# Introduction au choix des distributions *a priori* en inférence bayésienne

M. L. Delignette-Muller
VetAgro Sup - LBBE

5 mai 2019




# Quelle loi *a priori* ?

Une des spécificités de la statistique bayésienne : l'utilisation de lois *a priori*.

Mais comment définir ces lois *a priori* ?

# La probabilité en terme de fréquence

En statistique fréquentiste, le terme **"probabilité"** est associé à la **fréquence de réalisation d'un évènement aléatoire**.
**Signification parfois appelée objective**

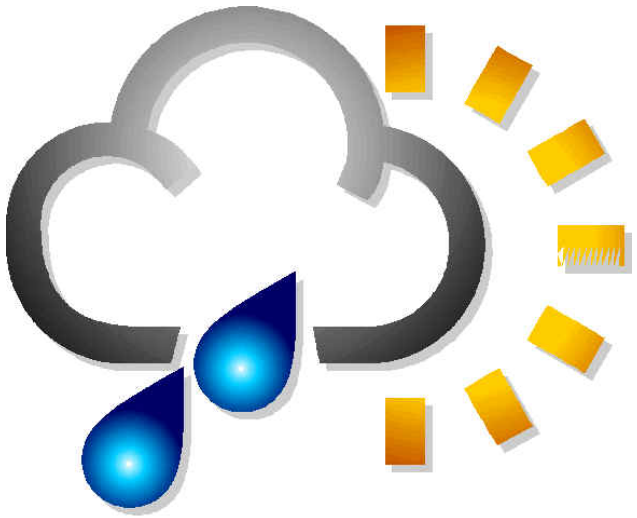
*Probabilité pour un mouton tiré au hasard d'être noir =*
*fréquence de moutons noirs ?*

# La probabilité en terme de pari

Le terme **"probabilité"** peut aussi être associé au **degré de croyance en la réalisation d'un évènement**, lié à un défaut de connaissance.

**Signification parfois appelée subjective**

*Probabilité de pluie demain,*
*ou avec quelle confiance on parie qu'il pleuvra demain ?*



# Cadre conceptuel de l'inférence fréquentiste

En statistique fréquentiste, **les paramètres d'un modèle sont supposés fixes mais inconnus**.

- On ne leur associe pas de loi de probabilité.
- On raisonne sur la probabilité des données connaissant les paramètres (vraisemblance, p-value).
- On ne peut jamais conclure sur la probabilité d'un paramètre d'appartenir à un intervalle ni sur la probabilité d'une hypothèse (égalité d'un paramètre à une valeur donnée), *même si parfois on aimerait bien !*
  C'est d'ailleurs une des sources des trop nombreuses erreurs d'interprétation des résultats des tests d'hypothèse.

# Cadre conceptuel de l'inférence bayésienne

En statistique bayésienne, on s'autorise à raisonner sur la **distribution de probabilité des paramètres d'un modèle** qui sont donc considérés comme aléatoires.

- On associe aux paramètres une loi de probabilité "au sens subjectif".
- On estime une loi *a posteriori* à partir des données et d'une loi *a priori* (ce que l'on savait avant de regarder les données).
- On peut raisonner de façon intuitive sur la probabilité d'un paramètre d'appartenir à tel intervalle.

Mais il n'est pas toujours facile de définir une loi *a priori*.

# Les lois *a priori* avant les MCMC

## Calcul de la loi *a posteriori*

$$P(\theta|Y) \propto P(Y|\theta) \times P(\theta)$$

- calcul analytique de la loi *a posteriori* limité aux lois conjuguées
- lois conjuguées disponibles uniquement sur quelques modèles simples
- information *a priori* pas forcément facile à modéliser sous la forme d'une loi conjuguée

**Limitation importante de l'inférence bayésienne et choix des lois parmi les lois conjuguées.**

# Utilisation classique de lois conjuguées

## Définition d'une loi conjuguée

$f(\theta)$ est une loi conjuguée pour la fonction de vraisemblance $f(y|\theta)$ si $f(\theta|y)$ est de la même forme que $f(\theta)$

**Quelques exemples de lois conjuguées**

| $f(y\|\theta)$ | $f(\theta)$ |
|---|---|
| Poisson | gamma |
| normale | normale pour $\mu$ et gamma pour $\tau$ |
| exponentielle | gamma |
| binomiale | beta pour p |
| négative binomiale | beta pour p |
| multinomiale | Dirichlet pour $\{p_1, p_2, \cdots, p_k\}$ |

# Les lois *a priori* à l'ère des MCMC

L'utilisation des MCMC (Markov Chain Monte Carlo) pour estimer par échantillonnage la loi *a posteriori* a donné un nouvel élan à la statistique bayésienne

- algorithme de Metropolis-Hastings : 1970
- échantillonneur de Gibbs : 1984
- projet BUGS (Bayesian inference Using Gibbs Sampling) : 1989

**L'utilisation des lois conjuguées, même si elle reste une pratique courante, n'est plus une nécessité. Le champ des possibles est devenu beaucoup plus large.**

# Loi informative ou non informative ?

- L'utilisation de lois informatives est parfois vue comme un défaut majeur de l'inférence bayésienne : source de subjectivité.
  D'où l'utilisation courante de lois *a priori* non informatives, pour se rapprocher de la statistique fréquentiste.
- D'autres revendiquent l'intérêt de pouvoir utiliser des lois informatives : permet de prendre en compte toute l'information dont on dispose pour estimer des paramètres lorsque le jeu de données est réduit.

**Y aurait-il deux types de "bayésiens", les objectivistes et les subjectivistes ?**

# Tout dépend de ce que l'on sait *a priori* et de la nécessité ou non d'ajouter cette information aux données

- L'utilisation de lois *a priori* non informatives est parfois plus délicate qu'on ne le croit.
- Dans les modèles biologiques, on sait souvent au moins un petit quelque chose et il vaut généralement mieux utiliser une loi *a priori* que fixer arbitrairement un paramètre difficile à estimer à partir des seules données ?
- Mais la définition de lois *a priori* informatives est aussi un peu délicate.

**Une démarche raisonnable consiste à choisir, pour chaque paramètre, entre lois a priori non informative, vaguement informative ou informative, suivant ce que l'on sait a priori et de la quantité de données dont on dispose.**

# Lorsqu'on ne sait rien ou presque

Est-ce simple de définir une loi non informative ?

# Exemple de l'estimation d'une proportion : modèle binomial

Ex. : estimation de la prévalence d'une maladie

## Pratique courante

Utilisation d'une loi *a priori* conjuguée plate :
$beta(1,1) = uniforme(0,1)$.
Soit $y$ observations positives sur $n$ observations, loi *a posteriori* :
$beta(y+1, n-y+1)$

"Mais plat est-il synomyme de non informatif ?"
Que signifie "non informatif ?"
Qui apporte le moins possible d'information.
Pour laquelle l'estimation *a posteriori* se rapproche le plus possible de l'estimation par maximum de vraisemblance (MLE) ?
Retenons cette dernière définition pour explorer cet exemple.

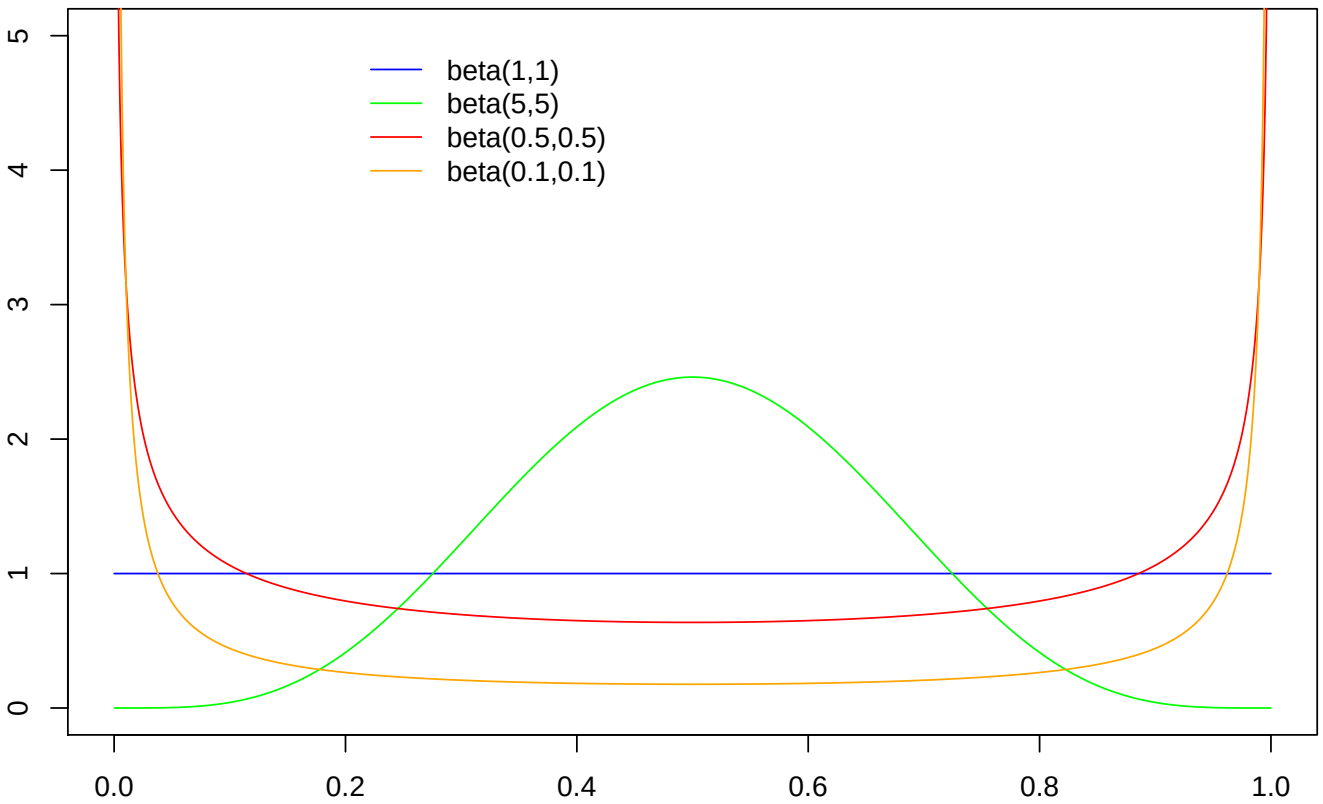
# La loi beta plate est-elle la plus non-informative ?

Plaçons nous dans un cas restreint de loi $beta(c, c)$ symétrique à un paramètre ($\alpha = \beta = c$).

- estimation MLE : $\frac{y}{n}$
- moyenne de la loi *a posteriori* $beta(y + \alpha, n - y + \beta) = beta(y + c, n - y + c)$ : $\frac{y+c}{n+2c}$ comme si on avait $2c$ observations en plus, dont la moitié de positives.

Donc plus $c$ est grand et plus la loi est informative.
On s'approche du MLE quand $c \to 0$.
On pourrait en théorie prendre $beta(\epsilon, \epsilon)$ avec $\epsilon$ très petit.
On utilise souvent la loi $beta(\frac{1}{2}, \frac{1}{2})$ (loi de Jeffrey, invariante par reparamétrisation).
La loi $beta(\epsilon, \epsilon)$ n'est pas plate mais de variance élevée.
"Donc plat n'est pas synomyme de non informatif."

# La première intuition peut être trompeuse !

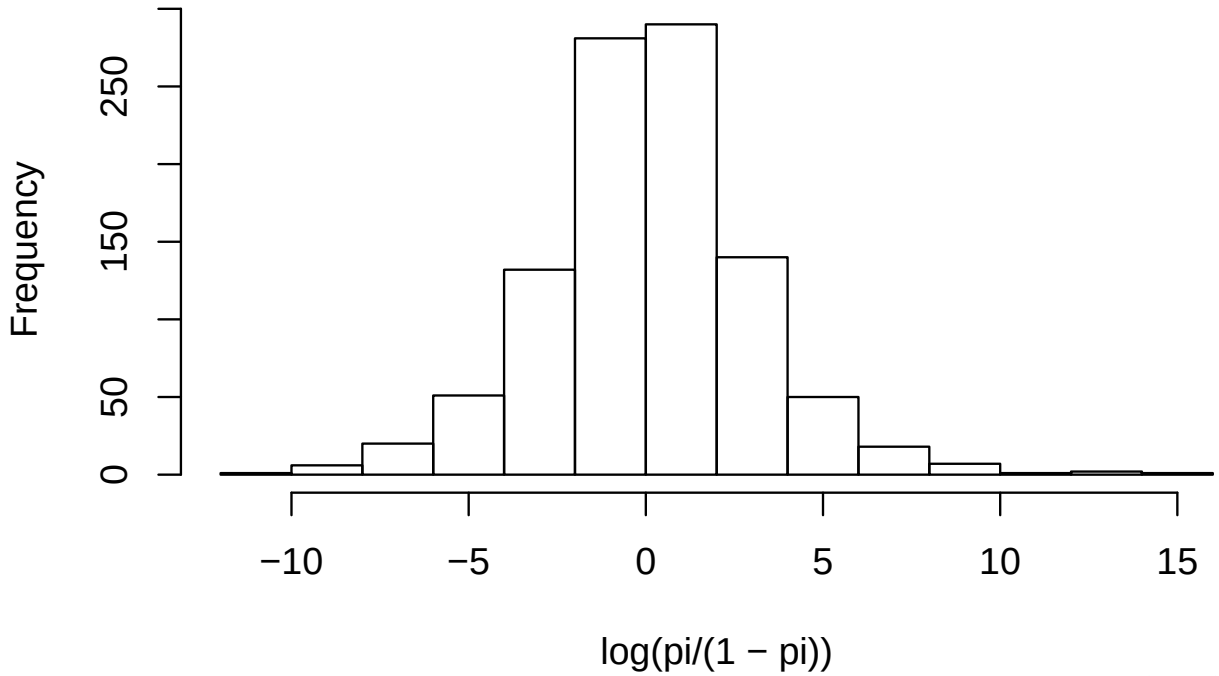



Il n'était pas forcément intuitif d'imaginer qu'il faille mettre un poids très important vers 0 et 1 pour rendre la loi moins informative, ceci étant dû au support compact de la loi.

# Plus intuitif si on change de paramétrisation

Visualisation de la loi de Jeffrey ($beta(\frac{1}{2}, \frac{1}{2})$) en échelle logit

```
> pi <- rbeta(1000, 0.5, 0.5)
> hist(log(pi/(1-pi)), main = "")
```



# Choix d'une loi *a priori* vaguement informative

Il est rare que l'on ne dispose pas d'une information au moins vague liée aux contraintes du modèle.
Le biologiste saura souvent dire qu'en dehors d'une certain gamme, les valeurs de tel paramètre ne sont pas réalistes.
**Une tendance actuelle est d'utiliser des lois vaguement informatives plutôt que non informatives.**
MAIS comment définir une loi informative ou vaguement informative à partir d'une information *a priori*
(**donc sans se servir des données bien sûr**) ?

*Traiter les questions 1, 2 et 3 de l'énoncé d'exercices.*

# Exemple de l'estimation de la prévalence d'une maladie rare

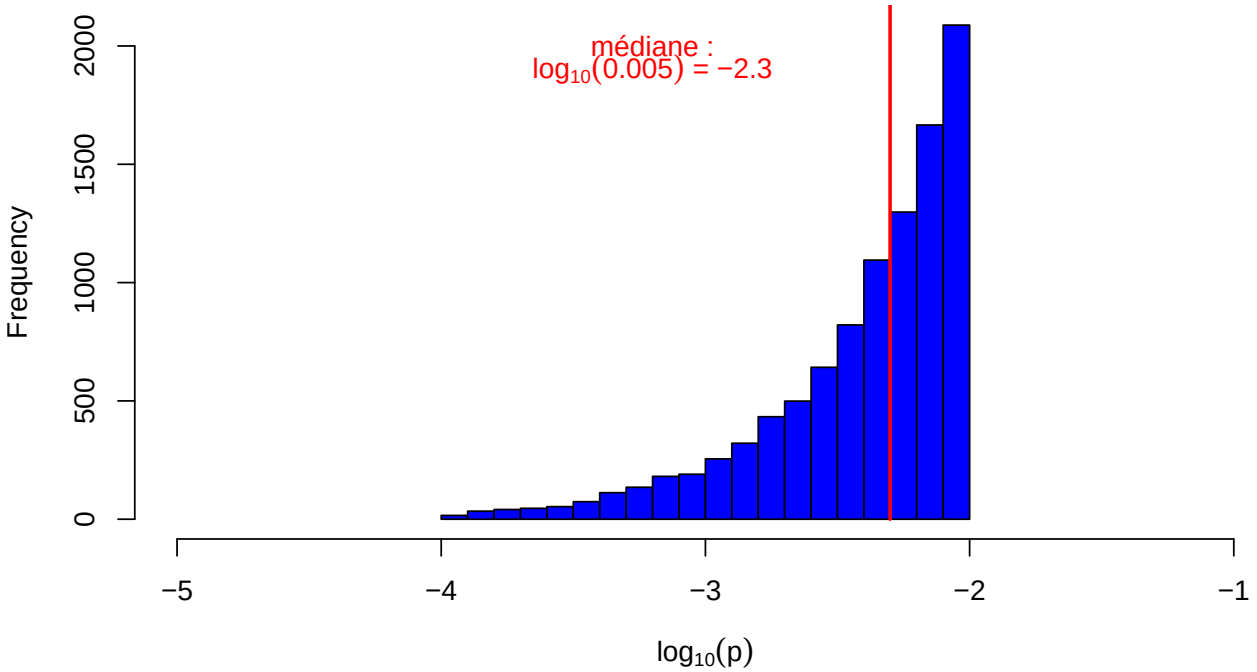
- Loi non informative $beta(\frac{1}{2}, \frac{1}{2})$ à préférer à la loi uniforme $beta(1, 1)$ surtout si $p$ est proche de 0 ou de 1 (ex. : estimation de la prévalence d'une maladie rare).
- Mais si l'on sait à l'avance que la maladie est rare, pourquoi ne pas définir une loi *a priori* vaguement informative en mettant un point fort uniquement du côté probable ?

# Choix d'une loi vaguement informative : échelle, support ?

Ex. : on sait que la prévalence d'une maladie rare est à peu près comprise entre 1 cas pour 10 000 et 1 cas pour 100. Quelle loi *a priori* pour modéliser cette information ?

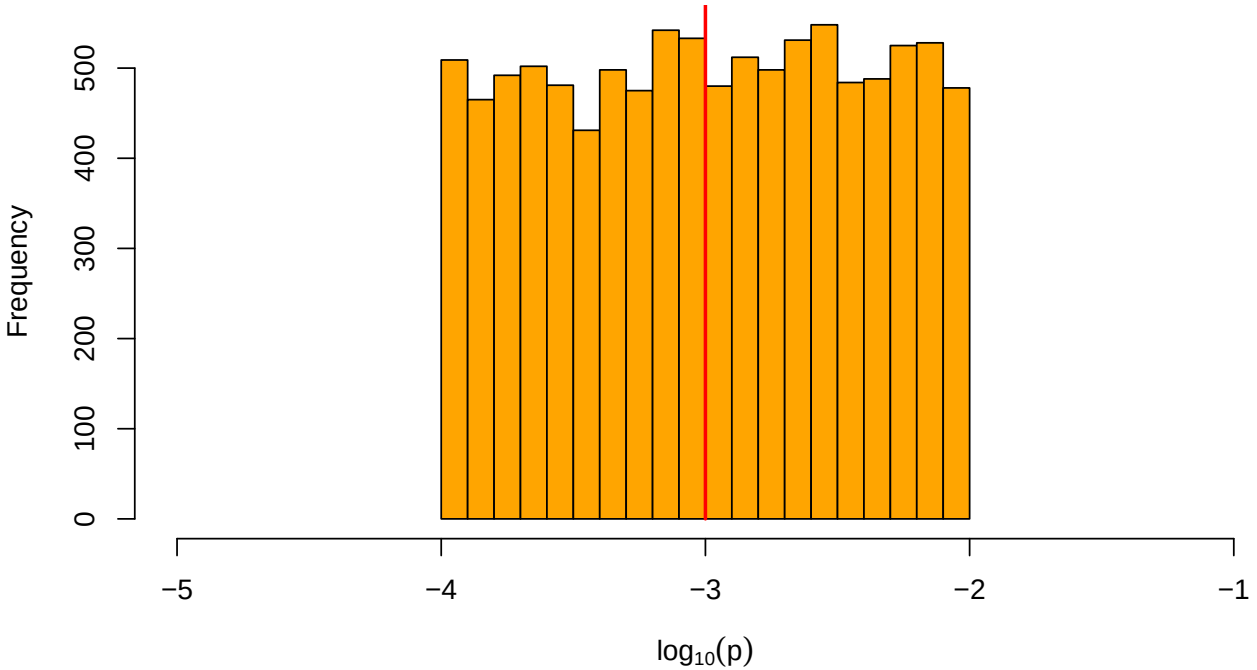
- loi uniforme sur $p$
  (ex. : $p \sim uniforme(10^{-4}, 10^{-2})$)
- loi uniforme sur $ln(p)$ ou $logit(p)$ (très proches pour $p < 10^{-2}$)
  (ex. : $log_{10}(p) \sim uniforme(-4, -2)$)
- loi normale sur $logit(p)$
  (ex. : $logit(p) \sim normale(\mu = -3ln(10), \sigma = \frac{ln(10)}{2})$)

# Loi 1 : $p \sim uniforme(10^{-4}, 10^{-2})$



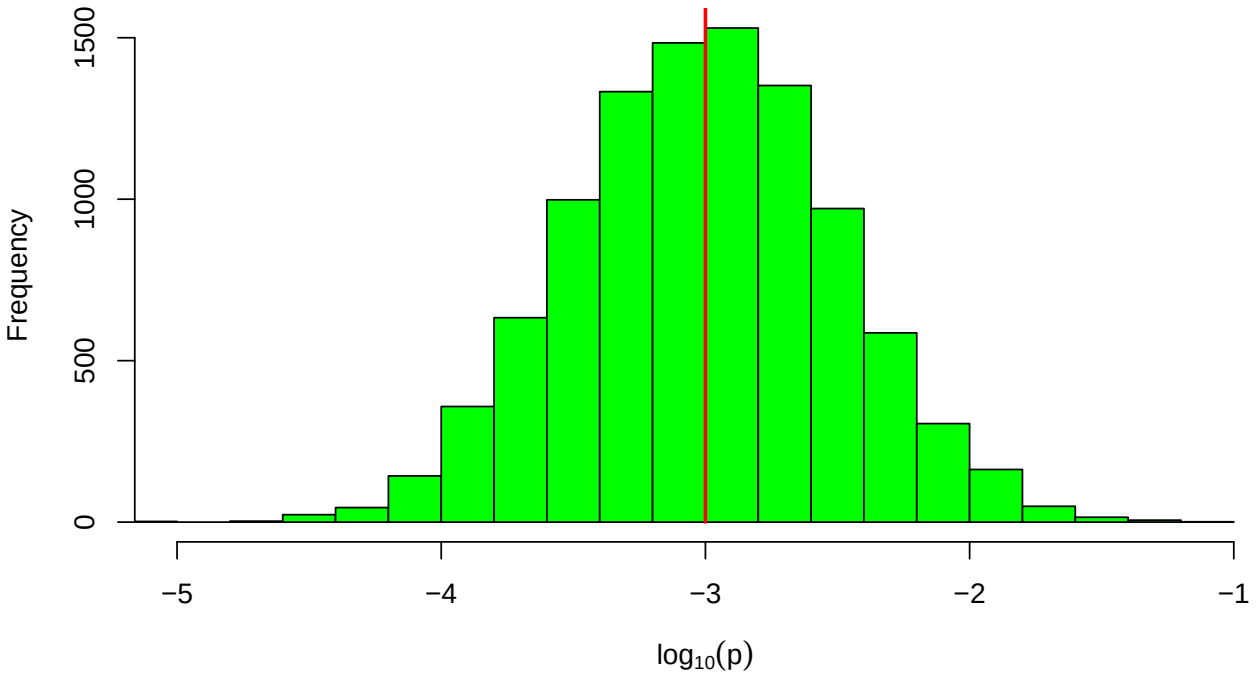

$Pr(p < 10^{-3}) = 0.09$ : il est dix fois plus probable que $p > 10^{-3}$.
Choix semble-t-il peu raisonnable !

# Loi 2 : $log_{10}(p) \sim uniforme(-4, -2)$



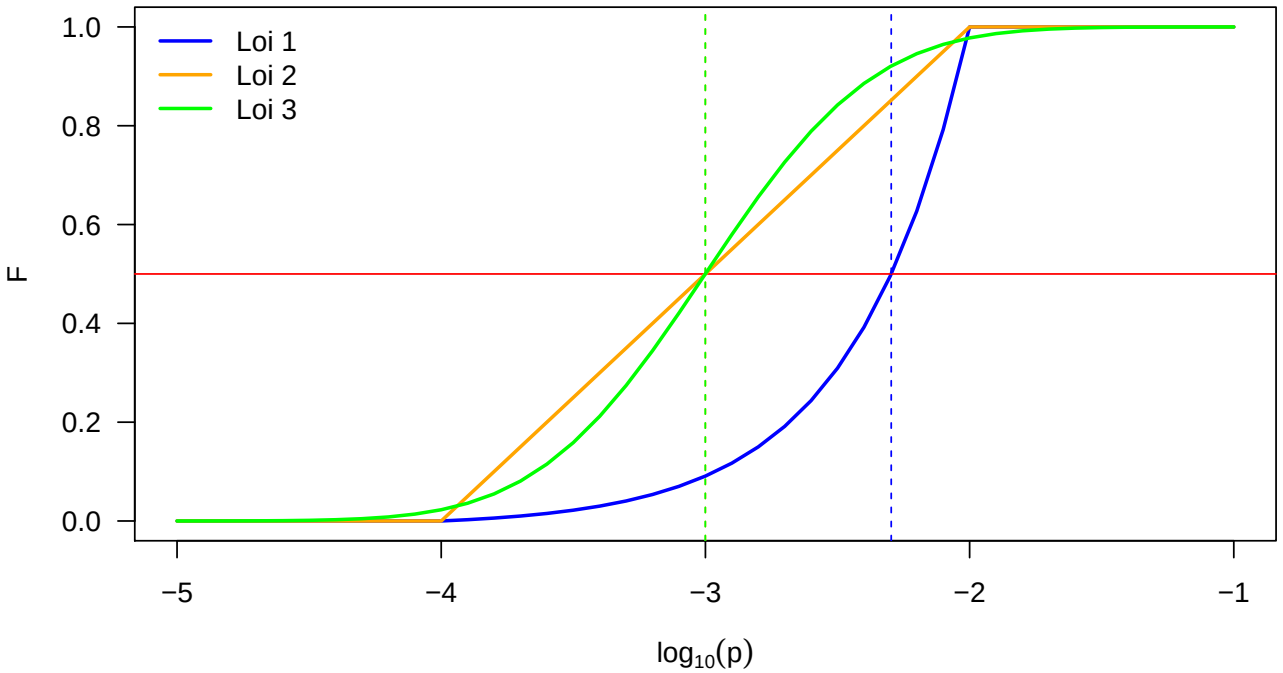

Choix plus raisonnable mais présentant l'inconvénient de fixer des limites strictes à $10^{-4}$ et $10^{-2}$.

# Loi 3 : $logit(p) \sim normale(\mu = -3ln(10), \sigma = \frac{ln(10)}{2})$




Choix le plus raisonnable ?

# Comparaison des 3 lois en courbe des fréquences cumulées




La définition d'une loi *a priori* ne peut pas se limiter à la définition de son support. Il est indispensable de définir (et de visualiser) sa forme.

# Quelques conseils pour le choix d'une loi informative ou vaguement informative

- **ATTENTION au choix de l'échelle utilisée !**
- **Quel type de loi, quel support ?**
  - **Lois peu informatives :**
    on peut en général utiliser une loi uniforme large sur le paramètre exprimé sur une échelle appropriée, en vérifiant que le support de la loi ne contraint pas trop la loi *a posteriori*.
  - **Lois plus informatives :**
    il est plus raisonnable d'utiliser une loi avec queues de distribution,
    éventuellement tronquée pour éviter les valeurs complètement irréalistes.

# Et comment obtenir une information a priori fiable

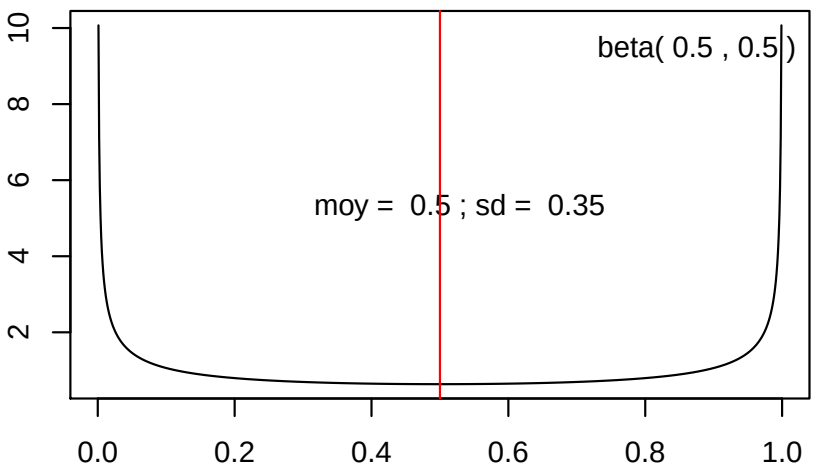
Comment recueillir une information *a priori* ?
A partir de quel questionnement de l'expert peut-on construire une loi *a priori* ?
Comment faire de l'"élicitation de priors" ?

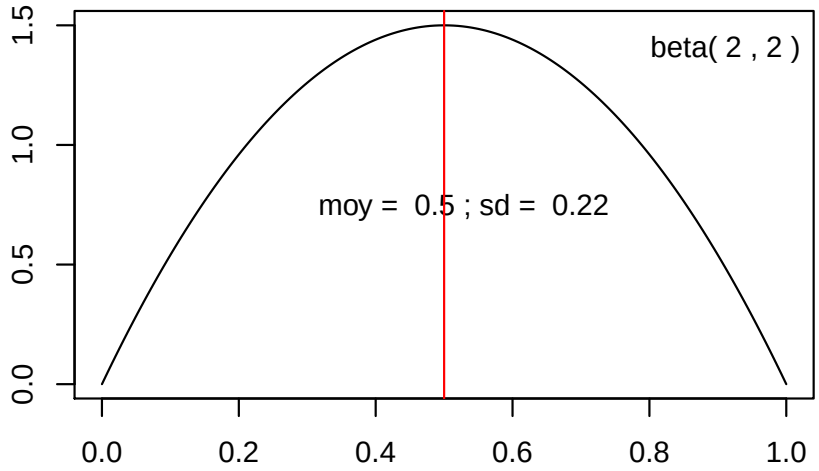
*Traiter les questions 4, 5 et 6 de l'énoncé d'exercices.*

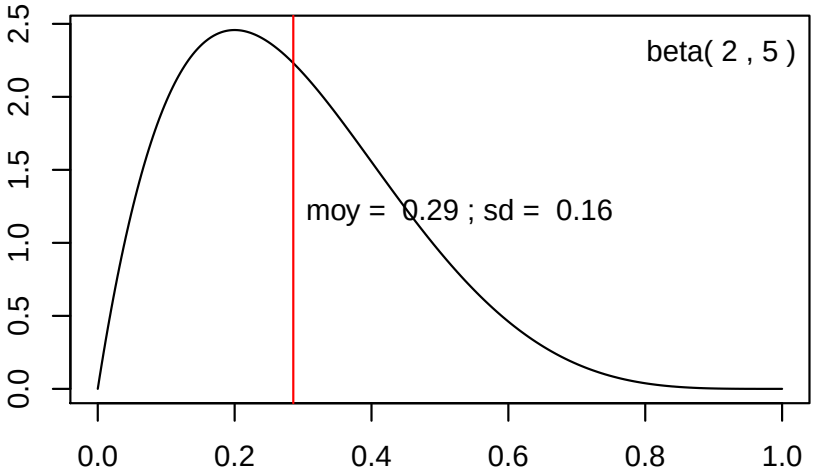
# Que sait-on estimer correctement ?

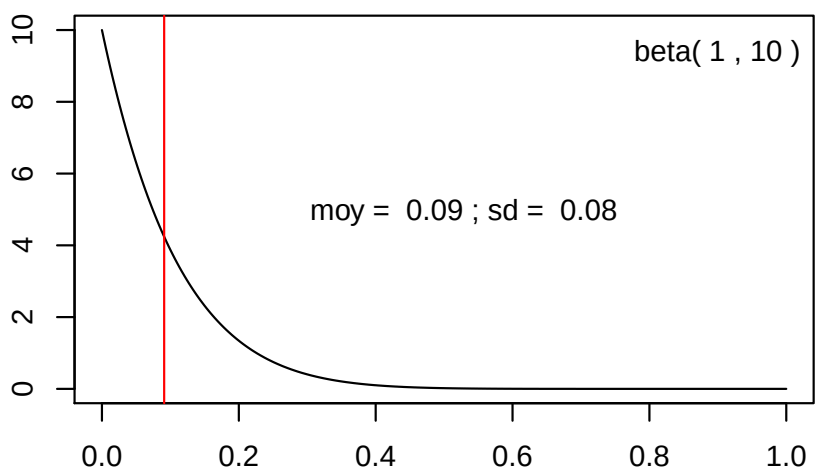
- On sait assez bien estimer une proportion.
- On sait bien estimer une tendance centrale sur une distribution symétrique, mais on donne généralement une estimation biaisée de la moyenne (vers la médiane) pour les distributions dissymétriques.
- On sait mal estimer une variance ou un écart type. **Attention à l'utilisation des moments !**
- On sait mieux estimer des quantiles (à partir d'un questionnement simple) mais on sous-estime généralement la crédibilité associé à un intervalle (excès de confiance)

# Ex. : difficulté d'éliciter les moments d'une loi bêta



# Questionnements possibles de l'expert

Elicitation de quantiles, si possible plus que nécessaires pour fixer les paramètres de la loi.

**Idées de questionnement :**

- Définition de la **gamme des possibles** souvent sous la forme d'un intervalle de crédibilité à 95%
- Définition de la **médiane** (valeur pour laquelle l'expert dirait qu'on a autant de chance d'être en-dessous ou au-dessus)
- Eventuellement définition des quartiles à 25 et 75% (même question si on suppose que la valeur est au-dessus de la médiane par ex.)

A partir de quantiles on peut paramétrer une loi en minimisant une distance entre les $F$ théoriques et élicités pour chaque quantile $Q$ (ex. distance quadratique de Cramer-von-Mises ($\sum (F_{theo} - F_{elicit})^2$)).

# Outils d'aide à l'élicitation de lois *a priori*

**Package R proposant notamment une application shiny pour ajuster diverses lois classiques sur des quantiles élicités**, décrits sous la forme d'un vecteur de quantiles $Q$ et du vecteur de fréquences cumulées associées $F$.

## SHELF

`https://cran.r-project.org/web/packages/SHELF/index.html`

**Interface web liée à cet outil : Match Tool**

`http://optics.eee.nottingham.ac.uk/match/uncertainty.php`

**Liens vers d'autres outils disponibles**

`http://www.expertsinuncertainty.net/Software/tabid/4149/Default.aspx`

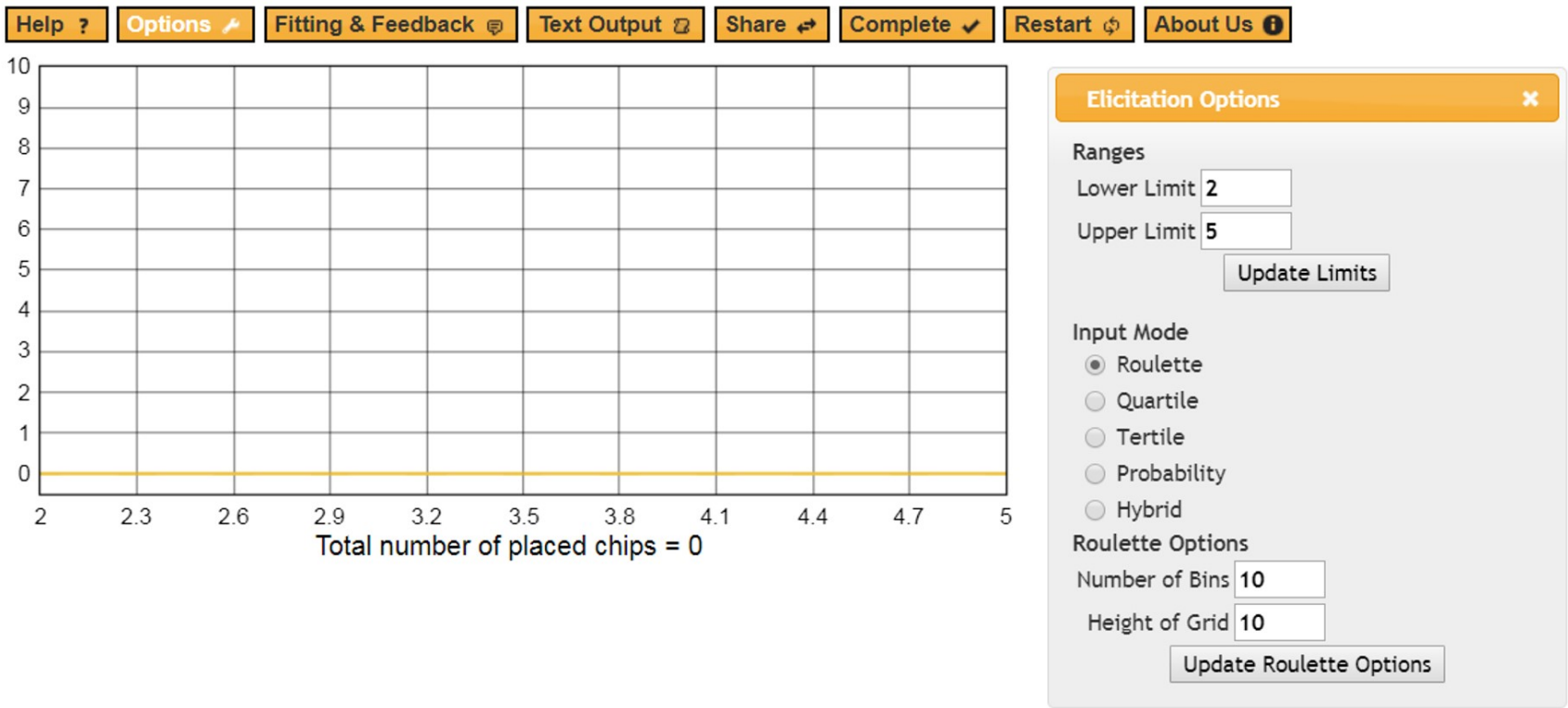
# Exemple d'utilisation de SHELF

Ex. de code utilisant SHELF : ajustement d'une loi bêta pour la prévalence de la maladie à partir d'une information *a priori* donnée sous la forme de la médiane (0.001) et d'un intervalle de crédibilité à 95% ([0.0001; 0.01]).

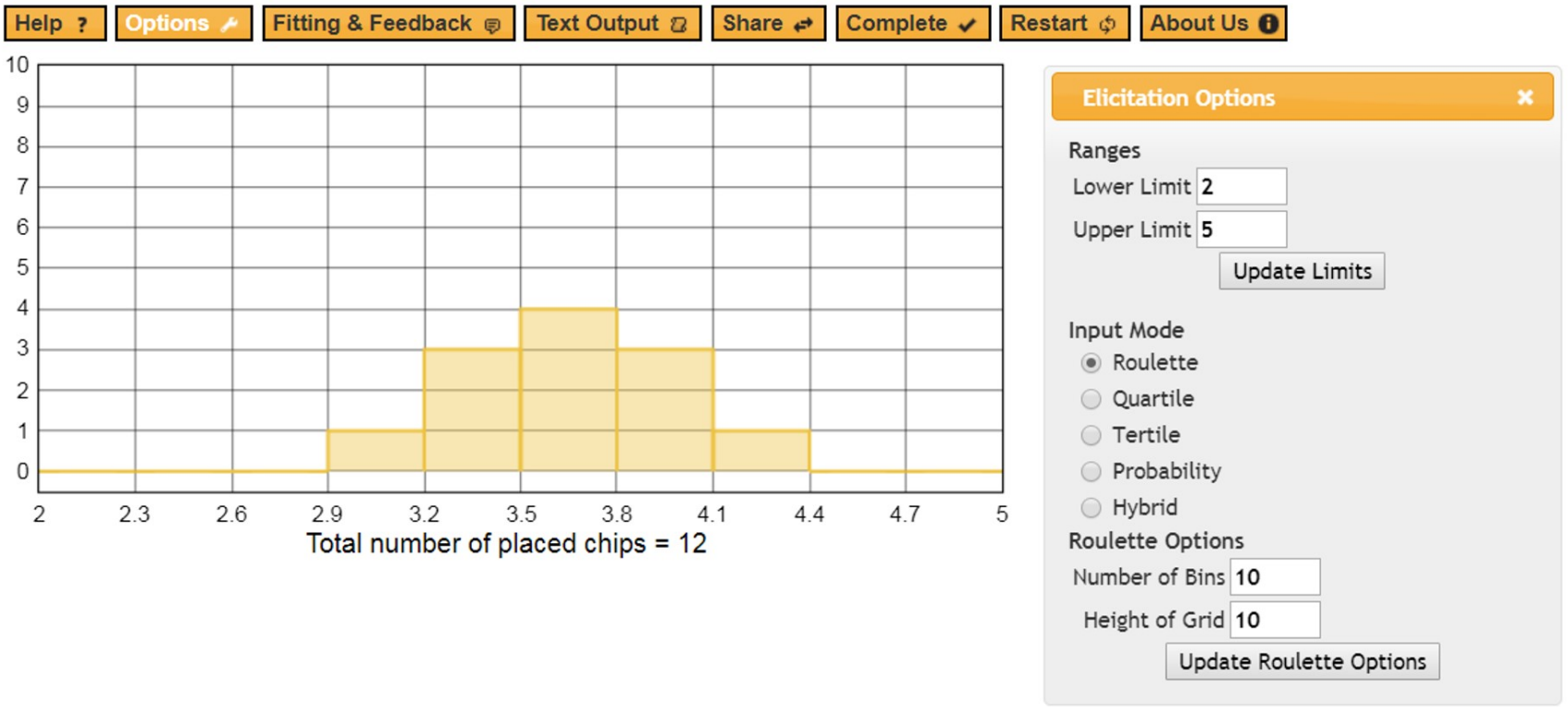
```
> require(SHELF)
> Q <- c(0.0001, 0.001, 0.01)
> F <- c(0.025,0.5,0.975)
> fit <- fitdist(vals = Q, probs= F, lower = 0, upper = 1)
> fit$Beta

    shape1   shape2
1 1.578272 1259.374
```

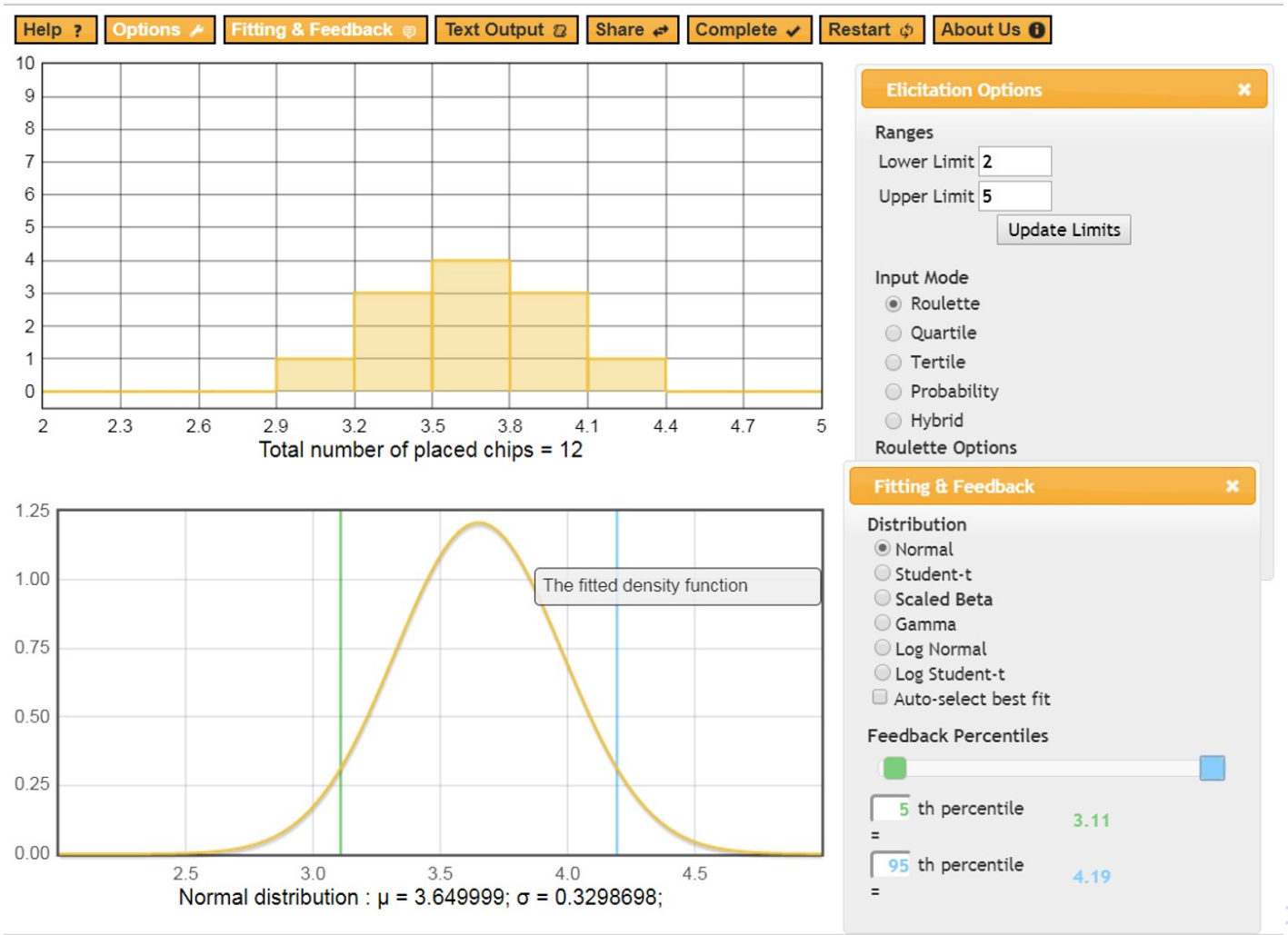
# Exploration de MATCH Tool sur un exemple d'élicitation du poids moyen garçons à la naissance



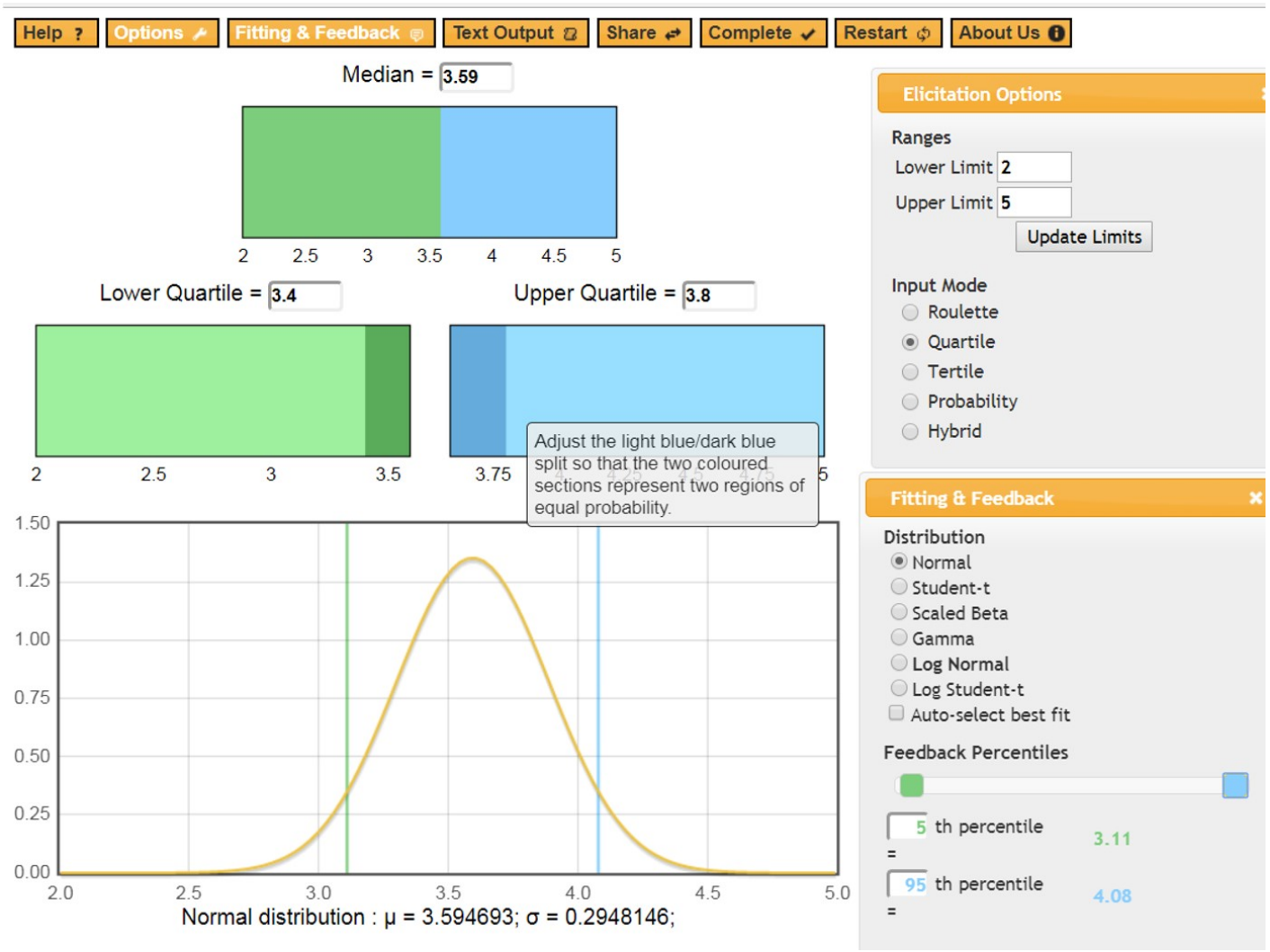
# Exploration de MATCH Tool : méthode de la roulette



# Exploration de MATCH Tool : Ajustement d'une loi normale



# Exploration de MATCH Tool : méthode des quartiles



# Quelques conseils pour définir des lois *a priori* vaguement informatives ou informatives

- Il est généralement préférable de raisonner sur les quantiles des distributions.
- Pour limiter le biais de centrage des experts on les interroge généralement d'abord sur les extrêmes puis sur la médiane.
- Pour corriger l'excès de confiance des experts on associe parfois les extrêmes élicités à des quantiles par ex. à 2.5 et 97.5%.

# Autres conseils généraux au sujet des lois *a priori*

- Il est important de visualiser les distributions *a priori* (avec JAGS il est très simple de faire tourner le modèle en Monte Carlo simple avant d'ajouter les données pour réaliser l'inférence).
- Il est important de confronter les distributions *a posteriori* aux distributions *a priori*.
- Il est important d'analyser la sensibilité de la loi *a posteriori* aux choix faits lors de la définition des lois *a priori*, notamment dans le cas de lois dites non informatives. ATTENTION, ceci n'est pas redondant avec le point précédent : même si la loi *a posteriori* est beaucoup moins dispersée que la loi *a priori*, cela ne garantit pas qu'elle soit peu sensible à cette dernière.

# Références utilisées


- Garthwaite P.H., Kadane J.B. and O'Hagan A.,2005. Statistical methods for eliciting probability distributions. *J. American Statistical Association*, 100, 680-701.
- Zhu M., 2004. The counter-intuitive non-informative prior for the Bernouilli family, *J. Statistics Education*.