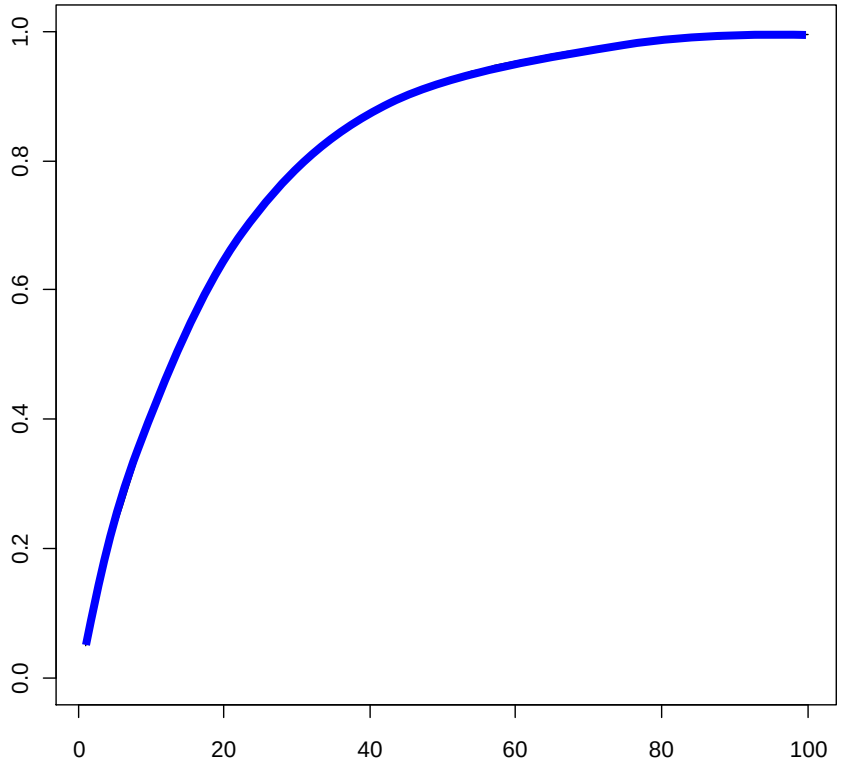
- On réalise k tests indépendants sur une série de données :
  - Rejet de H_0 du test i : R_i
  - Non-rejet de H_0 du test i : $\overline{R_i}$
  - Risque de première espèce (identique pour les k tests) : $\alpha_{indiv} = P(R_i / H_{0i})$
- Un « Signal » est observé si on rejette H_0 dans *au moins un* des k tests.
  - Signal : $Signal = R_1 \cup R_2 \cup ... \cup R_k$
  - Pas de signal : $\overline{Signal} = \overline{R_1} \cap \overline{R_2} \cap ... \cap \overline{R_k}$
- Supposons que toutes les hypothèses nulles H_{0i} soient vraies, quelle est la probabilité d'observer un signal ?
  - Pas de signal : $P(\overline{Signal}) = P(\overline{R_1}) \times P(\overline{R_2}) \times ... \times P(\overline{R_k})$

    $P(\overline{Signal}) = (1-\alpha_{indiv})^k$
  - Signal : $P(Signal) = \alpha_{total} = 1-(1-\alpha_{indiv})^k$
- On parle d'<u>inflation du risque α</u> :
  - Si k=1, $\alpha_{total}=\alpha_{indiv}$
  - Si k>1, $\alpha_{total}>\alpha_{indiv}$

Et quand on réalise plusieurs tests ?

- Inflation du risque alpha :

Risque alpha total si chaque test est réalisé au risque 5%



Nombre de tests (0-100)

1.A- Répétition des tests et inflation du risque alpha

Et quand on réalise plusieurs tests ?

- Pour conserver le α_{total} souhaité, il faut corriger le seuil de significativité de chaque test :
  - Correction de Šidák :
    - Utiliser $\alpha_{indiv} = 1-(1- \alpha_{total})^{1/k}$
    - Facile à retrouver, rappel : $a^b = e^{b.ln(a)}$
  - Correction de Bonferroni, très populaire :
    - Pour les risques α_{total} habituellement souhaités (1 à 10%), utiliser plus simplement : $\alpha_{indiv} = (\alpha_{total})/k$
    - Donc en pratique utiliser pour chaque test le seuil de significativité 0.05/k
    - Seuil légèrement inférieur au seuil proposé par Šidák, donc plus conservateur, mais très proche
    - Valable même si les k tests ne sont pas indépendants !

Quand réaliser une correction de Bonferroni ?

- A chaque fois qu'on lance une batterie de tests en espérant en voir au moins un significatif.
- *Ex 1 : Pour montrer la supériorité d'un somnifère sur un autre, on compare toutes les k caractéristiques du sommeil mesurables dans deux groupes de patients à l'aide de k tests de Student (délai d'endormissement, durée de sommeil profond, délai avant le premier réveil, délai jusqu'au lever effectif, nombre de cauchemars par semaine, etc.)*
- *Ex 2 : Pour prouver l'origine génétique d'une maladie, on séquence le génome entier de patients et on teste réalise k tests du Khi², entre chacun des k allèles examinés et la maladie.*
- => utiliser le seuil de significativité 0.05/k pour chaque test et non 0.05, sinon ce serait malhonnête !

Exercice 3

Exprimer le résultat d'une procédure de test

- On calcule d'abord une variable de décision
- Puis on teste cette variable de décision, deux manières d'exprimer le résultat :
  - La p valeur (significativité) :
    - Le test retourne le « p » obtenu en testant si cette variable est significativement différente de la valeur attendue sous H_0 (0 ou 1 en général)
    - On rejette H_0 si « p » est inférieur au risque α
  - L'intervalle de confiance :
    - Le test retourne l'intervalle de confiance à 1-α de cette variable
    - On rejette H_0 si cet intervalle de confiance ne contient pas la valeur attendue sous H_0 (0 ou 1 en général)
- Ces deux options sont équivalentes

1.B- Significativité versus taille de l'effet
Exemples des proportions

- Comparer deux proportions avec des lois normales au risque 5% :
  - H_0 : $\pi_1 = \pi_2$ donc $\pi_1 - \pi_2 = 0$
  - Variable de décision $\Delta = \pi_1 - \pi_2$, estimée par $d = p_1 - p_2$
  - p valeur :
    - « p » obtenu en testant si d est significativement différent de 0 (valeur attendue sous H_0)
    - => on rejette H_0 si $p < 0.05$
  - Intervalle de confiance :
    - L'intervalle de confiance à 95% de Δ
    - => on rejette H_0 si cet intervalle de confiance ne contient pas 0 (valeur attendue sous H_0)

} Plus rare, mais équivalent

Exemple des proportions

- Comparer deux proportions avec le risque relatif RR (au risque 5%) : cas particulier d'une cohorte avec une exposition E et une maladie M
  - H_0 : $P(M/E)=P(M/\overline{E})$ donc $P(M/E) / P(M/\overline{E})=1$
  - Variable de décision $RR= P(M/E) / P(M/\overline{E})$ estimée par $\widehat{RR}$
  - p valeur :
    - « p » obtenu en testant si le $\widehat{RR}$ est significativement différent de 1 (valeur attendue sous H_0)
    - => on rejette H_0 si $p < 0.05$
  - Intervalle de confiance :
    - L'intervalle de confiance à 95% du RR
    - => on rejette H_0 si cet intervalle de confiance ne contient pas 1 (valeur attendue sous H_0)

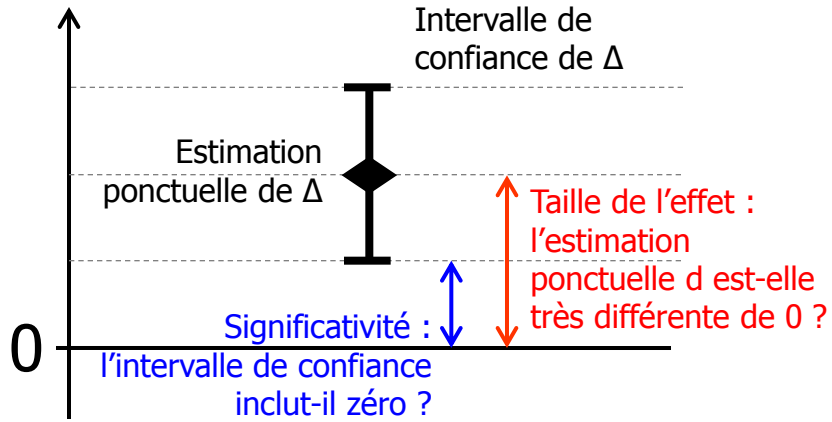
} Plus rare, mais équivalent

1.B- Significativité versus taille de l'effet

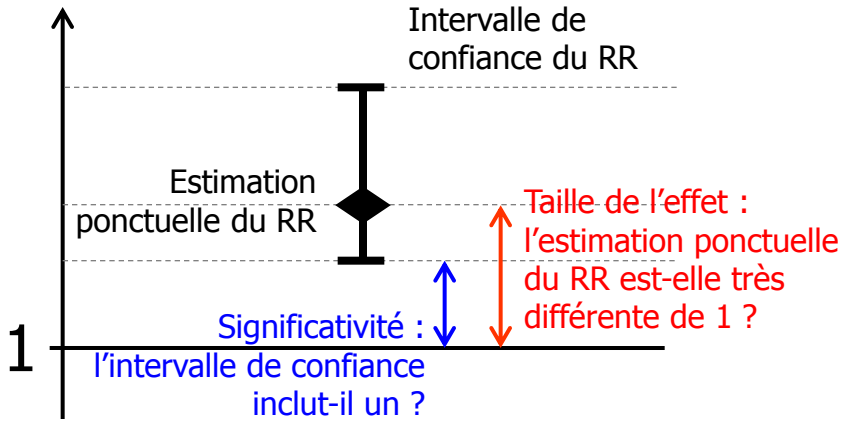
Exemple des proportions

- Représentons les intervalles de confiance précédents :

Intervalle de confiance de Δ :



Intervalle de confiance du RR :



- Grands effectifs :
  - rétrécissement des IC => volontiers résultats « très significatifs »
  - Ne change pas la « taille de l'effet », qui figure l'importance d'un facteur de risque

Exercice 4

2- Exemples de problèmes méthodologiques

A. Problème des définitions utilisées
B. Surajustement des modèles prédictifs

2.A- Problème des définitions utilisées

- Les problèmes méthodologiques des études rétrospectives concernent souvent des points non statistiques !

- Toujours vérifier les définitions, surtout lorsqu'on utilise des données rétrospectives, recueillies ou calculées pour d'autres exigences que la présente étude

- Plus encore dans les recueils multicentriques et les méta-analyses

Exercice 5

2.B- Surajustement des modèles prédictifs
Position du problème

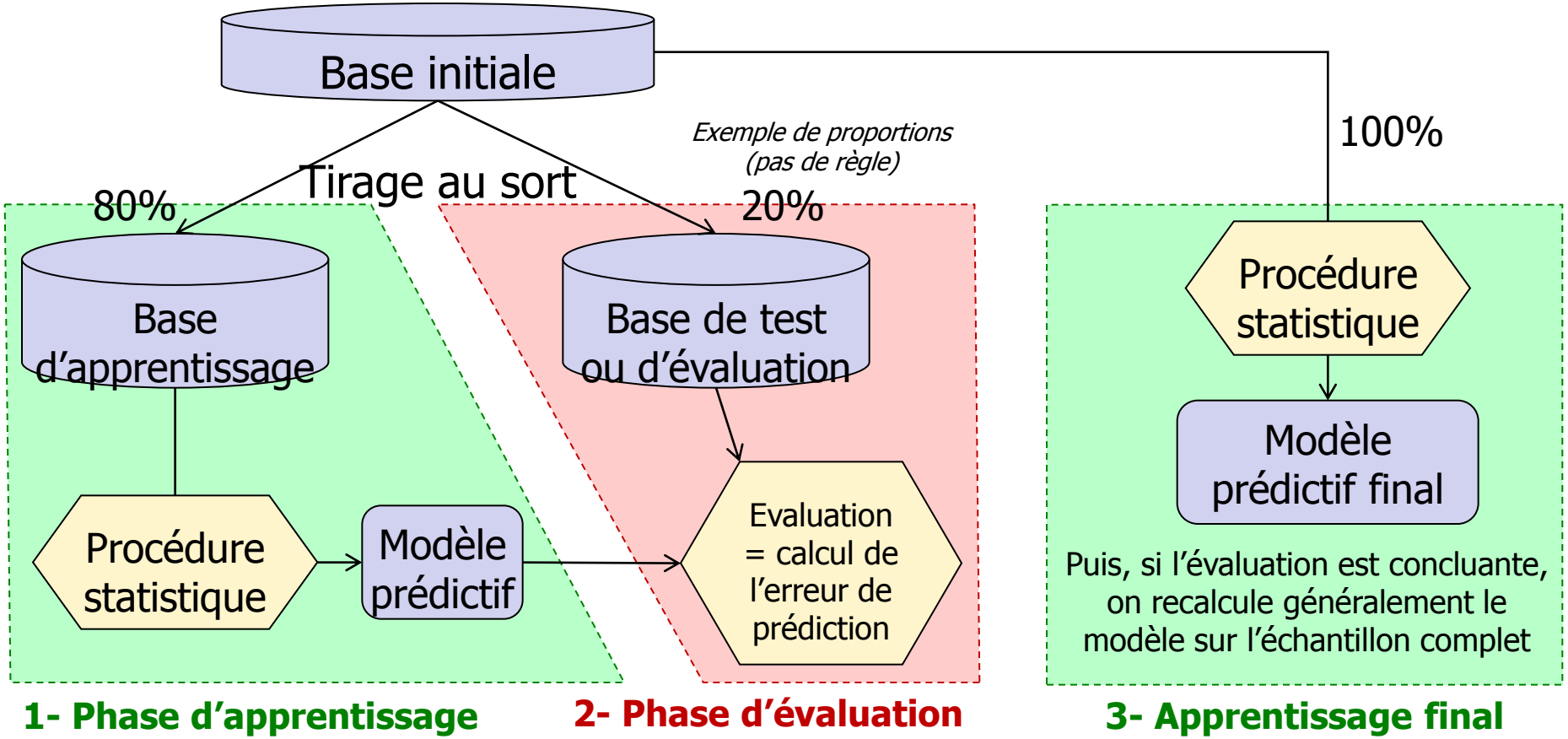
- *Exemple de modèle prédictif : poids = 0.71*taille + 8.4*sexeH - 57*
- Il est aisé de découvrir des modèles prédictifs en *Big Data* et en *Data Reuse* :
  - Approches de data mining en aveugle, sans hypothèse *a priori* (on découvre par l'observation des données : *data-driven approach*)
  - Inflation non corrigée du risque alpha (cf. exercice précédent, ou exemple des arbres de décision CHAID qui réalisent tous les tests du Khi² possibles)
  - Grands effectifs => tests avec résultats généralement très significatifs, même lorsque la taille de l'effet est assez faible
- Risque de « surajustement » du modèle :
  - *Exemple : dans l'amphithéâtre de 3100 étudiants de PACES, peut « améliorer » le modèle ci-dessus en ajoutant la longueur des cheveux, le nombre de couleurs utilisées pour écrire, le temps passé au petit-déjeuner, etc.*
  - Le modèle fonctionne bien sur l'échantillon : il est « optimisé » pour cet échantillon en particulier.
  - Mais cela peut être fortuit : le même modèle n'aurait pas forcément été découvert sur un autre échantillon

2.B- Surajustement des modèles prédictifs
Position du problème

- Solutions aux problèmes de surajustement, à appliquer simultanément :
  - Correction du risque alpha
  - Taille de l'effet : inclure des critères tenant compte de la taille de l'effet dans les algorithmes
  - Critères de parcimonie : inclure un indice augmenté par la précision de prédiction mais diminué par le nombre de variables utilisées, permettant de privilégier les modèles « parcimonieux ». Compromis entre précision et simplicité du modèle.
  - Tous les modèles prédictifs doivent être bâtis sur un échantillon et testés sur un autre. Procédure détaillée ci-après.

2.B- Surajustement des modèles prédictifs
Solution : apprentissage et évaluation

- Exemple de procédure pour les modèles explicatifs ou prédictifs :

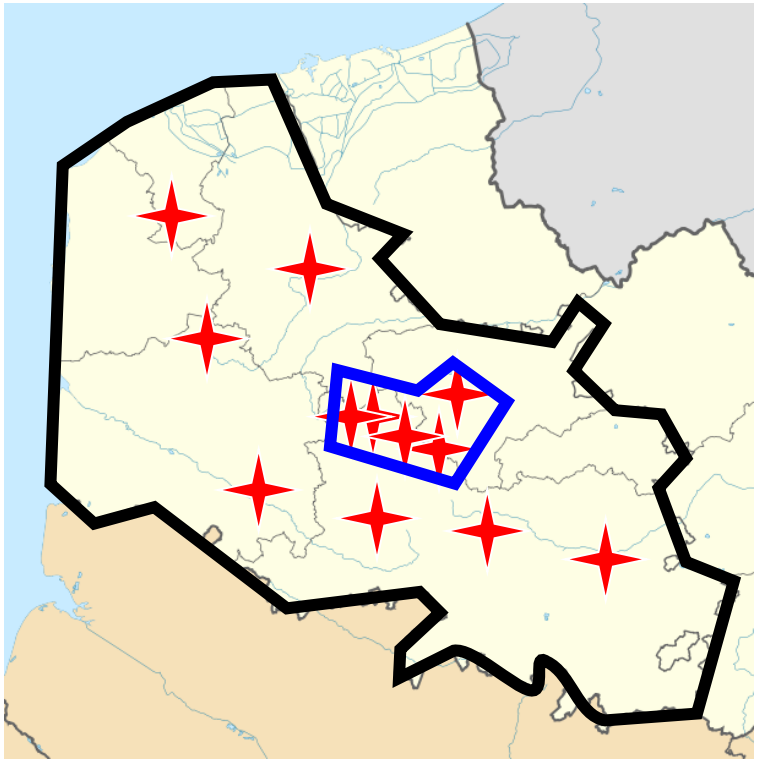


3- Exemples de problèmes liés à l'observation rétrospective

A. Problème des maxima locaux
B. Problème des études écologiques

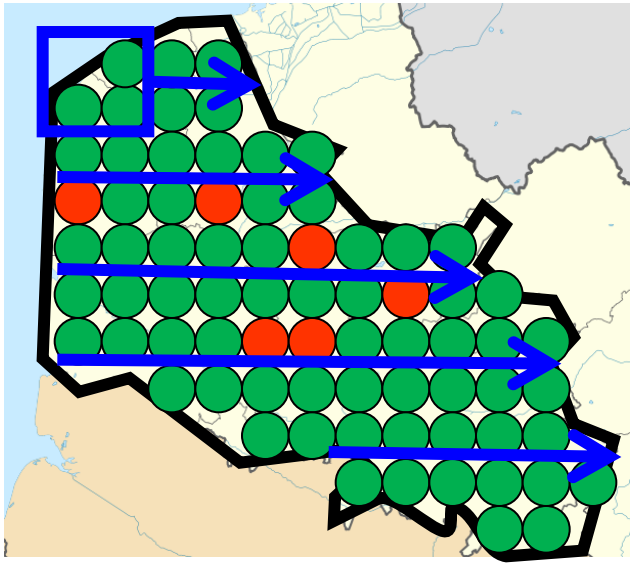
3.A- Problème des maxima locaux
Position du problème

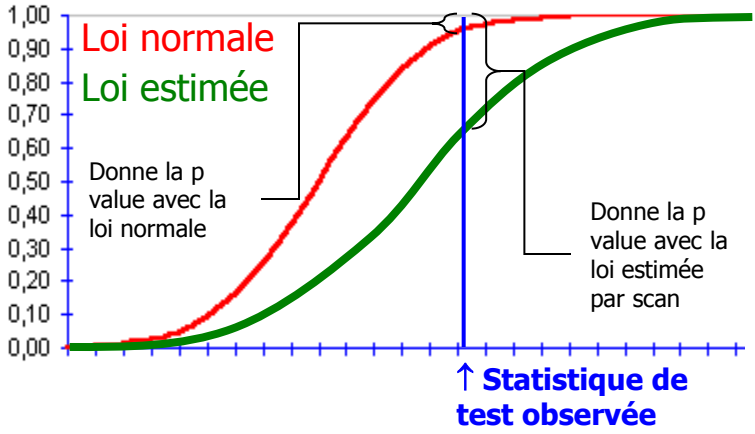
- Attitude classique (exemple fictif) :
  - On observe des cas de cancers sur une carte (✦), on observe une concentration de cas
  - On définit a posteriori une zone de forte prévalence
  - On teste cette zone contre le reste du plan (en tenant compte de la population)
  - Et, naturellement, le test est significatif !
- Amène à découvrir à tort des excès de cas :
  - Avant de voir la carte, on ne savait pas quelle zone tester
  - Or, si des cas sont distribués au hasard sur le plan, il n'est pas choquant d'observer des concentrations locales
  - => pas honnête ! Solution : scans statistiques



3.A- Problème des maxima locaux
Scans statistiques (simplifié)

- Dans l'exemple précédent :
  - On a calculé d=p_1-p_2 et une statistique de test, censée suivre *N(0,1)* sous H_0.
  - On a utilisé *N(0,1)* pour conclure
- Principe du Scan Statistique
  - Simulation n°1 (reproduit H_0 !) :
    - individus placés sur le plan
    - Aléatoirement malades ou non malades, selon le taux de prévalence observé réellement
    - On déplace une fenêtre de scan et on calcule à chaque position la statistique de test
    - On mémorise la valeur la plus extrême obtenue
  - … simulation n°1000000 (reproduit H_0 !) : idem
  - Enregistrement des valeurs prises par la statistique de test sous H_0
  - => estimation de la loi de distribution sous H_0
  - On calcule la p value d'après la loi estimée, et non d'après la loi normale
  - => conclusion généralement moins optimiste, p value supérieure



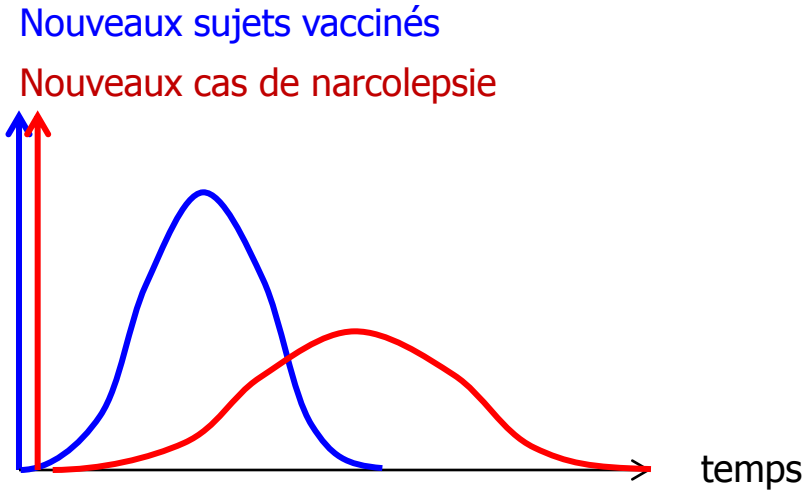


3.B- Problème des études écologiques
Position du problème

- Définition :
  - étude épidémiologique dans laquelle les critères analysés concernent une population plutôt que des individus
- Importance des biais :
  - cohorte < cas-témoin < étude écologique
  - beaucoup de fausses découvertes
  - emballement médiatique !
- Utilisation : c'est un « signal »
  - à accueillir avec beaucoup de réserves
  - et à confirmer par des méthodes plus robustes

3.B- Problème des études écologiques
Un exemple récent

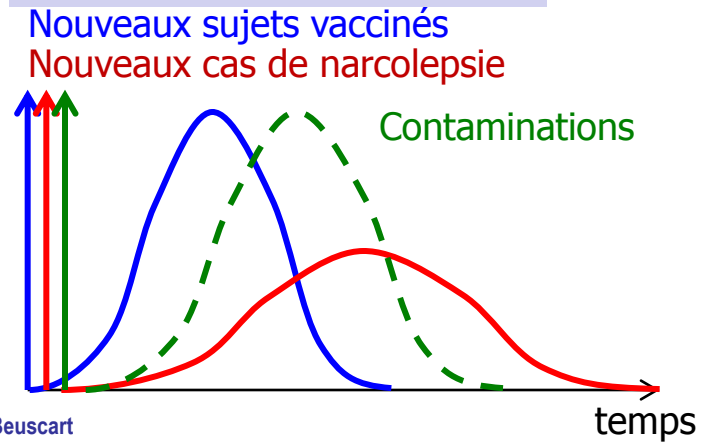
- Épidémie de narcolepsies après la campagne de vaccination contre le virus de la grippe H1N1
- Narcolepsie : maladie neurologique très rare et invalidante, accès de sommeil, et cataplexies lors des émotions fortes
- 51 cas confirmés en France après la campagne de vaccination 2009-2010 contre la grippe pandémique A (H1N1). Idem en Europe.
- Taux d'incidence très supérieur au taux habituel



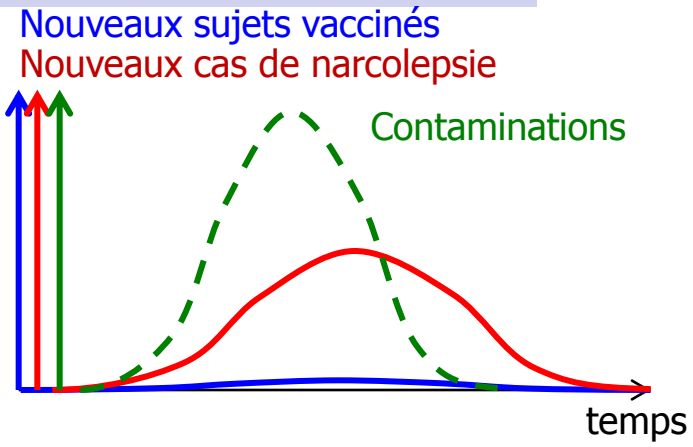
3.B- Problème des études écologiques
Un exemple récent

- Pour les media, le lien est évident. Pour beaucoup de Français, il ne faut plus se faire vacciner. Mais…
  - 51 cas confirmés pour 5,7 millions de vaccinations. Or la grippe saisonnière « simple » tue des milliers de personnes en France !
  - Même augmentation du taux d'incidence en Chine, en l'absence de campagne de vaccination
  - Un point commun mal mesuré : vague de contamination par le virus. Les personnes vaccinées sont elles aussi contaminées, même si elles développent pas ou peu de symptômes ! La narcolepsie pourrait donc être un effet du virus lui-même. A prouver…

FRANCE



CHINE



Conclusion de ce cours

- Objectifs :
  - Découvrir de nouvelles connaissances, parfois sans *a priori*
  - Détection précoce de signaux à confirmer
  - Par *data reuse* en *big data*
- Divers problèmes abordés ici :
  - Spécifiques à ce cadre ou accentués par ce cadre
  - Effet majeur : risque de fausse découverte de connaissance
- Solutions proposées :
  - Méthodes à appliquer intelligemment et honnêtement, c'est-à-dire dans le sens qui diminue le risque alpha, et augmente le risque bêta
  - En effet : rejet de $H_0 \Rightarrow$ décision
    non-rejet de $H_0 \Rightarrow$ simple *statu quo* (ne prouve pas que H_0 soit vraie)
  - Interprétation avec précautions, confirmation toujours nécessaire