Les cycles d'apprentissage de BioSP - Pôle M2SUD

Automne 2021 - Cycle 1 - Introduction à la statistique Bayésienne

06/09	Rappels en probabilité (Emily)
20/09	Principes de la démarche bayésienne en statistique (Samuel)
04/10	Initiation aux modèles graphiques (Emily)
18/10	Distributions a priori Présentation des logiciels WinBUGS JAGS STAN (Emily)
08/11	Vérification et évaluation de modèles (Samuel)
22/11	Estimation des distributions a posteriori avec des méthodes numériques (Samuel)
29/11	Séminaire 1 d'ouverture : Estimation de la distribution spatio-temporelle de mammifères marins (Matthieu Authier)
06/12	Séminaire 2 d'ouverture : Approche bayésienne pour la prévision probabiliste de températures à partir de post-traitements de données d'ensemble (Eric Parent)
13/12	Séminaire 3 d'ouverture : Reconstruction de l'histoire évolutive de langues humaines (Robin Ryder)

Les cycles d'apprentissage de BioSP - Pôle M2SUD

Automne 2021 - Cycle 1 - Introduction à la statistique Bayésienne

06/09	Rappels en probabilité (Emily)
20/09	**Principes de la démarche bayésienne en statistique**
04/10	Initiation aux modèles graphiques (Emily)
18/10	Distributions a priori Présentation des logiciels WinBUGS JAGS STAN (Emily)
08/11	Vérification et évaluation de modèles (Samuel)
22/11	Estimation des distributions a posteriori avec des méthodes numériques (Samuel)
29/11	Séminaire 1 d'ouverture : Estimation de la distribution spatio-temporelle de mammifères marins (Matthieu Authier)
06/12	Séminaire 2 d'ouverture : Approche bayésienne pour la prévision probabiliste de températures à partir de post-traitements de données d'ensemble (Eric Parent)
13/12	Séminaire 3 d'ouverture : Reconstruction de l'histoire évolutive de langues humaines (Robin Ryder)

Principes de la démarche bayésienne en statistique

Eric Parent

18 Mars 2019

Incluant des ajouts de Samuel Soubeyrand pour le cycle d'apprentissage « Introduction à la statistique Bayésienne » - 20 sept. 2021

Hier

Bayésien en réaction au Fréquentiste?

L'école traditionnelle fréquentiste, la pratique classiquement enseignée en France:

- présente une *boîte à outils* de recettes probabilistes dont l'utilisateur n'est pas toujours à même de comprendre facilement le fil directeur,
- raisonne avec *une seule* loi de probabilité: la loi d'échantillonnage.
  - elle s'interprête selon une perspective de répétition asymptotique,
  - seules les quantités observables peuvent être probabilisées,
  - une rupture de pensée de la modélisation à l'inférence.
- Qu'est ce qu'un intervalle de confiance? Pourquoi 95%?
- Qu'est ce qu'une erreur de type 1 ? Conditionnelle à quel événement?

La *confiance* ne porte pas sur la valeur éventuelle de l'inconnue, mais sur la recette statistique d'inférence choisie. Comment? Pourquoi?

Une parenthèse : l'estimation fréquentiste des paramètres

- $\theta \in \Theta$: paramètres inconnus intervenant dans l'écriture d'un modèle, symbolisé par $\mathcal{M}_\theta$
- Objectif : estimer les paramètres du modèle à partir de données
- En d'autres termes : évaluer sous quel $\theta_0 \in \Theta$ ont été obtenues les données sachant que le modèle qui les a générées est contenu dans la classe de modèles $\{\mathcal{M}_\theta : \theta \in \Theta\}$

- Cas du modèle de régression:

$$y = \mathcal{M}_\theta(x) + \epsilon$$

  où y: variable réponse, x: variable explicative, $\mathcal{M}_\theta$: fonction déterministe de x, ϵ: bruit
- Objectif : estimer θ en utilisant les données qui sont, dans le cas le plus simple, des observations répétées et indépendantes de (x, y), c'est-à-dire un échantillon $\{(x_i, y_i) : i = 1, \ldots, I\}$

L'estimateur des moindres carrés:

- Expression générale:

$$\hat{\theta} = \texttt{argmin}_{\theta \in \Theta} \sum_{i=1}^{I} \{y_i - \mathcal{M}_\theta(x_i)\}^2$$

- Cas particuliers: expressions exactes disponibles
  e.g., régression linéaire: $(y = x'\theta + \epsilon)$, $\hat{\theta} = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y}$
- Cas général: minimisation via un algorithme numérique itératif (Gauss-Newton, Nelder-Mead, recuit simulé...)

L'estimateur du maximum de vraisemblance:

- Hypothèse distributionnelle pour le bruit
  E.g., $Y_i = \mathcal{M}_\theta(x_i) + \epsilon_i$, les ϵ_i sont indépendants et identiquement distribués (i.i.d.) selon la loi $\mathtt{Normale}(0, \sigma^2)$ où $\sigma > 0$
  (Y_i est la variable aléatoire qui, à travers l'échantillonnage, prend la valeur observée y_i)
- La vraisemblance est, sous une telle hypothèse, la loi jointe des Y_i sachant les x_i
  Dans l'ex gaussien, loi conditionnelle de Y:

$$Y \mid X = x \sim \mathtt{Normale}(\mathcal{M}_\theta(x), \sigma^2).$$

  densité de probabilité de Y sachant $X = x$:

$$y \mapsto f_{Y|X}(y|x; \theta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - \mathcal{M}_\theta(x))^2}{2\sigma^2}\right).$$

  vraisemblance:

$$L(\theta; (y_i, x_i) : i = 1, \ldots, I, \sigma) = \prod_{i=1}^{I} f_{Y|X}(y_i|x_i; \theta, \sigma).$$

- Il faut ensuite maximiser la vraisemblance par rapport à θ:

$$\hat{\theta} = \texttt{argmax}_{\theta \in \Theta} L(\theta; (y_i, x_i) : i = 1, \ldots, I, \sigma)$$

- Expression analytique de $\hat{\theta}$ pour certains modèles
- Dans le cas général, maximisation via un algorithme numérique itératif
  - Gauss-Newton, Nelder-Mead, recuit simulé
  - EM: *Expectation-Maximization*
  - MCEM: *Monte Carlo Expectation-Maximization*
  - ...

Mesurer l'incertitude de $\hat{\theta}$

... et plus généralement décrire les propriétés de $\hat{\theta}$

- Une estimation ponctuelle ne suffit pas
- Le $\hat{\theta}$ est-il éloigné du θ "vrai", noté θ_0?
- **Il faut** fournir des mesures de l'incertitude de $\hat{\theta}$
  ex: variances d'estimation, intervalles de confiance, distributions...
- Les mesures d'incertitude doivent prennent généralement en compte:
  - l'aléa lié à l'échantillonnage
  - la structure probabiliste de l'aléa ϵ

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact:
  ex: $y = x'\theta + \epsilon$, $\epsilon \sim_{indep.}$ Normale$(0, \sigma^2)$, σ^2 connu, $\hat{\theta} \sim$ Normale$(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact:
  ex: $y = x'\theta + \epsilon$, $\epsilon \sim_{indep.}$ Normale$(0, \sigma^2)$, σ^2 connu, $\hat{\theta} \sim$ Normale$(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$
- A l'aide de l'asymptotique (quand $I \to \infty$):
  Sous des hypothèses de régularité, l'estimateur $\hat{\theta}$ du minimum de contraste est consistent et asymptotiquement normal:
  - $\hat{\theta}$ converge en probabilité vers θ_0
  - $\sqrt{I}(\hat{\theta} - \theta_0)$ converge en loi vers la loi $Normale(0, \mathcal{J}_{\theta_0}^{-1}\Gamma_{\theta_0}\mathcal{J}_{\theta_0}^{-1})$

Mesurer l'incertitude de $\hat{\theta}$

- A l'aide d'un argument exact:
  ex: $y = x'\theta + \epsilon$, $\epsilon \sim_{indep.}$ Normale$(0, \sigma^2)$, σ^2 connu, $\hat{\theta} \sim$ Normale$(\theta_0, \sigma^2(\mathcal{X}'\mathcal{X})^{-1})$
- A l'aide de l'asymptotique (quand $I \to \infty$):
  Sous des hypothèses de régularité, l'estimateur $\hat{\theta}$ du minimum de contraste est consistent et asymptotiquement normal:
  - $\hat{\theta}$ converge en probabilité vers θ_0
  - $\sqrt{I}(\hat{\theta} - \theta_0)$ converge en loi vers la loi $Normale(0, \mathcal{J}_{\theta_0}^{-1}\Gamma_{\theta_0}\mathcal{J}_{\theta_0}^{-1})$
- A l'aide de simulations (bootstrap (non-)paramétrique):
  1. Simuler B jeux de données
     - soit sous le modèle $\mathcal{M}_{\hat{\theta}}$
     - soit par ré-échantillonnage des données
  2. Estimer θ pour les B jeux de données
  3. Utiliser les B estimations de θ pour caractériser la loi de $\hat{\theta}$

Bayésien en réaction au Fréquentiste?

L'école traditionnelle fréquentiste, la pratique classiquement enseignée en France:

- présente une *boîte à outils* de recettes probabilistes dont l'utilisateur n'est pas toujours à même de comprendre facilement le fil directeur,
- raisonne avec *une seule* loi de probabilité: la loi d'échantillonnage. `->càd la loi des données`
  - elle s'interprète selon une perspective de répétition asymptotique,
  - `affirmation<- contestable` seules les quantités observables peuvent être probabilisées,
  - une rupture de pensée de la modélisation à l'inférence.
- Qu'est ce qu'un intervalle de confiance? Pourquoi 95%?
- Qu'est ce qu'une erreur de type 1 ? Conditionnelle à quel événement?

La *confiance* ne porte pas sur la valeur éventuelle de l'inconnue, mais sur la recette statistique d'inférence choisie. Comment? Pourquoi?

Aujourd'hui

Le raisonnement bayésien

- repose sur la notion de probabilité *personnelle*.
- la probabilité (bayésienne) doit être interprétée comme un degré de croyance, un pari quant à la connaissance d'une quantité incertaine. Toute probabilité est conditionnelle.
- la probabilité (bayésienne) se met à jour quand on dispose d'information (l'assimilation de données permet l'appentissage statistique)
- la cohérence du raisonnement bayésien est garantie grâce aux règles mathématiques du calcul des probabilités.
- Conditionner son raisonnement aux données et aux hypothèses aide à ne pas perdre de vue la différence entre le *petit monde* du formel et le *grand monde* du réel.

Petite histoire de la pensée bayésienne:

- Thomas Bayes (1701-1761) et Pierre-Simon de Laplace (1749-1827)
- R.O. *vs* Stats théoriques 18ème->Poincaré->Alan Turing->1950
- Wald->Definetti, Savage, Jeffrey, Lindley—>1990
- 1990 -> Berger, Robert, O'Hagan, Dawid, Marin, Gelman, etc. Computational Bayes & Machine Learning

Lire: S. Mac Grayne (2012): The theory that would not die...

Une mise en pratique facile

Pour construire un modèle bayésien, il faut une *loi a priori* et un *modèle d'échantillonnage*:

- L'existence mathématique de ces deux composantes a été prouvée dans des conditions de structuration a minima d'un raisonnement prédictif en présence d'incertitude (théorème dit de représentation de DeFinetti-Savage-Hewitt)
- La loi a posteriori est proportionnelle au produit vraisemblance $\times$ prior.
- La loi a posteriori est l'objet qui récapitule toute l'inférence.
- Conceptuellement simple: l'inférence cherche à exprimer ce que l'on sait sur les inconnues du problème sachant ce qui est connu (ou assumé comme tel).

Le raisonnement conditionnel

- Deux quantités seulement en Bayes : les non observables θ et les observables Y.
- Les valeurs manquantes, les paramètres, les valeurs à prévoir,... $\in \theta$!
- le modèle bayésien demande $[\theta]$ et $[Y|\theta]$
- l'inférence bayésienne : ce que l'on sait sur les inconnues au vu des observées $[\theta|Y]$
- mise à jour séquentielle facile !
- exemple des tirages binomiaux et normaux.
- réseaux bayésiens et DAG.(Éric)

La machinerie bayésienne

Inférence Bayésienne = Apprentissage à partir des données

- Qui : Vous!
- Quoi ? Notations: Y, y, θ, $[\bullet|\bullet]$
- Construction: Prior, vraisemblance

$$[\underset{Y}{y}|\theta] \times [\theta] = [\theta, Y] = [\theta|Y] \times [Y]$$

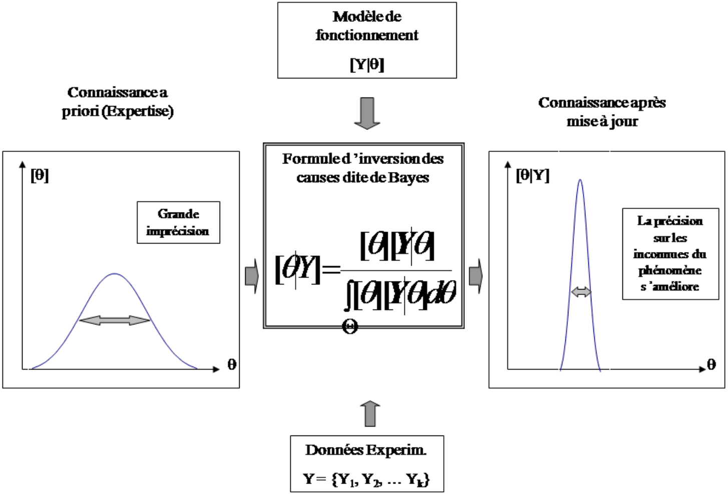
- Déduction? Posterior, prédictive

$$[\theta|Y = y] \propto [y|\theta] \times [\theta]$$

- Prédiction? Integration sur l'inconnue

$$[Y = y] = \int_\theta [Y = y|\theta] \times [\theta] d\theta$$

La machinerie bayésienne = apprentissage statistique



4

Exemple (1750 < <1990): conjugaison beta-binomiale

- $Y \sim dbin(\theta, n)$:

$$[Y\,|\theta] = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y}(1-\theta)^{n-y}$$

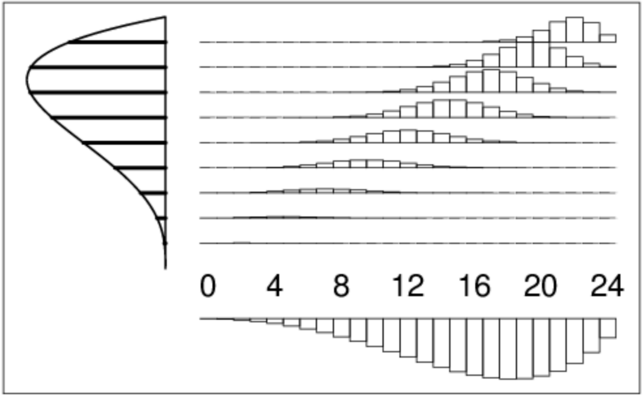
- $\theta \sim dbeta(a, b)$:

$$[\theta] = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

- $\theta|Y \sim dbeta(a+y, b+n-y)$
- Application $n = 24, y = 10,\ a = 3, b = 2$

$$
\begin{aligned}
[\theta|Y] &= [Y|\theta][\theta]/[Y] \\
&\propto [Y|\theta][\theta] \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y}(1-\theta)^{n-y} \\
&\quad \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}
\end{aligned}
$$

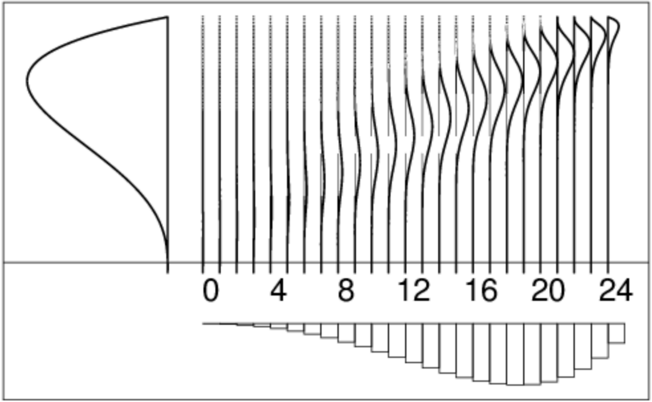
Exemple: conjugaison beta-binomiale $[\theta, Y] = [Y|\theta] \times [\theta]$



axe vertical: θ

axe horizontal: $Y = y$

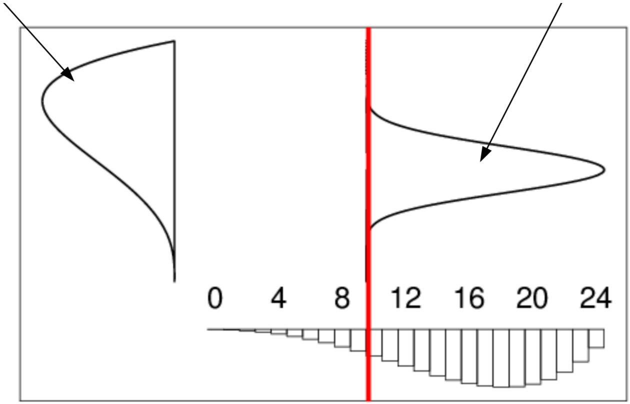
Exemple: conjugaison beta-binomiale $[\theta, Y] = [\theta|Y] \times [Y]$



axe vertical: θ

axe horizontal: $Y = y$

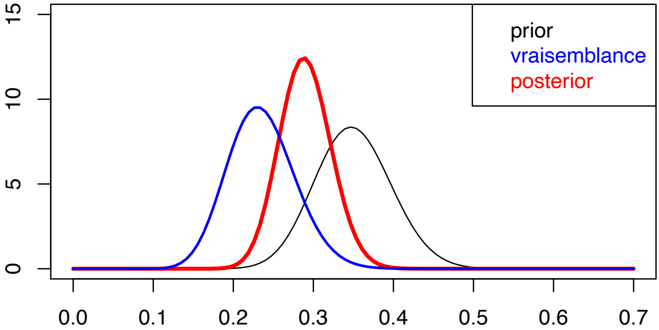
Exemple: conjugaison beta-binomiale $[\theta|Y] \propto [\theta, Y]$



axe vertical: θ

axe horizontal: $Y = y$

Apprentissage = Mise à jour



- Plus d'information dans la loi *a posteriori*
- Compromis de synthèse! cf. positionnement central
- Poids relatif des sources d'information

Le paradigme bayésien de l'apprentissage statistique

- *Le posterior d'aujourd'hui est le prior de demain*
- Natures de l'incertitude
- Cohérence probabiliste
- Théorèmes asymptotiques
- Théorie de la décision statistique
- Psychologie cognitive

Estimateurs bayésiens et autres inférences

Toutes les inférences sont obtenues à partir du posterior $[\theta|y]$ (Sophie)

- best guest? MAP, Posterior mean $\hat{\theta}(Y) = \int_\theta \theta[\theta|Y]d\theta$,

posterior variance, etc.

- [a,b] intervalle bayésien de crédibilité à 70% (le plus court)

$$[a < \theta < b|y] = 0.7$$

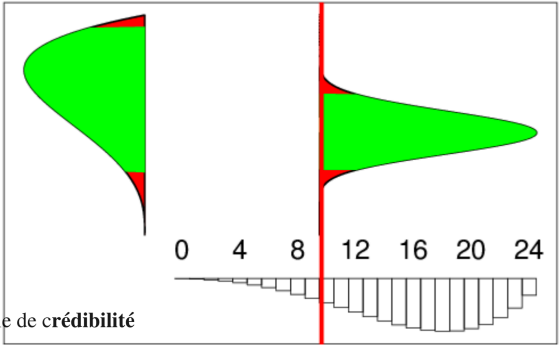
- décision sous critère d'utilité $u(\theta, d)$

$$d^\star = ArgMin_d \int_\theta u(\theta, d)[\theta|Y]d\theta$$

- prédictive et prévisions probabilistes

$$[Y_{new}|y] = \int_\theta [Y_{new}|\theta] \times [\theta|Y = y]d\theta$$

Exemple: conjugaison beta-binomiale $[\theta|Y] \propto [\theta, Y]$



Intervalle de **crédibilité**

axe vertical: θ

axe horizontal: $Y = y$

La loi a priori

La conjugaison normal-normale

- $Y|\mu,\sigma \sim dnorm(\mu,\sigma^{-2})$:

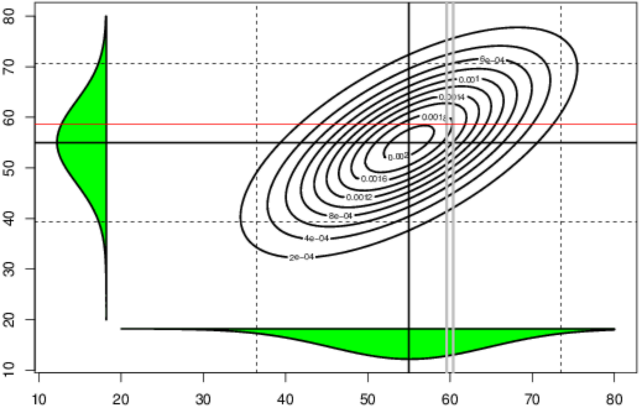
$$[Y\,|\mu,\sigma] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(Y-\mu)^2}{2\sigma^2}\right)$$

- $\mu|\sigma \sim dnorm(m_0, s_0^{-2})$
- $\mu|Y,\sigma \sim dnorm(m_1, s_1^{-2})$

$$s_1{}^{-2} = s_0{}^{-2} + \sigma^{-2}$$
$$s_1{}^{-2} m_1 = s_0{}^{-2} m_0 + \sigma^{-2} y$$

- Application: $\sigma = 5, m_0 = 55, s_0 = 8$

Exemple: conjugaison normale-normale
$[\mu, Y] = [Y|\mu] \times [\mu]$



axe vertical: μ

axe horizontal: $Y = y$

Le prior, un nouvel objet pas compliqué ?

- Souvent l'homme de l'art ont une opinion préalable à l'expérience à propos des inconnues.
- Encoder cette opinion à l'aide d'une loi a priori!
- la loi a priori peut être vague, multiple pour refléter diverses opinions, informative, non informative selon divers critères
- le prior peut souvent s'interpréter grâce une taille d'échantillon équivalent de données virtuelles.
- En plus du support des observations, l'approche bayésienne demande de probabiliser (conjointement) l'espace des inconnues,... mais tous les résultats d'estimation fréquentistes peuvent être retrouvés comme des cas particuliers de l'inférence bayésienne.

Distributions *a priori* informatives (Olivier)

`-> Emily`

Le processus d'élicitation : dialogue entre un expert et le statisticien pour encoder les jugements sous forme probabiliste.

- difficile : langage
- imprécis : quantitatif à partir de qualitatif
- robustesse : temps limité, tâche complexe

Une analyse de robustesse est nécessaire...

On dispose quelquefois de données historiques ou de données de phénomènes similaires

Distributions *a priori* NON-informatives (Olivier)

-> Emily

Un refus de s'engager sur la construction d'un prior pour un tas de raisons plus ou moins bonnes, mais. . .

PAS DE CONSENSUS sur ce que signifie l'ignorance *a priori*

- invariance par rapport à certaines transformations
- invariance du prior par reparamétrisation
- laisser parler les données = recours à une mesure de désordre (entropie)
- . . .

La loi a posteriori: $[\theta|y] \propto [y|\theta] \times [\theta]$

Comment obtenir la loi a posteriori? le bayésien computationnel! (Samuel)

- Avancées récentes extraordinaires des algorithmes de simulation Monte Carlo pour l'inférence!
- Outils commodes disponibles pour le praticien: BUGS & Jags, STAN, Greta, Nimble,... (Matthieu, Sophie) `-> Emily`
- De plus en plus de travaux bayésiens appliqués à cause de la facilité computationnelle. Les calculs théoriques d'intégration ou de transformation sont remplacés par des calculs empiriques sur échantillons générés.
- Des alternatives bayésiennes intéressantes aux fonctions standard de R: lm, glm, lmer, etc. mais aussi de nouvelles fonctionalités originales.
- Un effet de mode, pourquoi pas?

Et après?

Science et subjectivité

- Deux statisticiens face au même jeu de données aboutiront généralement à des conclusions différentes!
- Les questions scientifiques donnent lieu à des contreverses. Travail d'une communauté avec débat et reproductibilité pour accepter une interprétation.
- En statistique bayésiennes, les opinions a priori non *peremptoires* s'inclinent et convergent devant l'accumulation des données.

S'aventurer hors des sentiers battus... du modèle normal

- La loi normale est la pierre angulaire de la statistique fréquentiste (TCL asymptotique)
- Bayes pour l'étude de comportement avec peu de données
- Modélisation plus flexibles pour données Zéro inflatées, données "non missing at random", queues lourdes, outliers, etc...
- les modèles hiérarchiques: représenter les incertitudes et les quantifier. Tirer le meilleur parti de l'information disponible. (David)
- Les modèles simples sont difficiles à justifier en Bayes, mais les modèles complexes deviennent marginalement bien plus faciles à construire et à manier.

Just do it!

Si vous coincez...

- Réviser les notions d'intervalle de confiance *fréquentiste* et de *test*
- Revoir les lois de la famille exponentielle : normale, gamma, binomiale, beta.
- Dessiner sous R pour le modèle beta-binomial:
  - prior, vraisemblance, posterior, prédictive pour $n = 24, y = 10$, $a = 3, b = 2$
  - la collection de posteriors $[\theta|y]$
    - pour y allant de 0 à $n = 24$
    - pour n allant de 12 à 120, $\frac{y}{n}$ maintenu à $\frac{10}{24}$
- Aller surfer sur le blog de *Christian Robert*, sur celui d'*Andrew Gelman*

Bayesians versus Frequentists

Considérer les paramètres comme des variables aléatoires. . .

Et utiliser des priors non informatifs, tant qu'à faire. . .

- Hérésie fréquentiste: noter $f_\theta(y)$ plutot que $f(\theta|y)$
- Comment interpréter une probabilité?
  - incertitude épistémique
  - incertitude de répétabilité expérimentale
  - incertitude de variabilité *génétique*
- En bayésien θ est fixe et spécifique au problème
- Et si vous étiez un bayésien sans le savoir?
  - Q'avez vous compris du test d'hypothèse, type $\theta > 0$?
  - la mauvaise interprétation de l'intervalle de confiance

Paradoxe fréquentiste

- Le principe de vraisemblance
  - Si échantillon de $n = 100$ personnes , et observation $Y = 23$ alors $\hat{\theta}_1 = \frac{y}{n}$ sans *bias* (modèle binomial)
  - si observation de $N = n$ personnes interrogées jusque $y = 23$ alors l'estimateur sans approprié au modèle negatif binomial est $\hat{\theta}_2 = \frac{y-1}{n-1}$

Les écoles bayésienne et fréquentiste utilisent cet exemple pour défendre chacune leur *juste* interprétation.

En fait, on observe souvent N et Y jusque plus de ressources: difficile à modéliser en fréquentiste, facile en bayésien! Pour l'échantillonnage binomial à $n = 100$ fixé, l''estimateur fréquentiste $\hat{\theta}_3 = \frac{y}{n}$ si Y est impair et $\hat{\theta}_3 = \frac{y}{n+1}$ si Y est pair, est biaisé, ... mais donne la même valeur à l'estimation que l'estimateur, non biaisé, $\hat{\theta}_1$!

Annexe: Loi binomiale négative

- tirages avec succès de proba θ.
- Combien de tirages N jusqu'à obtention de y succès? (y fixé)?
- Appelons X le nombre d'échecs; $N = X + y$.

$$[X = k] = \left(\begin{array}{c} k + y - 1 \\ y - 1 \end{array} \right) \theta^y (1 - \theta)^k$$

- Il faut tirer obtenir $y - 1$ parmi $N - 1$ et terminer par un succès

$$[X = k] = \left(\left(\begin{array}{c} N - 1 \\ y - 1 \end{array} \right) \theta^y (1 - \theta)^{N - y - (y - 1)} \right) \times \theta$$

- $\hat{\theta} = \frac{y-1}{N-1}$ est sans biais pour la binomiale négative