

# Regression models for prediction (Part II Bayesian approaches)

Vincent Garin

# Recap

Standard linear regression model  $y = \mu + X\beta + e$

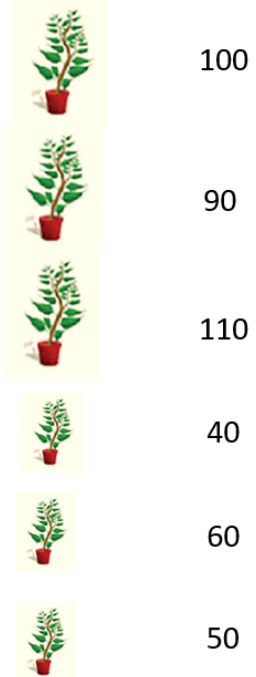
$$y = \mu + x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p + e$$

Genotype Phenotype

	L1	L2	L3	L4	L5	L6	L7
G1	AA	AA	AA	AA	AA	AA	AA
G2	CC	CC	CC	CC	CC	CC	CC
G3	AA	AA	AA	AA	AA	AA	AA
G4	AA	AA	AA	AA	AA	AA	AA
G5	CC	CC	CC	CC	CC	CC	CC
G6	AA	AA	AA	AA	AA	AA	AA
G7	CC	CC	CC	CC	CC	CC	CC
G8	CC	CC	CC	CC	CC	CC	CC

$$y = \begin{bmatrix} 100 \\ 90 \\ 110 \\ 75 \\ \dots \\ 80 \end{bmatrix}$$

$$X = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 1 & 2 & \dots & 0 \\ 2 & 2 & \dots & 1 \\ 0 & 0 & \dots & 2 \\ \dots & \dots & \dots & 2 \\ 0 & 0 & \dots & 1 \end{bmatrix}$$



# Recap

Standard linear regression model

$$y = \mu + X\beta + e$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$V(y|X) = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \sigma_e^2$$

Generalized least square model

$$y = \mu + X\beta + e$$

$$\beta_{GLS} = (X'V^{-1}X)^{-1}X'V^{-1}y$$





$$V(y|X) = \begin{pmatrix} v_{1,1} & & v_{1,n} \\ & \ddots & \\ v_{n,1} & & v_{n,n} \end{pmatrix}$$

Mixed models (GBLUP)

$$\underline{y} = X\underline{\beta} + Z\underline{u} + \underline{\epsilon} \quad u \sim N(0, G)$$

$$\epsilon \sim N(0, R)$$

**K**: Realized or molecular co-ancestry

	P	G	L1	L2	L3	L4	
	100	G1	AA	CC	CC	AA	~ Similar (50%)
	50	G2	CC	AA	CC	CC	
	?	G3	AA	CC	AA	CC	~ Similar (75%)
	?	G4	CC	AA	CC	AA	

	G1	G2	[...]	Gn
G1	1	0,42	0,58	0,48
G2		1	0,2	0,28
[...]			1	0,24
Gn				1

Regularisation methods (Rg, LASSO)

$$\hat{\beta}_{Ridge} = \underset{\beta}{\operatorname{argmin}} \left[ \sum_{i=1}^n (y_i - \mu - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 \right]$$

# Bayesian logic

---



Reverend Thomas Bayes (1702 - 1761)

*Essay Towards Solving a Problem in the Doctrine of Chances* - 1763 (Posthumous)

“The Bayes theorem is to the theory of probability what Pythagora’s theorem is to geometry” (Jeffreys Harold)



Pierre-Simon Laplace (1749 - 1827)

The French Newton

Bayesian interpretation of probabilities

# Bayes theorem – elements of probability (set theory)

---

The Bayes theorem is the probabilistic root of the Bayesian approach. The Bayes theorem is a rule about conditional probability.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(Y \cap X)}{P(Y)} = \frac{P(Y|X) * P(X)}{P(Y)}$$

where,

$X, Y$  are two events (realisation, hypothesis, observation).

$P(X|Y)$  is the conditional probability of the  $X$  event given that  $Y$  occurred.

$P(Y \cap X)$  is the joint probability of the two events (probability that  $Y$  and  $X$  are true).

$P(Y|X)$  is also a conditional probability that can be interpreted as the **likelihood** of observing  $Y$  if  $X$  happens.

$P(X)$  and  $P(Y)$  are the **prior** probability of observing  $X$ , respectively  $Y$  without conditions.

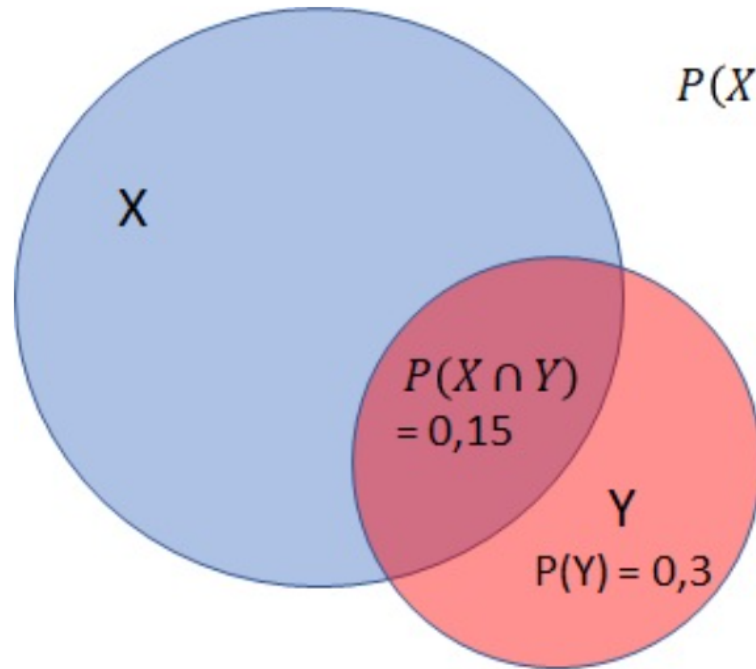
# Bayes theorem elements – Conditional probability

---

The conditional probability can be defined as an axiom of probability.

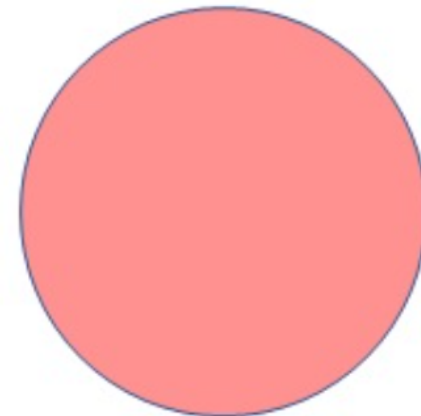
$$P(X \cap Y) = P(X|Y) * P(Y)$$

Another possibility is to use a graphical illustration. The conditional probability is defined as the quotient of the joint probability  $P(X \cap Y)$  where  $X$  and  $Y$  occur together and  $P(Y)$ .



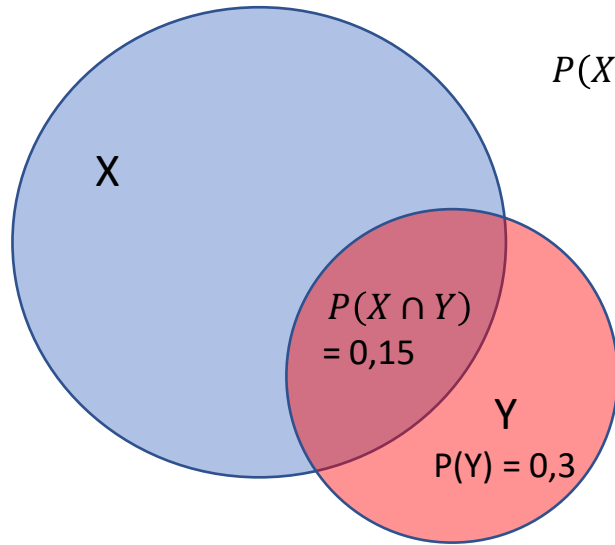
$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{0,15}{0,3} = 0,5$$

$$P(X|Y) = \frac{\text{Intersection}}{\text{Y}}$$

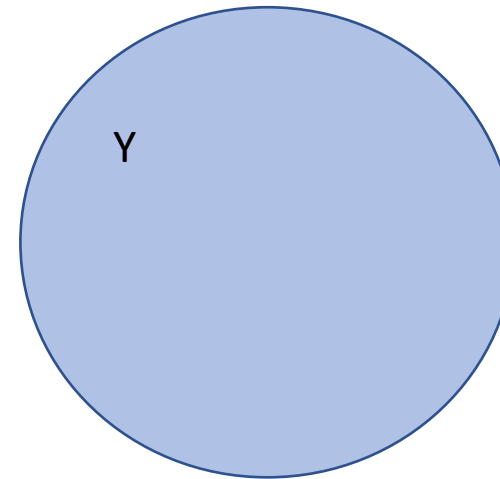


# Bayes theorem – elements of probability (set theory)

---

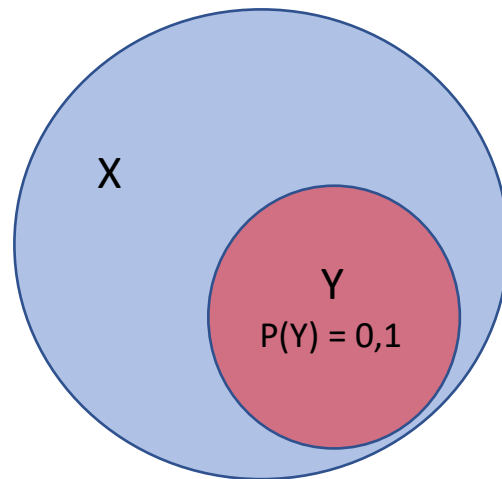
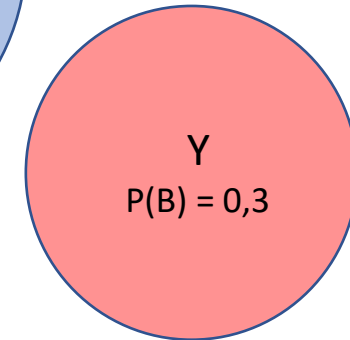


$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{0,15}{0,3} = 0,5$$



$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = ?$$

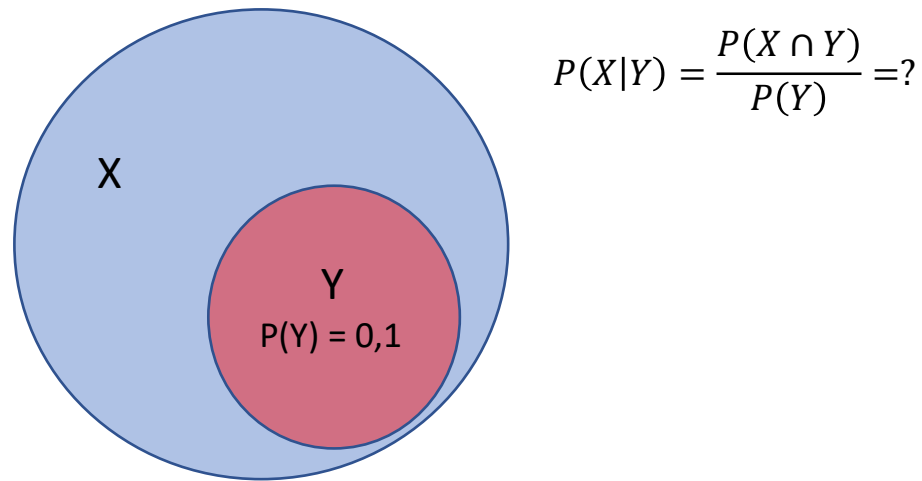
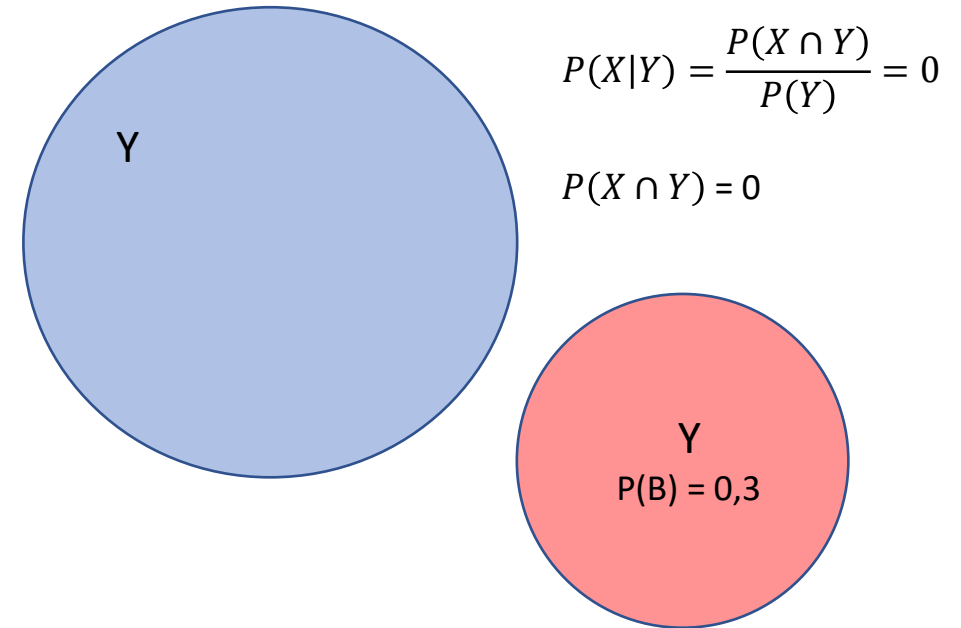
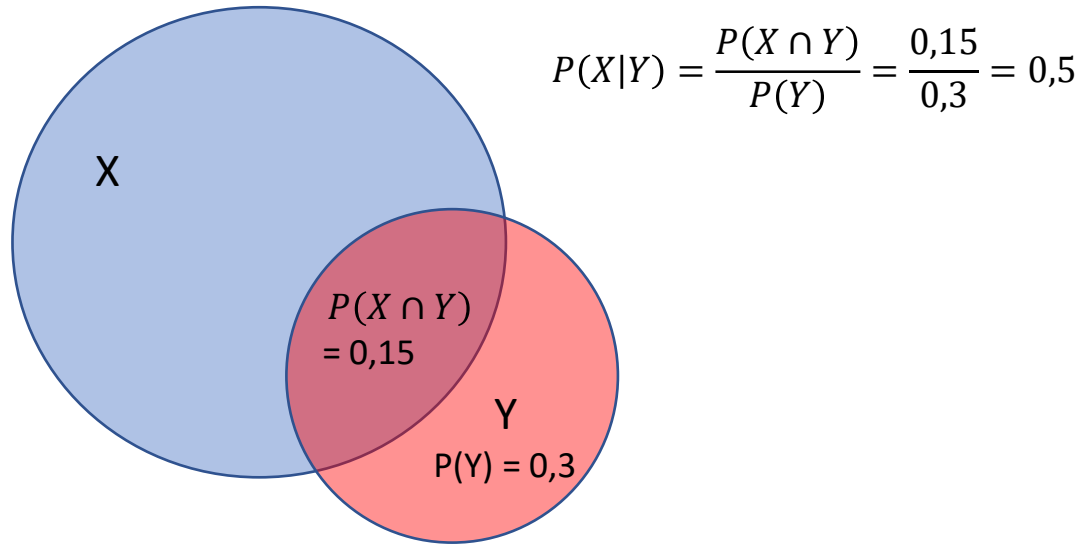
$$P(X \cap Y) = ?$$



$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = ?$$

# Bayes theorem – elements of probability (set theory)

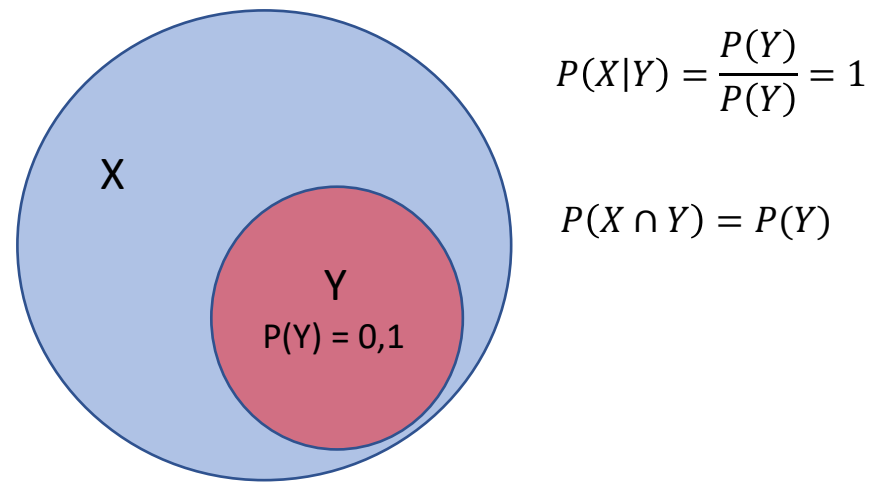
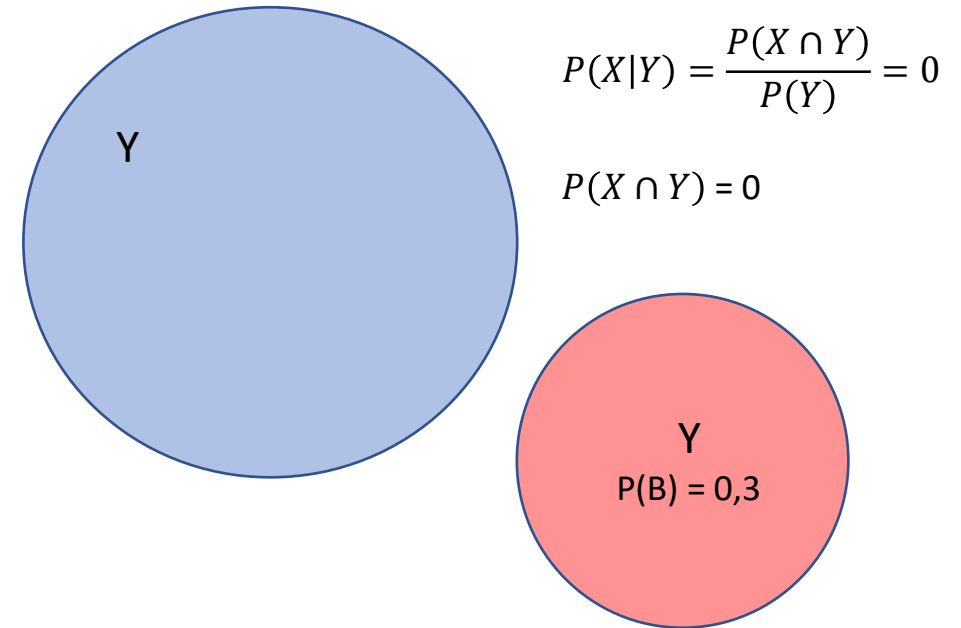
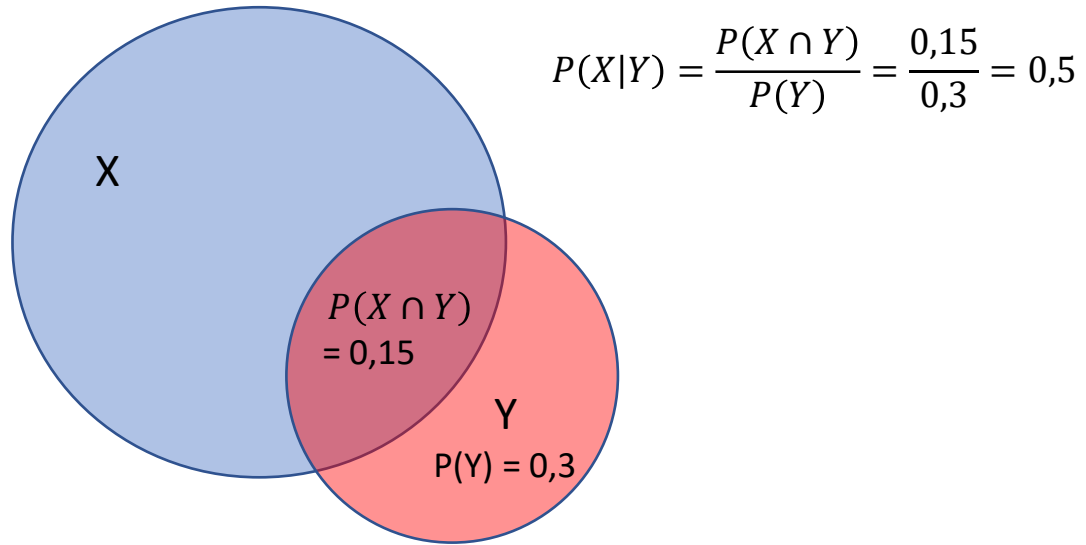
---





# Bayes theorem – elements of probability (set theory)

---



# Bayes theorem – Derivation

---

Given the conditional probability definition

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} \rightarrow P(Y \cap X) = P(Y|X) * P(X)$$

If we substitute the last expression in the first version of the conditional probability definition, we have the Bayes theorem

$$\overbrace{P(X|Y)}^{\text{Posterior}} = \frac{P(Y|X) * P(X)}{P(Y)} = \frac{\overbrace{P(Y|X)}^{\text{Likelihood}}}{\underbrace{P(Y)}_{\text{Marginal}}} * \overbrace{P(X)}^{\text{Prior}}$$

The Bayes theorem allows to express the **posterior** probability that a certain hypothesis ( $X$ ) holds given the observation of certain effects ( $Y$ ).

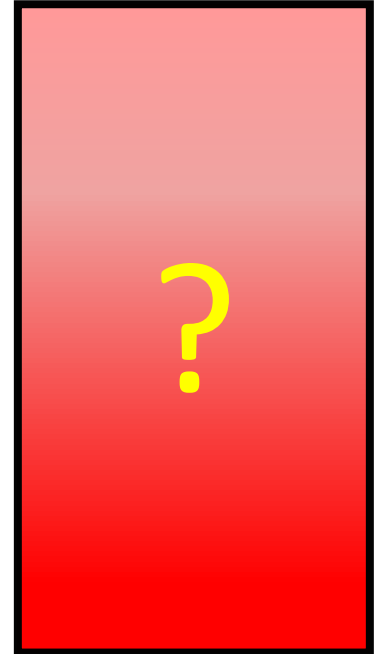
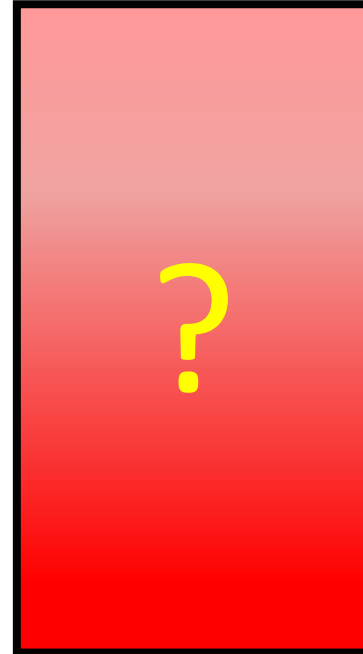
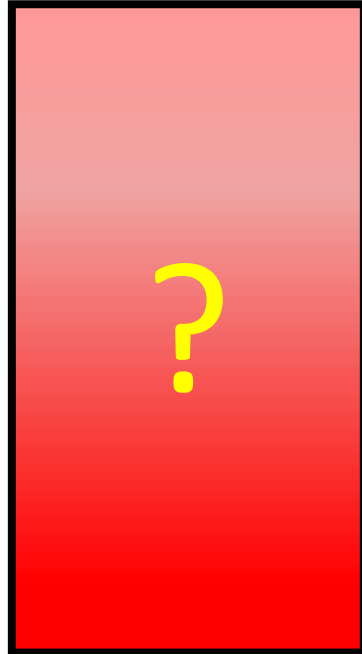
The conditional probability expressed in terms of the **joint** distribution of the hypothesis  $X$  and the event  $Y$  ( $P(Y, X)$ ), which is equal to the conditional probability or **likelihood** (distribution) of the effect given the hypothesis ( $P(Y|X)$ ) time the **prior** probability of the hypothesis ( $P(X)$ ).

## **Marginal distribution**

$$P(Y) = \sum_{i=1}^n P(Y|X_i) * P(X_i)$$

# Bayes inference illustration – The Monty Hall problem

---



# Bayes inference illustration – The Monty Hall problem

---

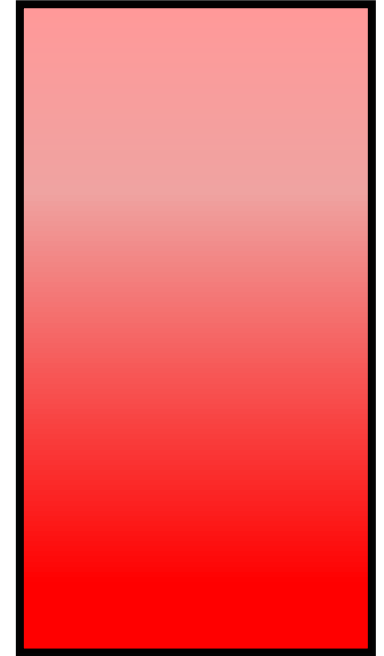
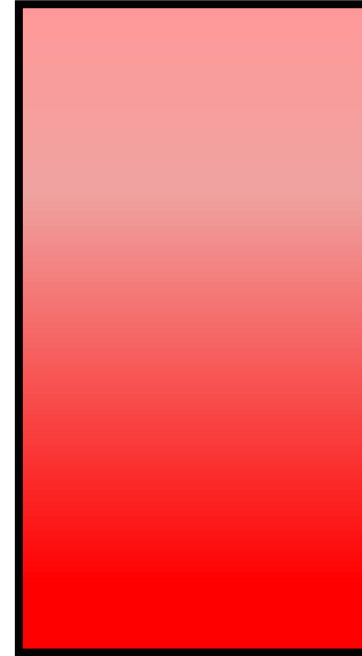
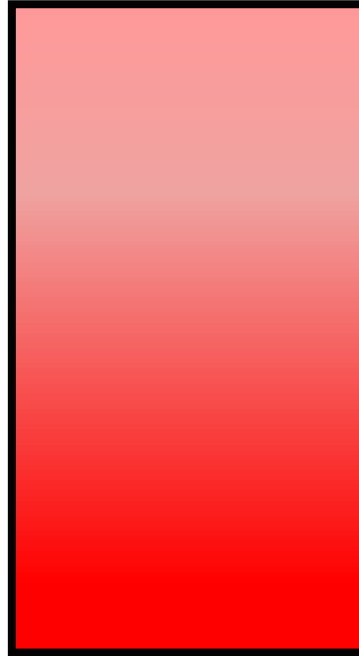
## Randomization of the prices



# Bayes inference illustration – The Monty Hall problem

---

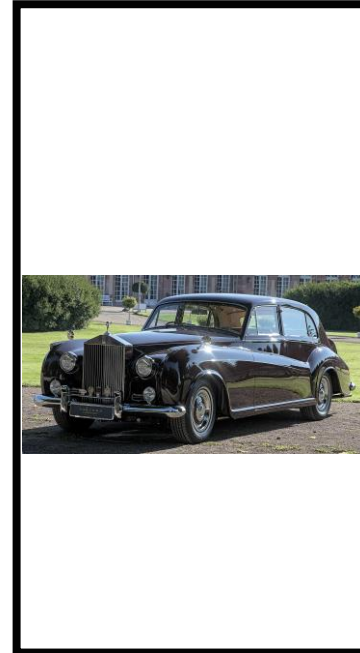
## Randomization of the prices



# Bayes inference illustration – The Monty Hall problem

---

## Randomization of the prices

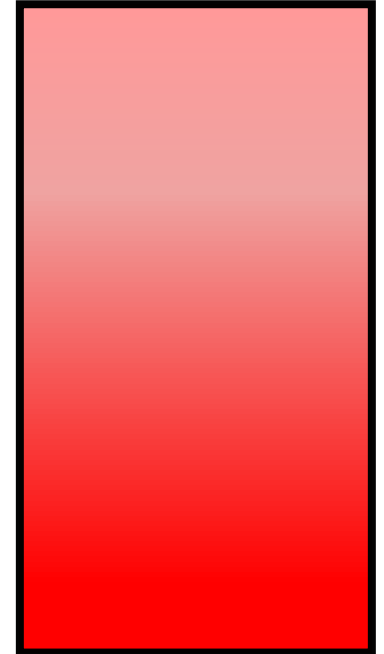
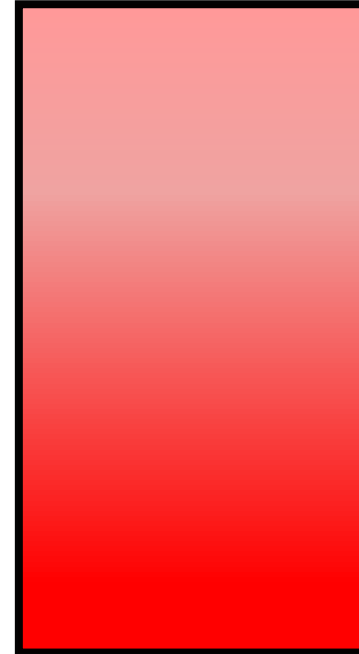
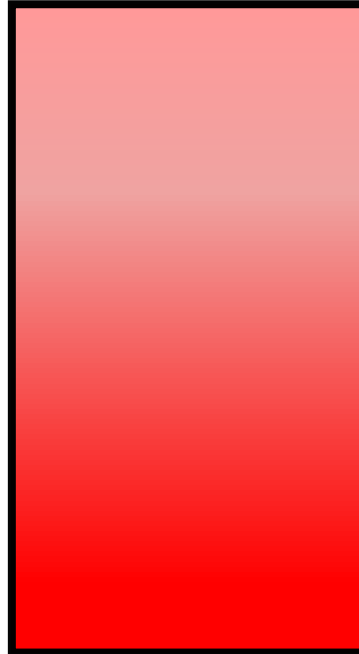




# Bayes inference illustration – The Monty Hall problem

---

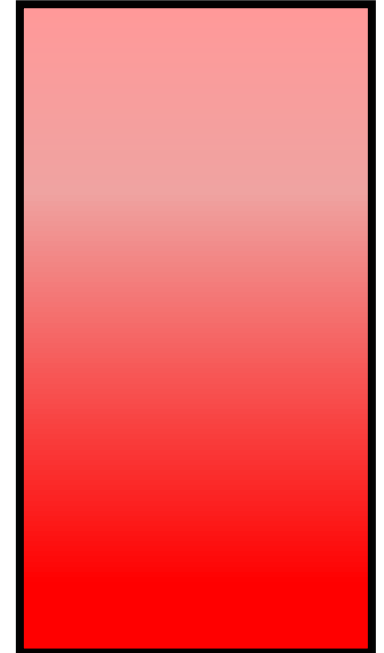
**Select one door**



# Bayes inference illustration – The Monty Hall problem

---

**Moderator opens one door**

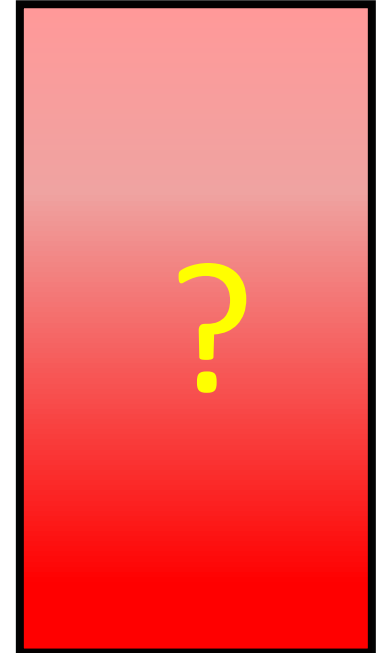
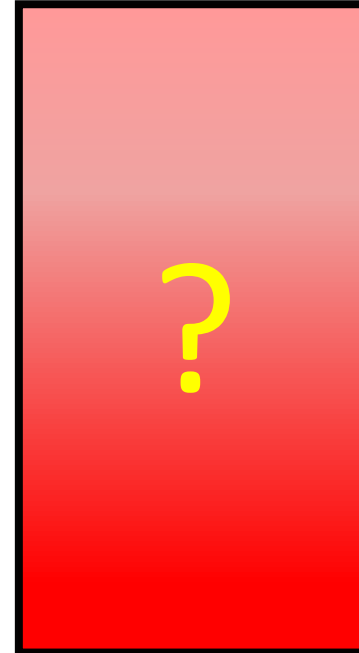




# Bayes inference illustration – The Monty Hall problem

---

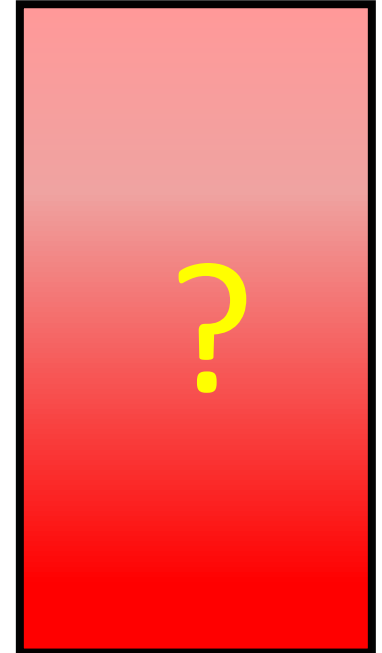
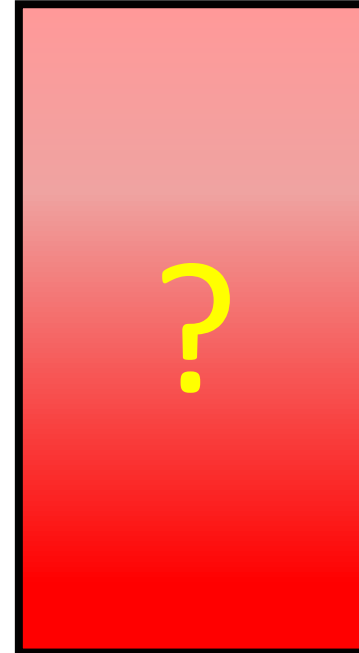
**Keep your choice or switch?**



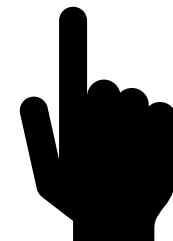
# Bayes inference illustration – The Monty Hall problem

---

**Keep your choice or switch?**

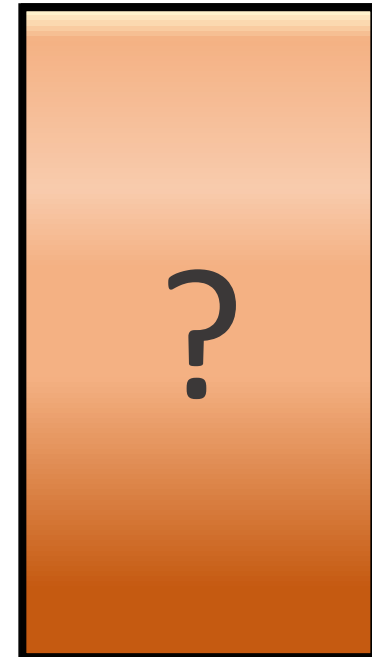
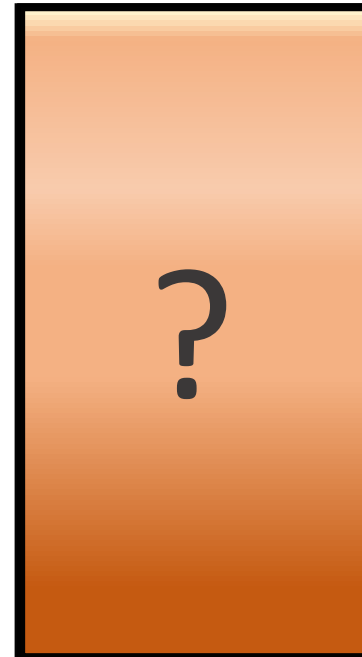
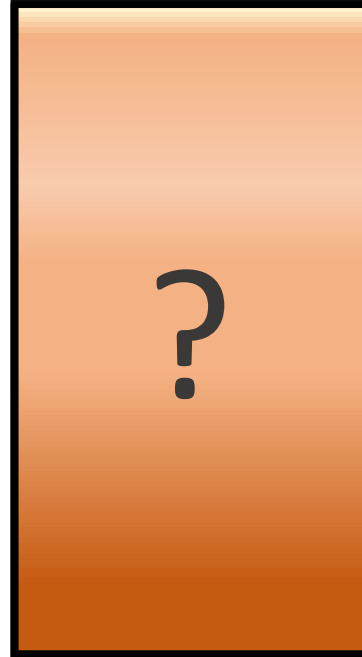


- A) Keeping initial choice increase your chance of winning
- B) Switching increase your chance of winning
- C) Both strategy have the same probability
- D) In the long term you should sometimes keep initial choice sometime change.



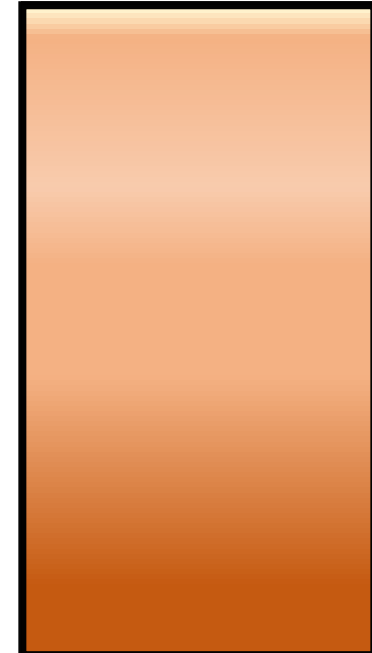
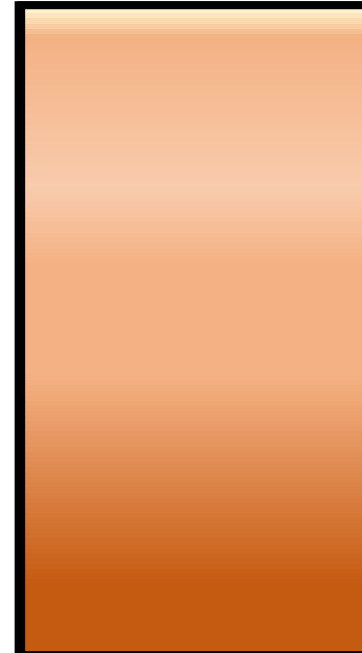
# Bayes inference illustration – The Monty Hall problem

---



# Bayes inference illustration – The Monty Hall problem

---

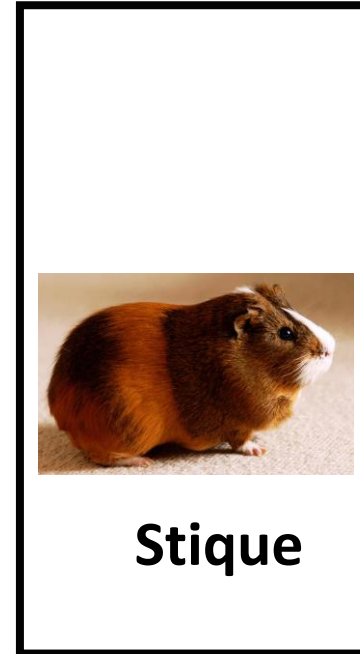
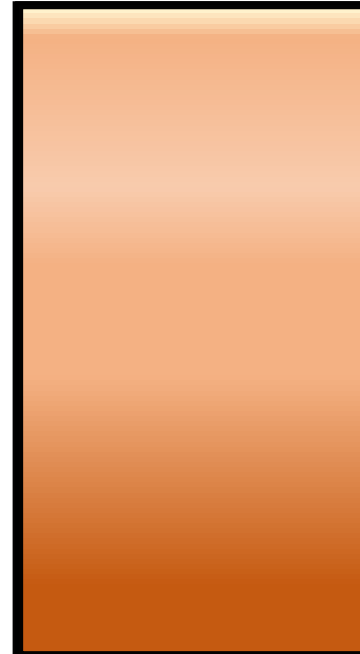


# Bayes inference illustration – The Monty Hall problem

---



Stati

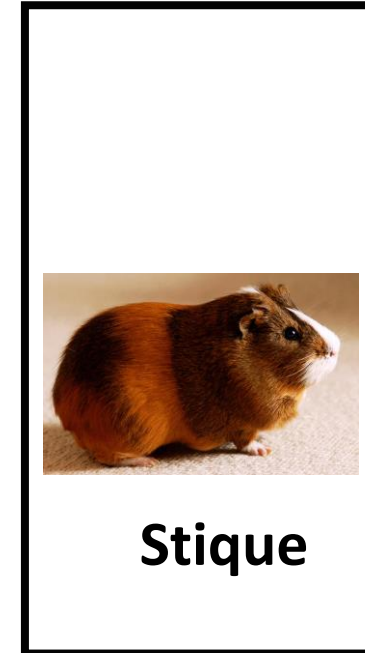
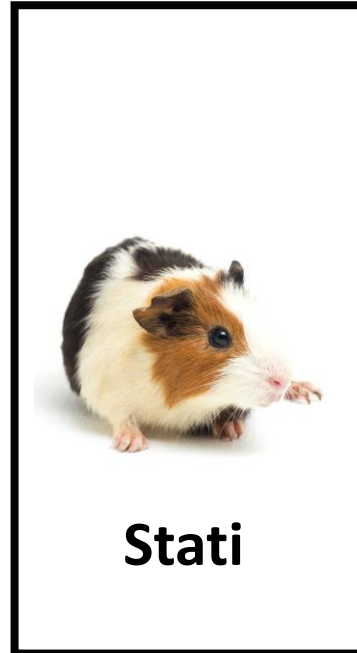


Stique



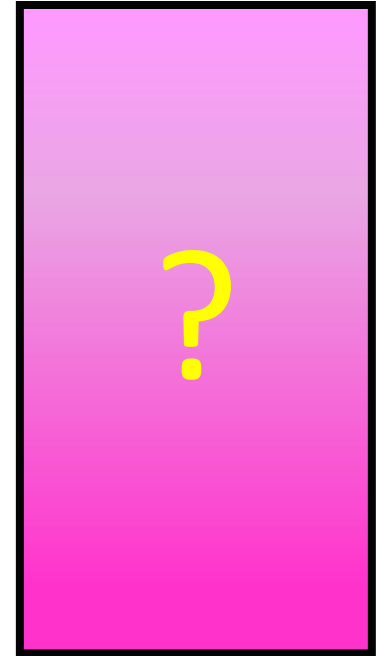
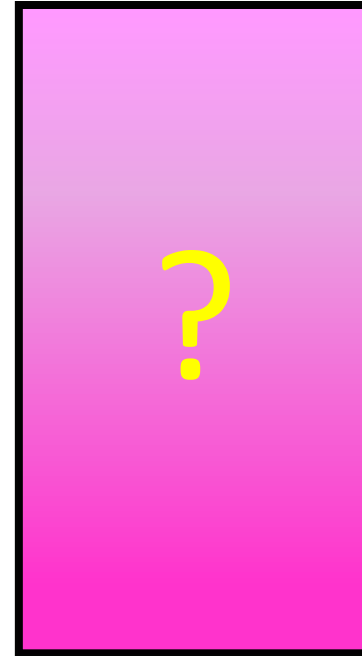
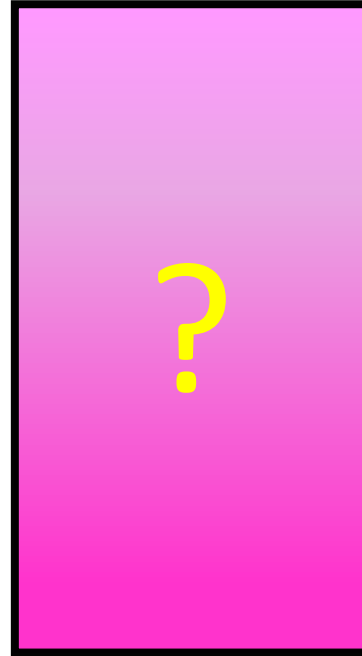
# Bayes inference illustration – The Monty Hall problem

---



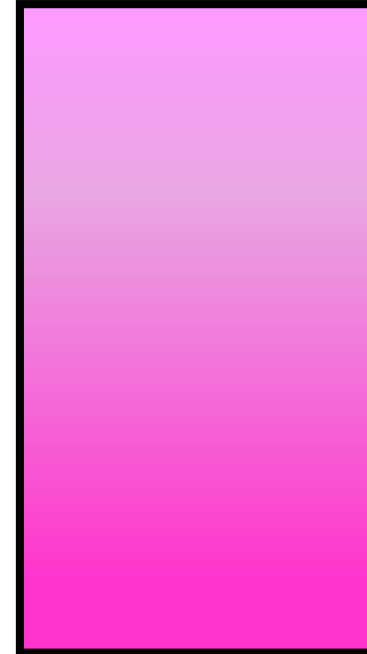
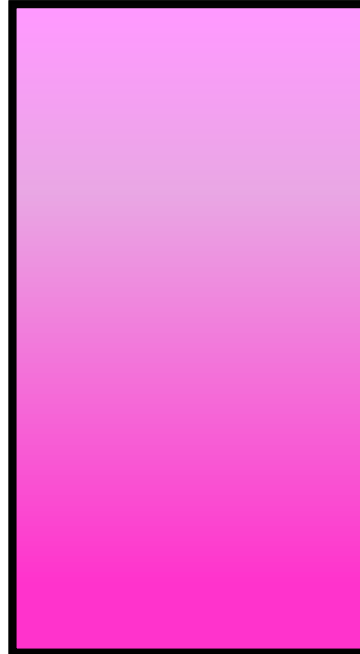
# Bayes inference illustration – The Monty Hall problem

---



# Bayes inference illustration – The Monty Hall problem

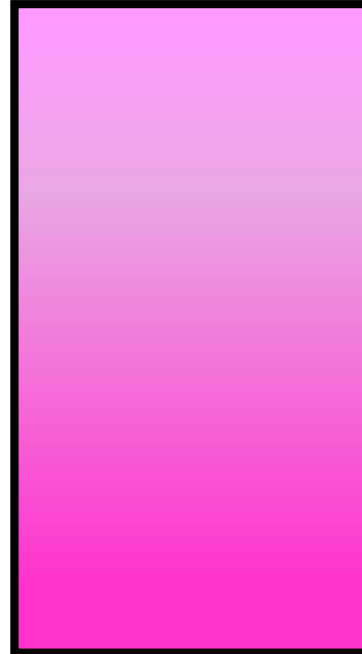
---





# Bayes inference illustration – The Monty Hall problem

---



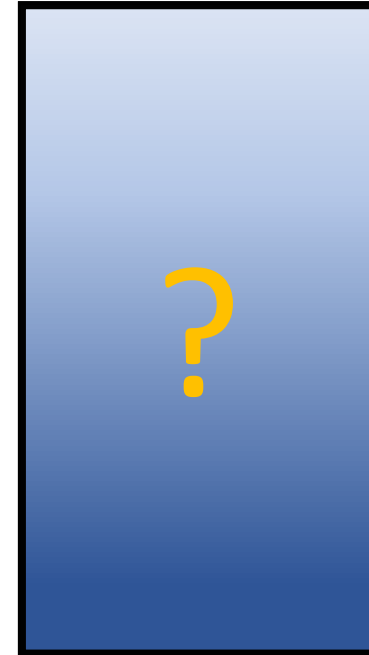
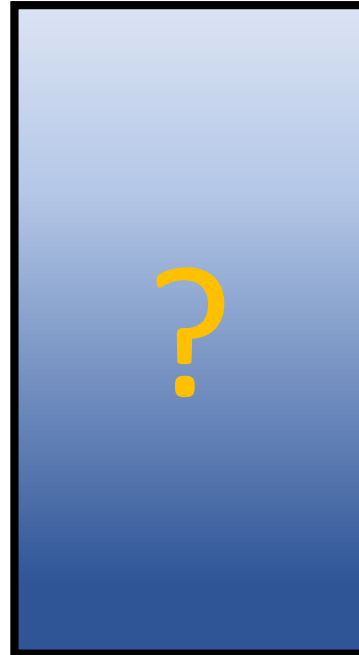
# Bayes inference illustration – The Monty Hall problem

---



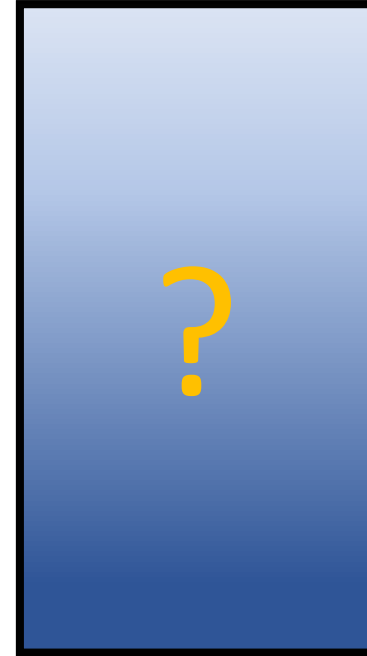
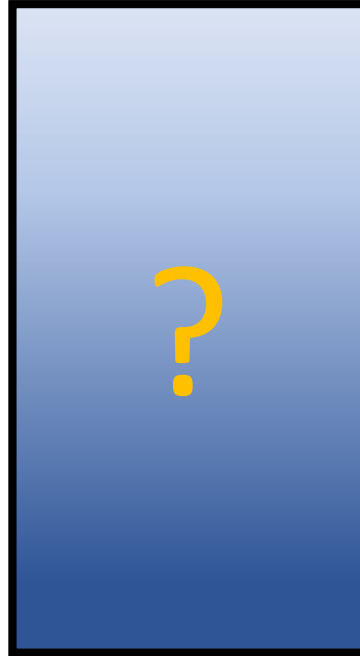
# Bayes inference illustration – The Monty Hall problem

---



# Bayes inference illustration – The Monty Hall problem

---



# Bayes inference illustration – The Monty Hall problem

---



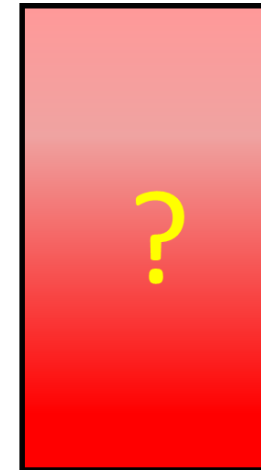
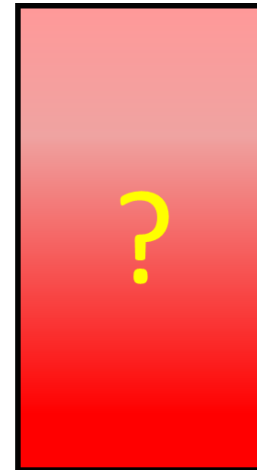


# Bayes inference illustration – The Monty Hall problem

---



**Keep your choice or switch?**



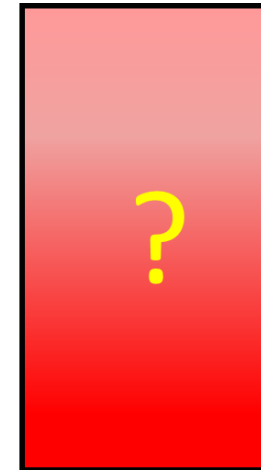
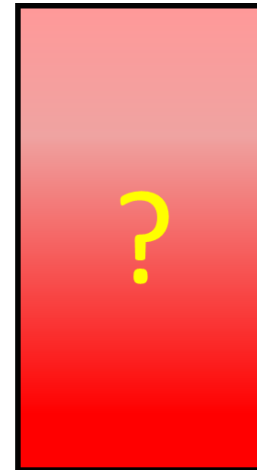
- A) Keeping initial choice increase your chance of winning
- B) Switching increase your chance of winning
- C) Both strategy have the same probability
- D) In the long term you should sometimes keep initial choice sometime change.

# Bayes inference illustration – The Monty Hall problem

---



**Keep your choice or switch?**



- A) Keeping initial choice increase your chance of winning
- B) Switching increase your chance of winning by a factor 2**
- C) Both strategy have the same probability
- D) In the long term you should sometimes keep initial choice sometime change.

# The Monty Hall problem – Bayesian resolution

---

Let us assume that the doors are labeled  $A, B, C$ , and that the candidate select the  $A$  door. We call this event  $C1_A$

$C1_A$  : the candidate select the A door as a first choice

Then the moderator open the  $C$  door which reveal a goat. We call this event  $O_c$

$O_c$  : the moderator open the C door after candidate first choice

We need to estimate the probability to get the car in the situation where the candidate remain with his initial choice

$$P(A = car | C1_A \cap O_c)$$

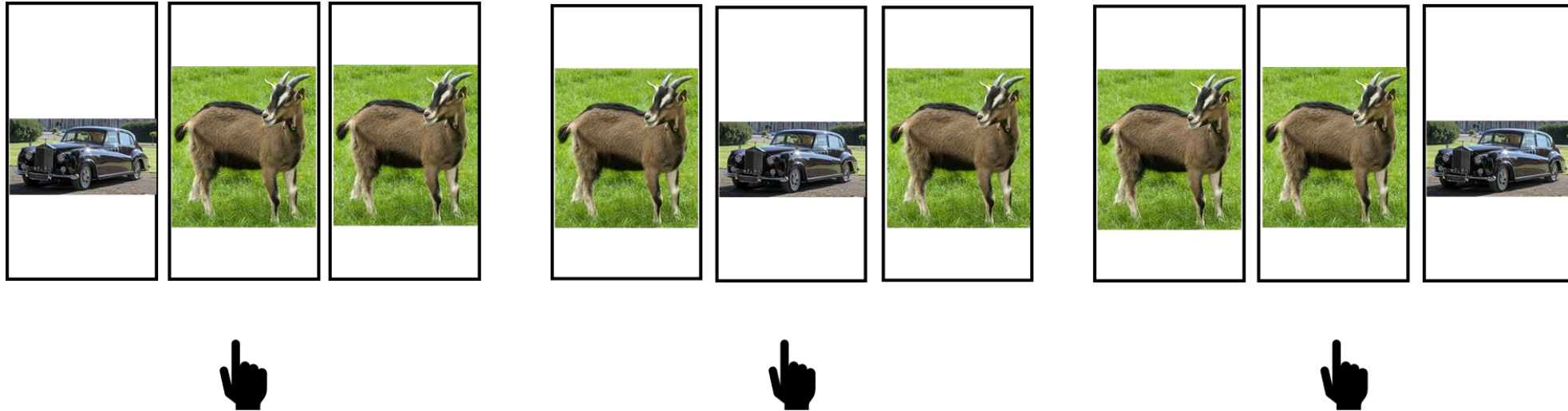
and the situation where the candidate switch to  $B$  door.

$$P(B = car | C1_A \cap O_c)$$



# The Monty Hall problem – Bayesian resolution

---



In the first option, the candidate ignore the extra information given by the fact that the moderator opened the  $C$  door.

$$P(A = car | C1_A \cap O_c) = P(A = car | \emptyset) = P(A = car) = \frac{1}{3}$$

# The Monty Hall problem – Bayesian resolution

---

In the second switching case, we can use the conditional probability

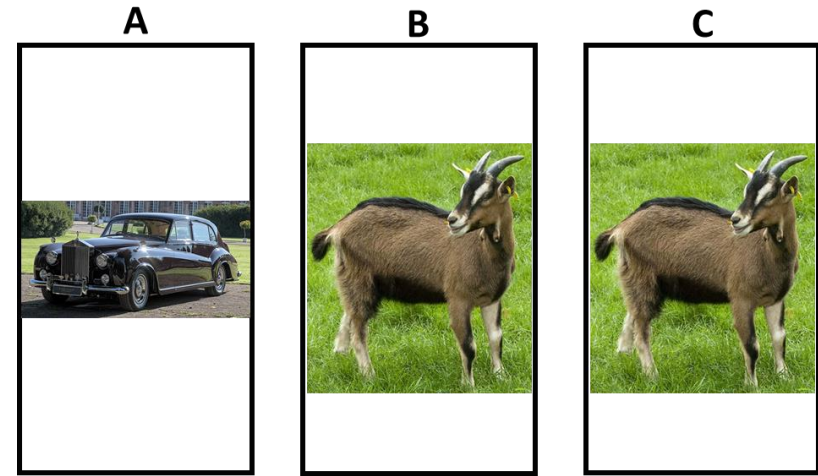
$$P(B = car | C1_A \cap O_c) = \frac{P(O_c \cap C1_A | B = car) * P(B = car)}{P(O_c \cap C1_A)}$$

We can first develop the denominator

$$P(O_c \cap C1_A) = \sum_{i=1}^{n_E} P(O_c \cap C1_A | E_i) * P(E_i) = P(O_c \cap C1_A | A = car) * P(A = car) +$$
$$P(O_c \cap C1_A | B = car) * P(B = car) +$$
$$P(O_c \cap C1_A | C = car) * P(C = car)$$

# The Monty Hall problem – Bayesian resolution

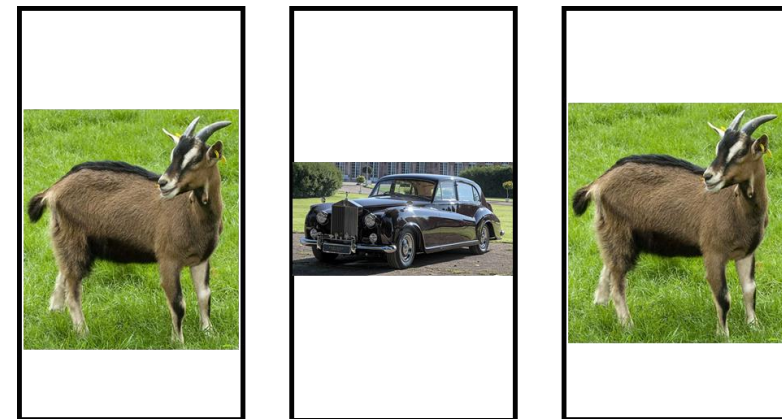
$$P(O_c \cap C1_A | A = car) = \frac{1}{2}$$



P = 1/2

P = 1/2

$$P(O_c \cap C1_A | B = car) = 1$$



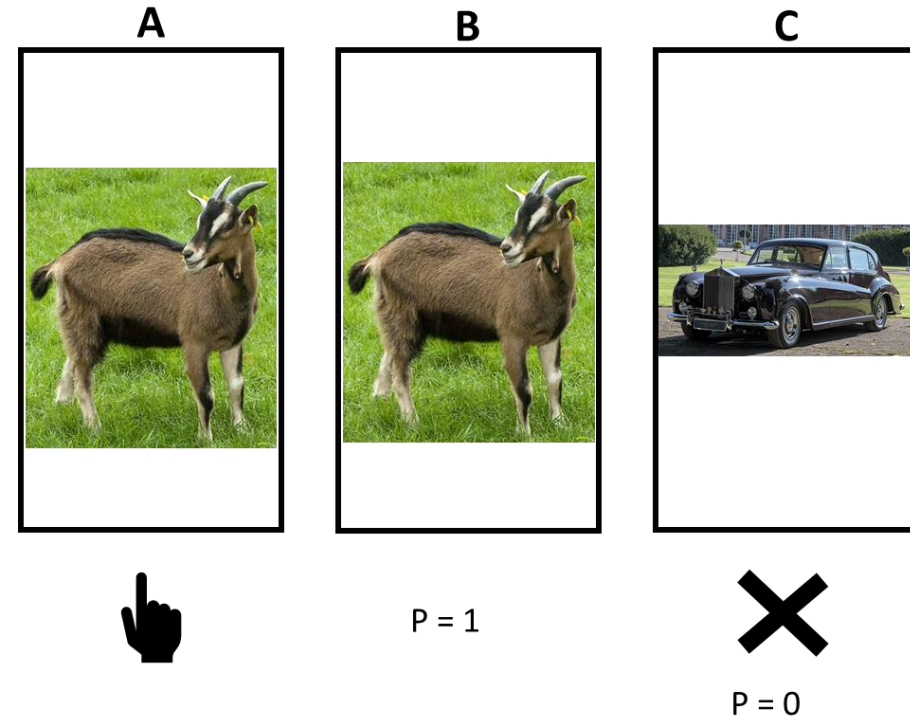
P = 0

P = 1

# The Monty Hall problem – Bayesian resolution

---

$$P(O_c \cap C1_A | C = car) = 0$$



$$P(O_c \cap C1_A) = \left(\frac{1}{2} \times \frac{1}{3}\right) + \left(1 \times \frac{1}{3}\right) + \left(0 \times \frac{1}{3}\right) = \frac{1}{6} + \frac{2}{6} = \frac{3}{6} = \frac{1}{2}$$

# The Monty Hall problem – Bayesian resolution

---

The numerator  $P(O_c \cap C1_A | B = car) * P(B = car) = 1 \times \frac{1}{3} = \frac{1}{3}$ , so the posterior probability is

$$\begin{aligned} P(B = car | C1_A \cap O_c) &= P(B = car | \text{use info to switch}) \\ &= \frac{P(O_c \cap C1_A | B = car)}{P(O_c \cap C1_A)} * P(B = car) \\ &= \frac{1}{\frac{1}{2}} * \frac{1}{3} \\ &= \frac{2}{3} \end{aligned}$$

The change in probability between the non switch situation  $P(A = car) = \frac{1}{3}$  and the switch situation  $P(B = car | \text{use info to switch}) = \frac{2}{3}$  come from the fact that since the moderator do not open one door (here  $B$ ) after the candidate select  $A$ , it increase the chance that this door contain the car because, to keep the suspens, the moderator will necessarily open a door that contain a goat.

# The Monty Hall problem – Bayesian logic (posterior = likelihood \* prior)

---

(Bayesian) Logic and probability can be highly counter-intuitive

**Bayesian logic combine two sources of information:**

- initial information/guess (prior)
- Data/evidence (e.g. the non-open door)

$$\overbrace{P(X|Y)}^{\textit{Posterior}} = \frac{P(Y|X) * P(X)}{P(Y)} = \frac{\overbrace{P(Y|X)}^{\textit{Likelihood}}}{\underbrace{P(y)}_{\textit{Marginal}}} * \overbrace{P(X)}^{\textit{Prior}}$$

# Bayesian logic (posterior = likelihood \* prior)

---

$$\overbrace{P(X|Y)}^{\text{Posterior}} = \frac{P(Y|X) * P(X)}{P(Y)} = \frac{\overbrace{P(Y|X)}^{\text{Likelihood}}}{\underbrace{P(y)}_{\text{Marginal}}} * \overbrace{P(X)}^{\text{Prior}}$$

After the experiment is realized the  $P(Y)$  will be a constant. So the Bayes theorem can be rewritten:

$$P(X|Y) \propto P(Y|X) * P(X)$$

Calculating the posterior distribution  $P(X|Y)$  can be challenging. Different approaches are possible.

1. Choose a prior distribution  $P(X)$  that make the derivation easier. For example *conjugate priors* have the property that their combination with a certain distribution of the data gives a specific distribution. For example, normal prior with normally distributed data gives normal posterior.
2. Simulate the posterior distribution by sampling many realisations from this distribution. The construction of the posterior distribution is one of the central task of the Bayesian algorithm construction.

# Bayesian theorem: discrete to continuous case

---

The evidence is now given by a vector of observations  $\mathbf{y} = y_1, \dots, y_N$  and the hypothesis is a vector of unknowns  $\theta$ . We can derive the Bayes theorem in the continuous case using the definition of the conditional distribution, which is similar to the discrete case.

$$p(\mathbf{y}|\theta) = \frac{p(\mathbf{y}, \theta)}{p(\theta)} \rightarrow p(\mathbf{y}, \theta) = p(\mathbf{y}|\theta) * p(\theta)$$

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}, \theta)}{p(\mathbf{y})}$$

If we substitute  $p(\mathbf{y}, \theta) = p(\mathbf{y}|\theta) * p(\theta)$  in the last expression, we have

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta) * p(\theta)}{p(\mathbf{y})}$$

Where,

$p(\theta|\mathbf{y})$ : **posterior** density or distribution of the parameter ( $\theta$ ) given the observation ( $\mathbf{y}$ ).

$p(\mathbf{y}|\theta)$ : **likelihood** distribution of the observation ( $\mathbf{y}$ ) given parameter ( $\theta$ ).

$p(\theta)$ : **prior** distribution of the parameters ( $\theta$ ).

$p(\mathbf{y})$ : data **marginal** density

The data marginal density can be obtained by integration of the joint distribution over the parameters

$$p(\mathbf{y}) = \int p(\mathbf{y}, \theta) d\theta = \int p(\mathbf{y}|\theta) p(\theta) d\theta$$

The prior  $p(\theta)$  must properly be chosen for the integral to exist.

In our case,

$$\mathbf{y} = \mu + x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p + e$$

$$\theta = [\beta_1, \beta_2, \dots, \beta_p]$$



# Bayesian approach steps

---

1. Define a distribution for the data under a certain model (**likelihood**) using a certain number of unknown parameters
2. Define prior distributions for all unknown parameters (**prior elicitation**).
3. Arrive at conditional distribution of all unknown parameters given the data (**posterior**).
4. Derive marginal or conditional posterior distribution for all parameters using probability theory.
5. Use an algorithm (Metropolis-Hasting, Gibbs sampler) to sample from the marginal posterior (**(MCMC) sampling**).
6. Use the sampled data to reconstruct the parameter distribution and interpret in a probabilistic way.

# Bayesian approach illustration – model (likelihood) definition

---

Example from: Clyde M., Çetinkaya-Rundel M., Rundel C., Banks D., Chai C., Huang L. (2022) An Introduction to Bayesian Thinking A Companion to the Statistics with R Course. <https://statswithr.github.io/book/>

The Poisson distribution was used to model the number of cavalrymen kicked to death by horse every year in a unit by Bortkiewicz.

This variable (# of men kicked to death) is a count variable that can be supposed to be independent, so the Poisson distribution is a possible choice.

## Model definition

Let us assume a vector of observation  $\mathbf{y} = (y_1, \dots, y_n)$  for which  $y_i$  follows a Poisson distribution. with expectation  $\lambda$ .

$$y \sim \text{Poisson}(\lambda)$$
$$P(Y = y_i | \lambda) = \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

With  $E[y] = V[y] = \lambda$



Ladislaus Bortkiewicz

# Bayesian approach illustration – prior elicitation

---

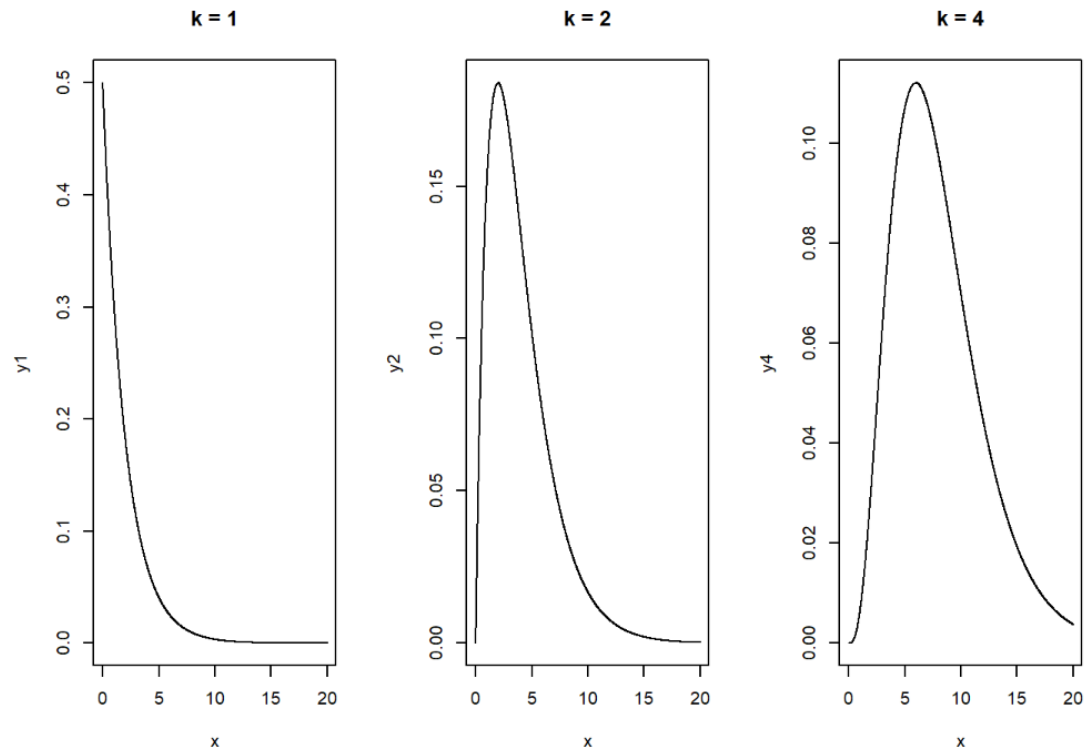
Prior elicitation consists of formalizing prior knowledge in the form of a probability distribution.

It is convenient to assume that the prior of the  $\lambda$  is a Gamma distribution because there is a Poisson-Gamma conjugacy. Therefore

The prior distribution of  $\lambda$  is assumed to be a Gamma distribution

$$\lambda|k, \theta = \Gamma(k, \theta) = \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\frac{\lambda}{\theta}}$$

The  $k$  parameter control the shape of the distribution.

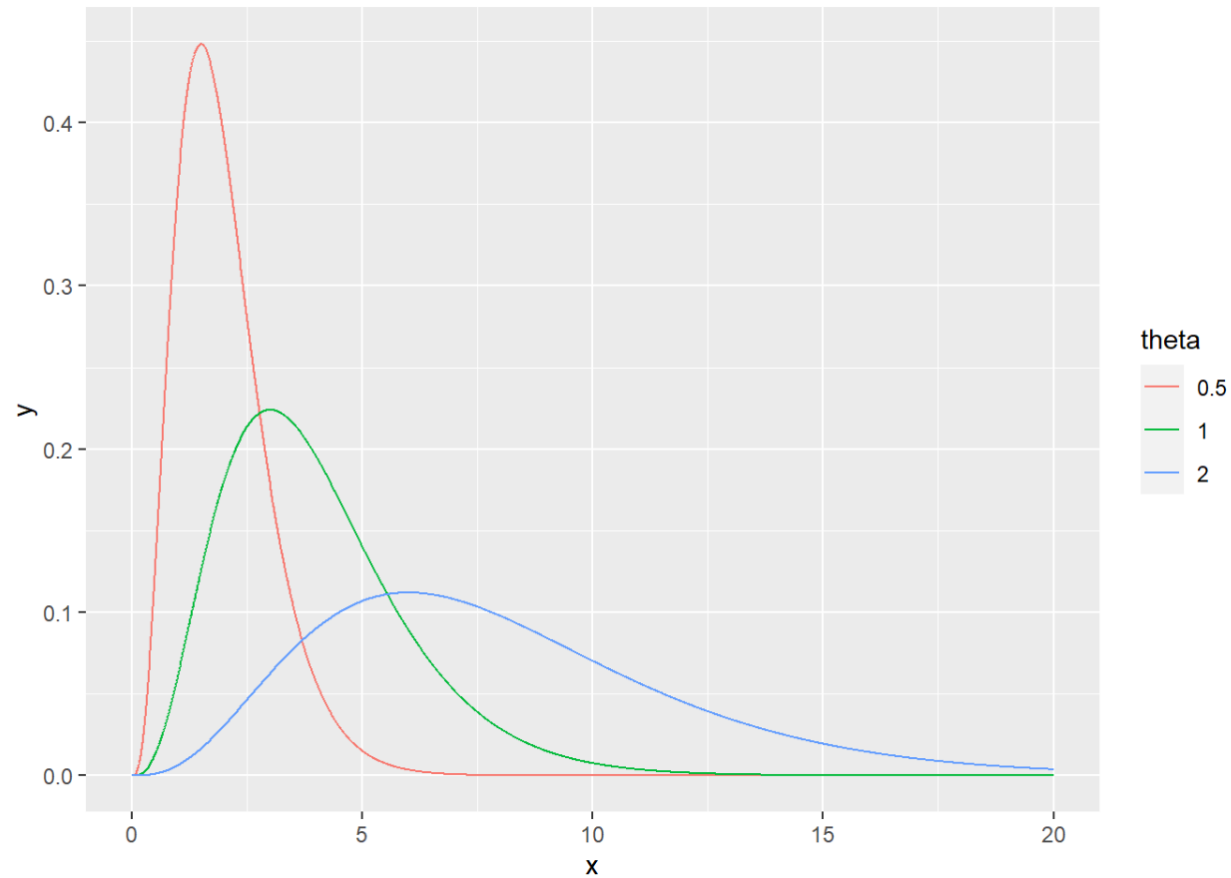


# Bayesian approach illustration – prior elicitation

---

$$\lambda|k, \theta = \Gamma(k, \theta) = \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\frac{\lambda}{\theta}}$$

while the  $\theta$  parameter controls the scale (height/flatness) of the distribution

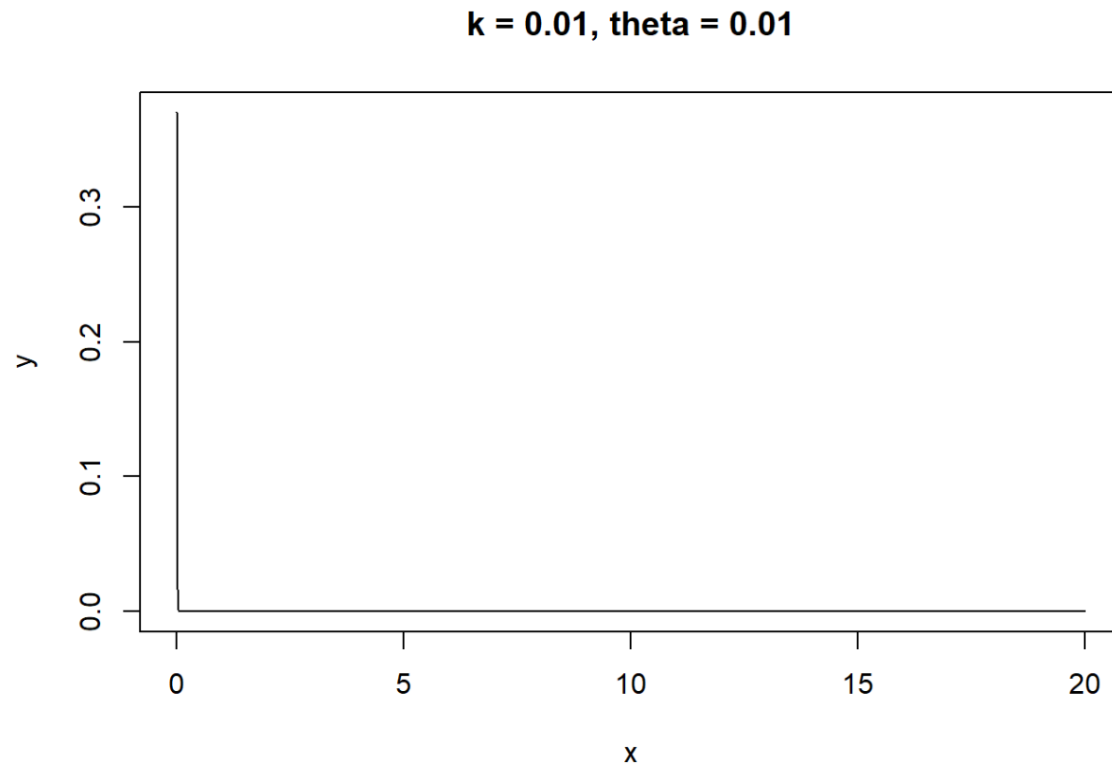


# Bayesian approach illustration – prior elicitation

---

$$\lambda|k, \theta = \Gamma(k, \theta) = \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\frac{\lambda}{\theta}}$$

It is not possible to get a completely uninformative “flat” prior with the Gamma distribution. Using small values for  $k$  and  $\theta$  allows to come close from such a distribution



# Bayesian approach illustration – prior elicitation

---

$$\lambda|k, \theta = \Gamma(k, \theta) = \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\frac{\lambda}{\theta}}$$

The determination of **hyper-parameters** values (the parameters that are fixed by the user before the estimation of the parameter starts), is part of the prior elicitation activity. Let us assume that before looking at the data, we assume that  $\lambda = 0.75$  with a standard deviation of 1. We need to use those values and the definition of the Gamma distribution expectation  $E(\lambda) = k\theta$  and variance  $V(\lambda) = k\theta^2$

$$k\theta = 0.75$$

$$k\theta^2 = 1$$

$$k\theta = 0.75 \rightarrow k = \frac{3}{4\theta}$$

$$\frac{3}{4\theta} \theta^2 = 1 \rightarrow \theta = \frac{4}{3} \quad \text{and} \quad k = \frac{9}{16}$$



# Bayesian approach illustration – posterior distribution derivation

---

Using model (likelihood) and the prior we can derive the posterior distribution

$$\begin{aligned} p(\lambda|k, \theta, \mathbf{y}) &\propto p(\mathbf{y}|\lambda, k, \theta)p(\lambda|k, \theta) \\ &\propto L(\lambda|\mathbf{y})p(\lambda|k, \theta) \end{aligned}$$

We use here the fact that the posterior can be expressed as a scaled likelihood ( $p(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$ ). The likelihood expression is given by

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^n \frac{\lambda^{y_i} e^{-\lambda}}{y_i!} = \frac{\lambda^{\sum y_i} e^{-N\lambda}}{\prod_{i=1}^n y_i!} \propto \lambda^{\sum y_i} e^{-N\lambda}$$

Using this we can continue to derive the posterior distribution

$$\begin{aligned} p(\lambda|k, \theta, \mathbf{y}) &\propto \lambda^{\sum y_i} e^{-N\lambda} \frac{1}{\theta^k \Gamma(k)} \lambda^{k-1} e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\sum y_i} e^{-N\lambda} \lambda^{k-1} e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\sum y_i} \lambda^{k-1} e^{-N\lambda} e^{-\frac{\lambda}{\theta}} \\ &\propto \lambda^{\sum y_i + k - 1} e^{-(\lambda \frac{N\theta + 1}{\theta})} \\ &\propto \lambda^{\sum y_i + k - 1} e^{-\left(\frac{\lambda}{\frac{\theta}{N\theta + 1}}\right)} \end{aligned}$$

The final expression correspond to a Gamma distribution with parameter  $\sum y_i + k$  and  $\frac{\theta}{N\theta + 1}$ .

$$p(\lambda|k, \theta, \mathbf{y}) \sim \Gamma(k^* = \sum y_i + k; \theta^* = \frac{\theta}{N\theta + 1})$$

# Bayesian approach illustration – posterior distribution derivation

---

$$p(\lambda|k, \theta, \mathbf{y}) \sim \Gamma(k^* = \sum y_i + k; \theta^* = \frac{\theta}{N\theta + 1})$$

Therefore, we can use the available data to get the posterior. Let us assume that over 20 years in 15 unit 200 cavalrymen died. Therefore,  $\sum y_i = 200$  and  $N = 300$ . This means that

$$k^* = 200 + 9/16 = 200.5625$$

and

$$\theta^* = \frac{4/3}{(300 * (4/3)) + 1} = 0.0033$$

$$p(\lambda|k, \theta, \mathbf{y}) \sim \Gamma(k^* = 200.5625; \theta^* = 0.0033)$$

The distribution allow also to check the change of value from prior values to posterior estimates.

Estimates	Prior	Posterior
$E[\lambda]$	0.75	$200.56 * 0.0033 = 0.66$
$V[\lambda]$	1	$200.56 * 0.0033^2 = 0.047$

# Bayesian approach illustration – MC(MC) sampling

---

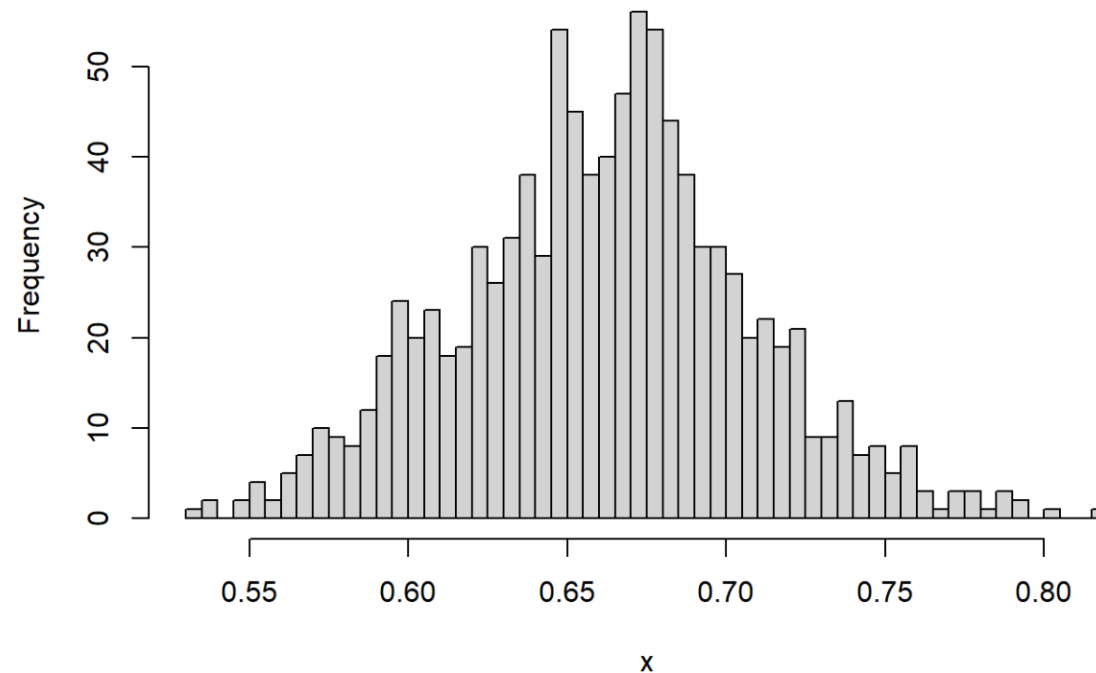
In this example, the use of a conjugate prior allowed to get a closed form for the posterior distribution.

This results gives us a probability distribution that we can use to get a probabilistic interpretation about the parameter of interest.

In many case, however, it is difficult or impossible to get a closed form for the posterior distribution. In that case, we can use a numerical integration strategy which consist of randomly sampling from a conditional posterior distribution. This random sampling strategy is called Monte Carlo (MC) sampling.

For example, in our situation, we could still sample from the posterior distribution

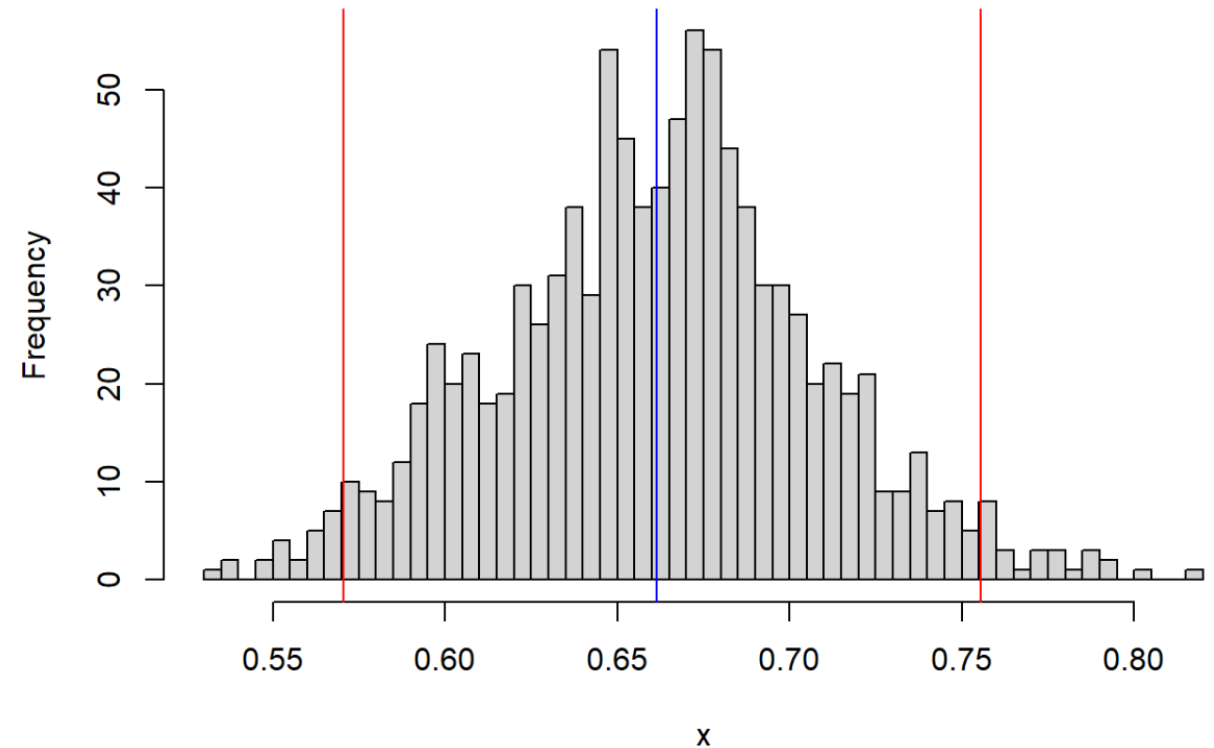
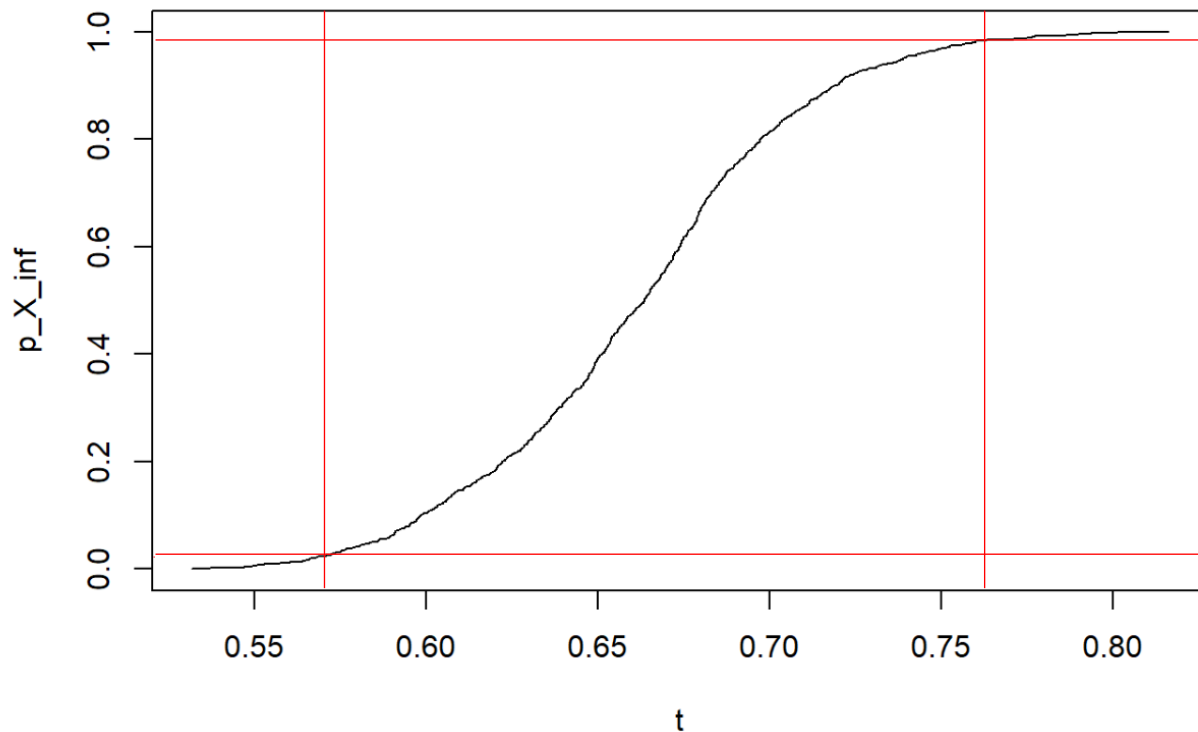
$$p(\lambda|k, \theta, \mathbf{y}) \sim \Gamma(k^* = 200.5625; \theta^* = 0.0033)$$



# Bayesian approach illustration – Probabilistic interpretation

Using this distribution we make probabilistic inference about the parameter estimate. For that purpose, we can use the empirical cumulative distribution function. for a grid of values  $t$ , we can estimate the probability that the random sample values  $x$  have an inferior or equal value

$$\hat{F}(t) = \frac{\sum_{i=1}^n I(x_i \leq t)}{n}$$



# Bayesian approach illustration – MCMC sampling

---

Markov Chain Monte Carlo (MCMC) methods encompass methods introduced by Metropolis and Hasting for Monte Carlo integration. Monte Carlo integration approximate complex integration by the expectation of a sample from a target distribution. Many applications of MCMC methods arise in Bayesian inference, especially the problem of integrating the parameter ( $\theta$ ) conditional posterior distribution.

Given the Bayes theorem, we know that

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}$$

where,

$f(\theta|y)$ : posterior distribution

$f(y|\theta)$ : likelihood

$f(\theta)$ : prior

Using the formula of the expectation of a function from a continuous variable ( $E[g(X)] = \int g(x)f(x)dx$ ), we can write

$$E[f(\theta|y)] = \int g(\theta)f(\theta|y)d\theta = \frac{\int g(\theta)f(y|\theta)f(\theta)}{\int f(y|\theta)f(\theta)d\theta}$$

# Bayesian approach illustration – MCMC sampling

---

$$E[f(\theta|y)] = \int g(\theta) f(\theta|y) d\theta = \frac{\int g(\theta) f(y|\theta) f(\theta)}{\int f(y|\theta) f(\theta) d\theta}$$

This kind of integration are often too difficult or impossible to calculate. MCMC provide solutions for this kind of integration. MCMC reduce the problem to estimate  $E[g(\theta)] = \int g(\theta) f(\theta|y) d\theta$  using the sample mean

$$\bar{g} = \frac{1}{n} \sum_{i=1}^n g(x_i)$$

where  $x_1, x_2, \dots, x_n$  are sampled from the posterior distribution  $f(\theta|y)$ . By the law of large numbers when  $n \rightarrow \infty$ ,  $\bar{g} \rightarrow E[g(\theta)]$ . MCMC allows to draw this large number of replications. The Markov chain provide the sampler that generate the chain. More formally, if  $X_0, X_1, \dots, X_n$  are realisation of an irreducible, ergodic Markov chain with stationary distribution equal to  $f(\theta|y)$ , then

$$\bar{g}(X) = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

converge to  $E[g(X)]$  if  $n \rightarrow \infty$ .



# Bayesian approach illustration – Metropolis Hasting algorithm

---

The Metropolis-Hastings algorithm is a class of MCMC methods that include special cases of the Metropolis sampler, the Gibbs sampler, the independence sampler and the random walk. The main objective is to generate a Markov chain with a stationary distribution that is the **target distribution**. One of the main task is to choose a **proposal distribution** that allow to meet Markov chain properties (irreducibility, positive recurrence, and aperiodicity) and converge to the target distribution.

The M-H algorithm generate the Markov Chain as follow

1. Choose the proposal distribution  $g(\cdot | X_t)$  subject to regularity conditions
2. Generate  $X_0$  from  $g$
3. repeat the following sampling steps until convergence:
  - a. Generate  $y$  from  $g(\cdot | X_t)$
  - b. Generate  $u$  from  $U[0, 1]$
  - c. If

$$u \leq \frac{f(y)g(X_t|y)}{f(X_t)g(y|X_t)}$$

Accept  $y$ , so  $X_{t+1} = y$  otherwise  $X_{t+1} = X_t$

- d. increment  $t$

We can interprete the selection rules in the following way, if  $f(y)g(X_t|y) \geq f(X_t)g(y|X_t)$  the newly sampled point  $y$  has a higher probability than the previous one  $X_t$  (could be more precisely defined). In that case, we always accept the point otherwise, we accept with a probability  $u$ .

This is equivalent to say that the candidate point is accepted with probability

$$\alpha(X_t, y) = \min\left(1, \frac{f(y)g(X_t|y)}{f(X_t)g(y|X_t)}\right)$$

# Bayesian approach illustration – Metropolis Hasting algorithm illustration

---

Example from Rizzo M. L. (2008) Statistical Computing with R

We can illustrate the M-H algorithm using the approximation of a Rayleigh density

$$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$$

with  $x \geq 0$  and  $\sigma > 0$

The Rayleigh distribution is the target distribution. Here we will assume that  $\sigma = 4$

1. Choose the proposal distribution  $g(\cdot | X_t)$

We can use the chi-squared distribution with the degree of freedom equal to  $X_t$  (the previous value of the chain)

$$g(\cdot | X_t) \sim \chi_{df=X_t}^2$$

2. Generate  $X_0$  from  $g$

To generate the first value of the chain we assume that  $X_t = 1$ , so  $X_0 \sim \chi_{df=1}^2$

3. Chain

- a. generate  $y \sim \chi_{df=X_{t-1}}^2$
- b. generate  $u \sim U[0, 1]$
- c. calculate  $r(X_t, y) = \frac{f(y)g(X_t|y)}{f(X_t)g(y|X_t)}$  with  $f(\cdot)$  the Rayleigh distribution and  $g(\cdot | X_t) \sim \chi_{df=X_t}^2$

# Bayesian approach illustration – Metropolis Hasting algorithm illustration

---

```
library(VGAM)
```

```
# set the chain
iter <- 10000
x <- rep(NA, iter)
sigma <- 4
k <- 0

# initialize the first value of the chain
x[1] <- rchisq(n = 1, df = 1)

for(i in 2:iter){

  # generate variable from proposal distribution
  xt <- x[i-1]
  y <- rchisq(n = 1, df = xt)

  # generate  $u \sim U[0, 1]$ 
  u <- runif(n = 1, min = 0, max = 1)

  # compute ratio
  numerator <- drayleigh(x = y, scale = sigma) * dchisq(x = xt, df = y)
  denominator <- drayleigh(x = xt, scale = sigma) * dchisq(x = y, df = xt)
  r_Xt_Y <- numerator/denominator

  # move the chain
  if(r_Xt_Y >= u){
    x[i] <- y # accept the proposition (move the chain)
    k <- k + 1
  } else {
    x[i] <- x[i-1] # reject the chain (keep it at the same stage)
  }
}

1 - (k/iter)
```

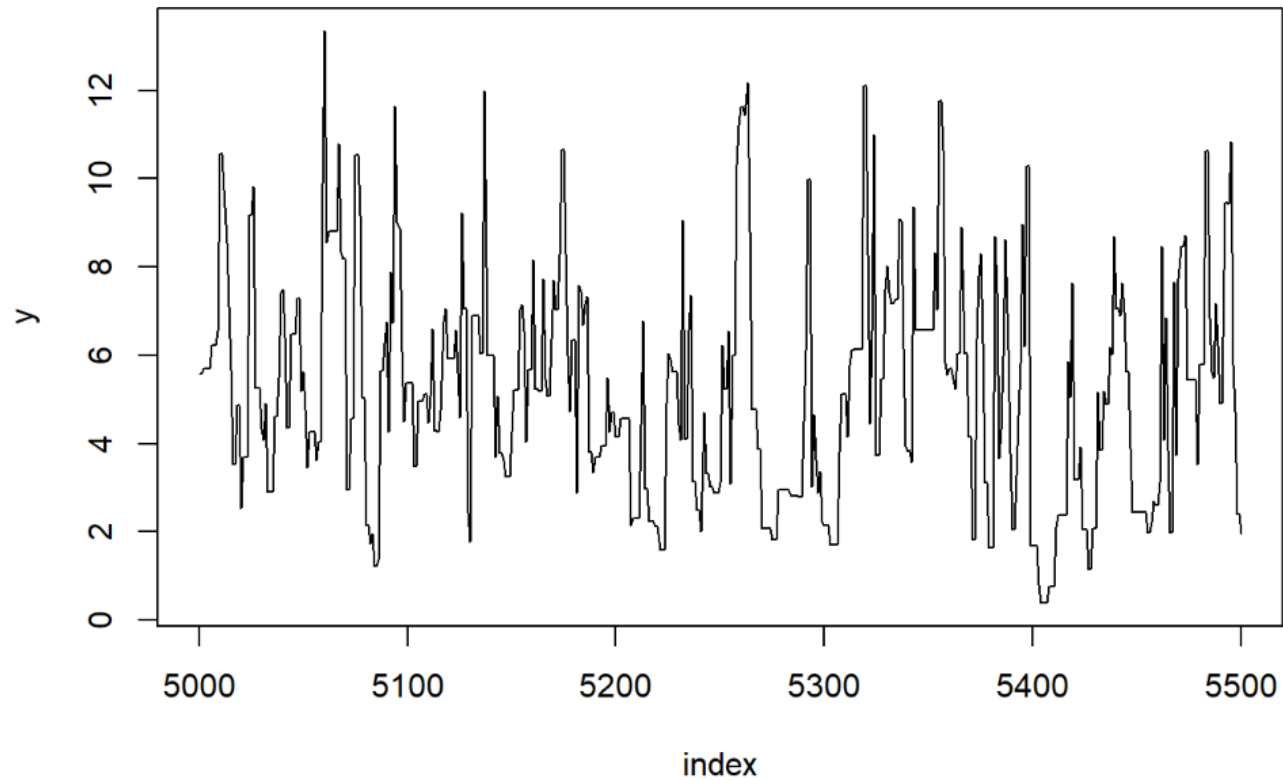
```
## [1] 0.4074
```

# Bayesian approach illustration – Metropolis Hasting algorithm illustration

---

```
# around 40% of the points are rejected
```

```
index <- 5000:5500  
y <- x[index]  
plot(x = index, y, type = "l")
```



# Bayesian approach illustration – Gibbs sampling

---

The Gibbs sampler is a special case of the M-H sampler. In the Gibbs sampler all candidates points are accepted.

Let us assume the following multi-parameter model

$$y \sim p(y|\theta_1, \theta_2, \dots, \theta_p)$$

with the following joint conditional distribution

$$y \sim p(\theta_1, \theta_2, \dots, \theta_p | y) \propto p(y|\theta_1, \theta_2, \dots, \theta_p)p(\theta_1, \theta_2, \dots, \theta_p)$$

It can be difficult to integrate that distribution to derive marginal posterior

$$p(\theta_1 | y) \propto \int_{\theta_{j \neq 1}} p(\theta_1, \theta_2, \dots, \theta_p | y) d\theta_{j \neq 1}$$

In that case, we can use MCMC sampling more particularly the Gibbs sampler to reconstruct the parameter posterior distribution through sampling in the fully conditional posterior distribution. Let us assume a model with three parameters  $\theta = [\theta_A, \theta_B, \theta_C]$ . The full conditional posterior are the following

$$p(\theta_1 | \theta_2, \theta_3, y)$$

$$p(\theta_2 | \theta_1, \theta_3, y)$$

$$p(\theta_3 | \theta_1, \theta_2, y)$$

# Bayesian approach illustration – Gibbs sampling

---

$$p(\theta_1|\theta_2, \theta_3, y)$$

$$p(\theta_2|\theta_1, \theta_3, y)$$

$$p(\theta_3|\theta_1, \theta_2, y)$$

To use the Gibbs sampler, we need to know the full conditional posterior distribution for every parameters  $\theta_i$ . Gibbs sampling: sample from a joint posterior  $p(a, b, c|y)$  each coordinate is drawn from a marginal posterior  $p(a|(b, c), y)$ ,  $p(b|(a, c), y)$ ,  $p(c|(a, b), y)$ .

If full conditional distributions are not recognizable, we need to use other sampling algorithms like the Metropolis-Hasting algorithm.



# Bayesian approach illustration – Gibbs sampling

---

1. Initialize parameter values (e.g.  $\theta_1 = \theta_2 = \theta_3 = 0$ )

2. Iterate through the following operation

- a. Sample the first parameter given other parameter (initial values) and data

$$\theta_1^{[1]} \sim p(\theta_1 | \theta_2^{[0]}, \theta_3^{[0]}, y)$$

+ (b) Store the  $\theta_1^{[1]}$  and update the other conditional distributions with  $\theta_1^{[1]}$

- c. Sample the second parameter given other parameter (initial values), first parameter updated value and data

$$\theta_2^{[1]} \sim p(\theta_2 | \theta_1^{[1]}, \theta_3^{[0]}, y)$$

+ (d) Store the  $\theta_2^{[1]}$  and update the other conditional distributions with  $\theta_2^{[1]}$

- e. Sample the third parameter given other (updated) parameters and data

$$\theta_3^{[1]} \sim p(\theta_3 | \theta_1^{[1]}, \theta_2^{[1]}, y)$$

+ (f) Store the  $\theta_3^{[1]}$  and update the other conditional distributions with  $\theta_3^{[1]}$

- g. repeat the sampling and storing operation for all parameters and  $N$  iteration of the chain

A burnin can be applied to remove initial values before expected convergence and a thinning rate can applied to avoid auto-correlation between points.

# Bayesian approach illustration – Gibbs sampling

---

Illustration from Rizzo M. L. (2008) Statistical Computing with R

Gibbs sampler is often used when multivariate normal distribution is the target distribution. We can illustrate the Gibbs sampler through the sampling of points from a bivariate normal distribution via sampling of the conditional distribution. Therefore, the target distribution is

$$p(x_1, x_2) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1}(x - \mu)\right)$$

Where,  $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$  and  $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$

The conditional distribution of  $x_1|x_2$  is a univariate normal distribution with

$$E[x_1|x_2] = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2)$$

$$V[x_1|x_2] = (1 - \rho^2) \sigma_1^2$$

Therefore the two proposal distributions are

$$f(x_1|x_2) \sim N\left(\mu_1 + \frac{\rho\sigma_1(x_2 - \mu_2)}{\sigma_2}, (1 - \rho^2)\sigma_1^2\right)$$

$$f(x_2|x_1) \sim N\left(\mu_2 + \frac{\rho\sigma_2(x_1 - \mu_1)}{\sigma_1}, (1 - \rho^2)\sigma_2^2\right)$$

We assume that the parameters  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$ ,  $\sigma_2$  and  $\rho$  are known.

# Bayesian approach illustration – Gibbs sampling

---

## 0. Set parameters and constant

```
# 0. create space
N <- 5000
burn <- 1000
X <- matrix(0, nrow = N, ncol = 2)

# 1. Define the parameters. We assume that the parameters are known.
mu1 <- 0
mu2 <- 2
s1 <- 1
s2 <- 0.5
r <- -0.75

# those parameters do not vary
sd1 <- sqrt((1 - (r^2))) * s1
sd2 <- sqrt((1 - (r^2))) * s2
```

## 1. Initialize the chain

```
# 2. Initialise the chain (with the expectation values)
X[1, c(1, 2)] <- c(mu1, mu2)
```

# Bayesian approach illustration – Gibbs sampling

---

2. Sample successively in the conditional distribution and update the parameter values

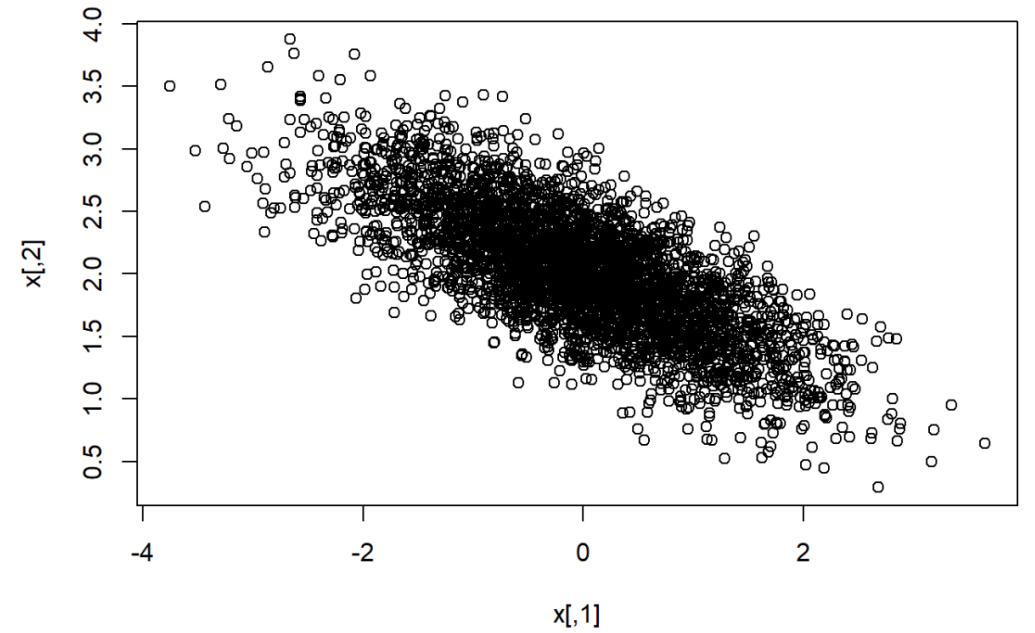
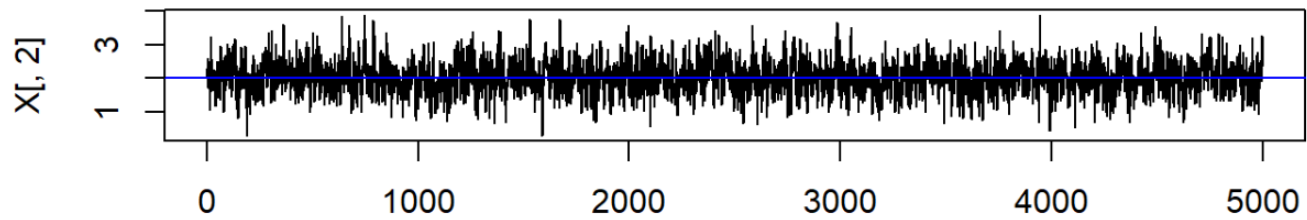
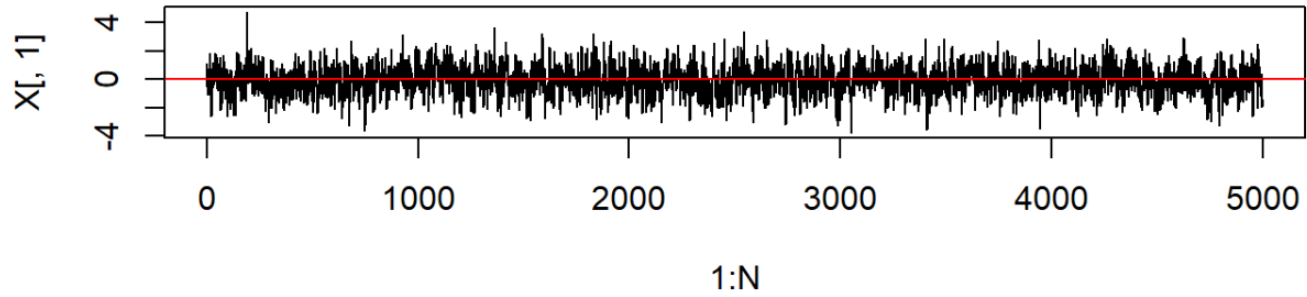
```
for(i in 2:N){  
  
  # generate variable from proposal distribution  
  
  # x1  
  x2 <- X[i-1, 2] # condition on the previous value of x2  
  m1 <- mu1 + (r * (s1/s2) * (x2 - mu2))  
  X[i, 1] <- rnorm(n = 1, mean = m1, sd = sd1)  
  
  # x2  
  x1 <- X[i, 1] # condition on the new value of x1 (sampled in the last run)  
  m2 <- mu2 + (r * (s2/s1) * (x1 - mu1))  
  X[i, 2] <- rnorm(n = 1, mean = m2, sd = sd2)  
  
}
```

3. Check the convergence of the distribution

```
par(mfrow=c(2, 1))  
plot(x = 1:N, y = X[, 1], type = "l")  
abline(h = mu1, col = "red")  
plot(x = 1:N, y = X[, 2], type = "l")  
abline(h = mu2, col = "blue")
```

# Bayesian approach illustration – Gibbs sampling

---



# Bayesian alphabet (A, B, C)

---

**Bayesian** application of the regression model with prior reflecting knowledge on genetic architecture.

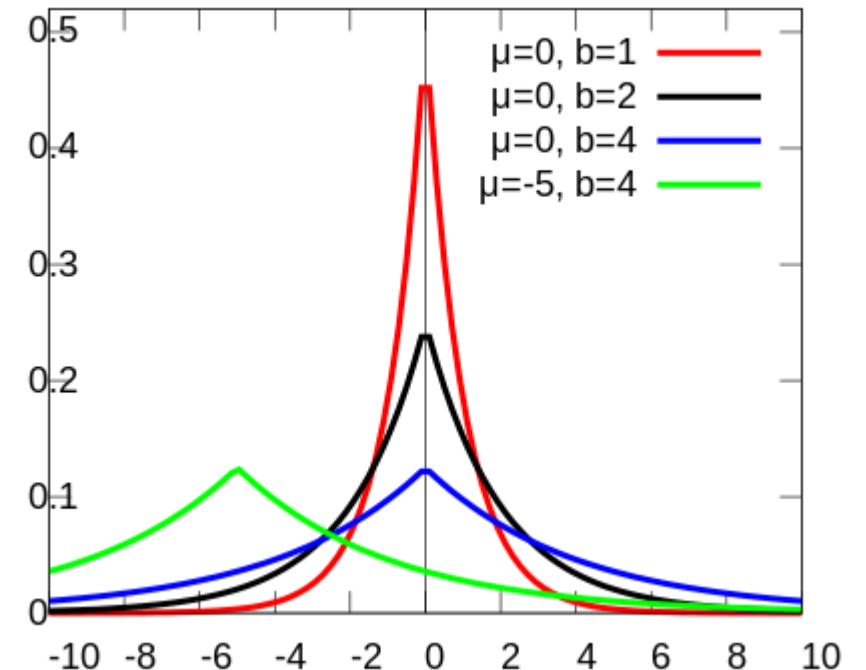
Prior distribution as a way to

- Integrate biological knowledge
- Perform shrinkage/variable selection

Prior of the models studied so far:

**GBLUP** (Ridge regression): All markers have a small non-zero effect. The value of those effects is normally distributed. Bayesian version of the GBLUP use a gaussian (Normal) prior

**LASSO** : Maximum  $n$  markers have a non-zero effect. The effect is centered on a reduced number of markers. Bayesian version of the LASSO uses a double exponential prior.



Double exponential (Laplace) distribution



# Bayes A – Model (likelihood) definition

---

The Bayes A approach assume the same regression model as the Ridge regression

$$y = \mu + X\beta + e$$
$$y = \mu + \sum_{j=1}^p X_j\beta_j + e$$

Where  $X$  represent the matrix with genotype marker scores and  $\beta = (\beta_1, \dots, \beta_p)'$  the vector of marker effects.

The Bayesian approach often use the same model as frequentist approach (e.g. linear regression, mixed model). A major difference is that it considers all parameters as random.

## Hierarchichal model definition

The Bayesian approach gives some flexibility about parameter definition and model specification. For example in Bayes A, we allow the marker effect  $\beta$  to vary via the modeling of their variance  $\sigma_{\beta_i}^2$ . This strategy request to define models at two levels at the data levels and at the SNP variance level

$$y = \mu + \sum_{j=1}^p X_j\beta_j + e$$
$$\beta_j \sim N(0, \sigma_{\beta_i}^2)$$
$$\sigma_{\beta_i}^2 \sim \chi^{-2}(v, S)$$
$$e \sim N(0, \sigma_e^2)$$

Therefore, at the first stage, the model likelihood is

$$y|\mu, \beta_j, \sigma_e^2 \sim N(\mu + X\beta, I\sigma_e^2)$$

# Bayes A – Prior elicitation

---

$\mu$  is assumed to have a “flat” uninformative prior (e.g. uniform distribution).

$$\mu \sim U(a, b)$$

Approaches like the GBLUP assume that all locus explain a small equal fraction of the genetic variance  $\frac{\sigma_g^2}{n_{SNP}}$ . RR-BLUP model assume that marker effects are normally distributed and that they all have a small non-zero effect. The Bayesian approach allow to modify the prior information considering that marker effect can be modulated by differences of linkage and that some markers have no effect.

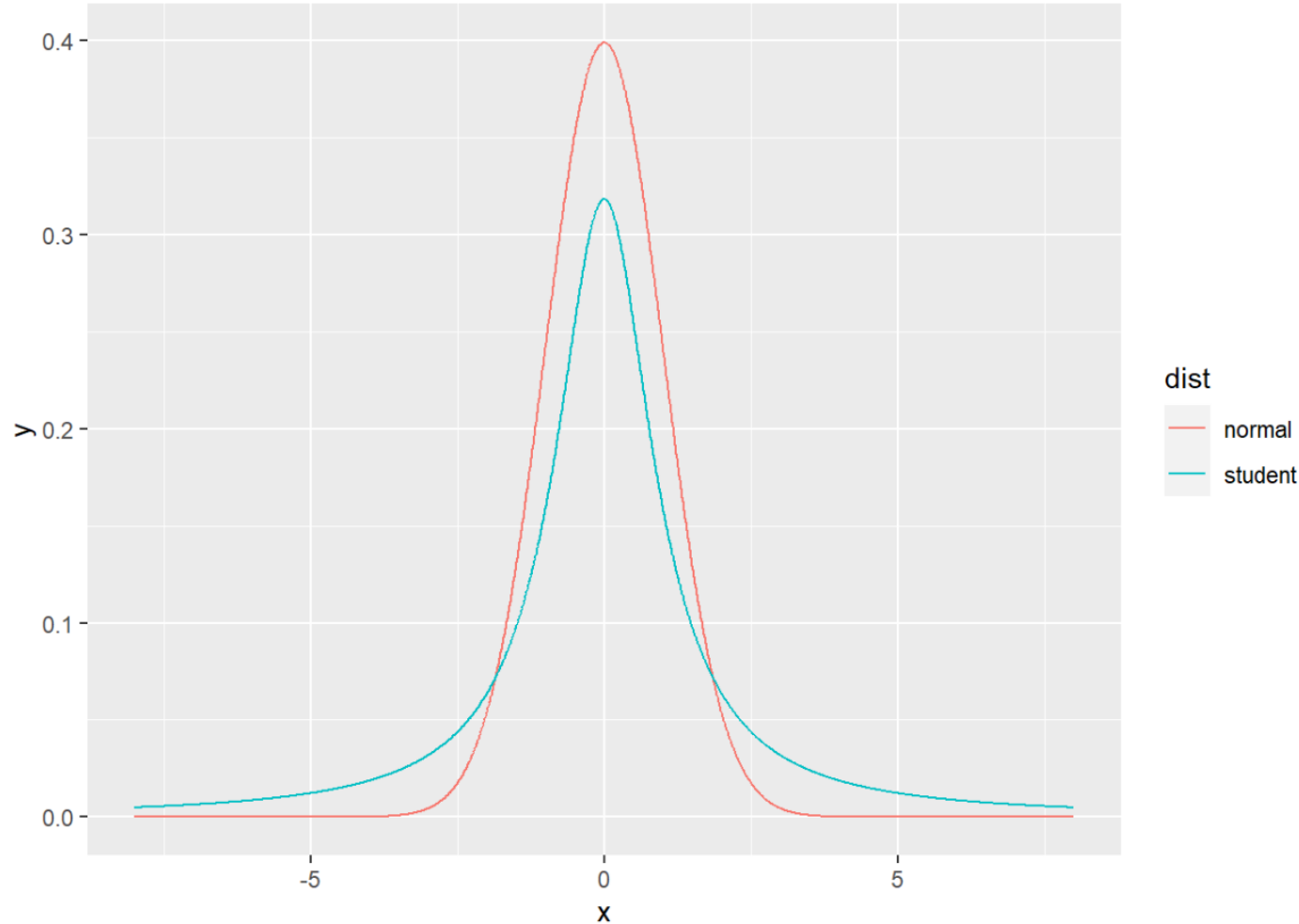
An alternative possibility is to assume that the marker effects follow a  $t$  distribution with “thicker” tail, which means that the probability of medium to large effects is bigger than in a normal distribution.

$$\beta_j | \sigma_{\beta_j}^2 \sim N(0, \sigma_{\beta_j}^2)$$

The distribution of the  $\beta_j$  is conditioned on the variance of the QTL effect  $\sigma_{\beta_j}^2$ . For that term, we assume a scaled inverse chi-squared distribution. This prior assumption is useful because it is a conjugate prior (Normal \* scaled-Inv chi-squared = scaled-Inv chi-squared)

$$\sigma_{\beta_j}^2 \sim \chi^{-2}(v_\beta, S_\beta)$$

# Bayes A – Prior elicitation



$$\beta_j | \sigma_{\beta_j}^2 \sim N(0, \sigma_{\beta_j}^2)$$

$$\sigma_{\beta_j}^2 \sim \chi^{-2}(v_\beta, S_\beta)$$

$$v = 4.012 \text{ and } S = 0.0020$$

Meuwissen et al. (2001)

$S$  is an hyper-parameter that control the amount of shrinkage on the marker effects

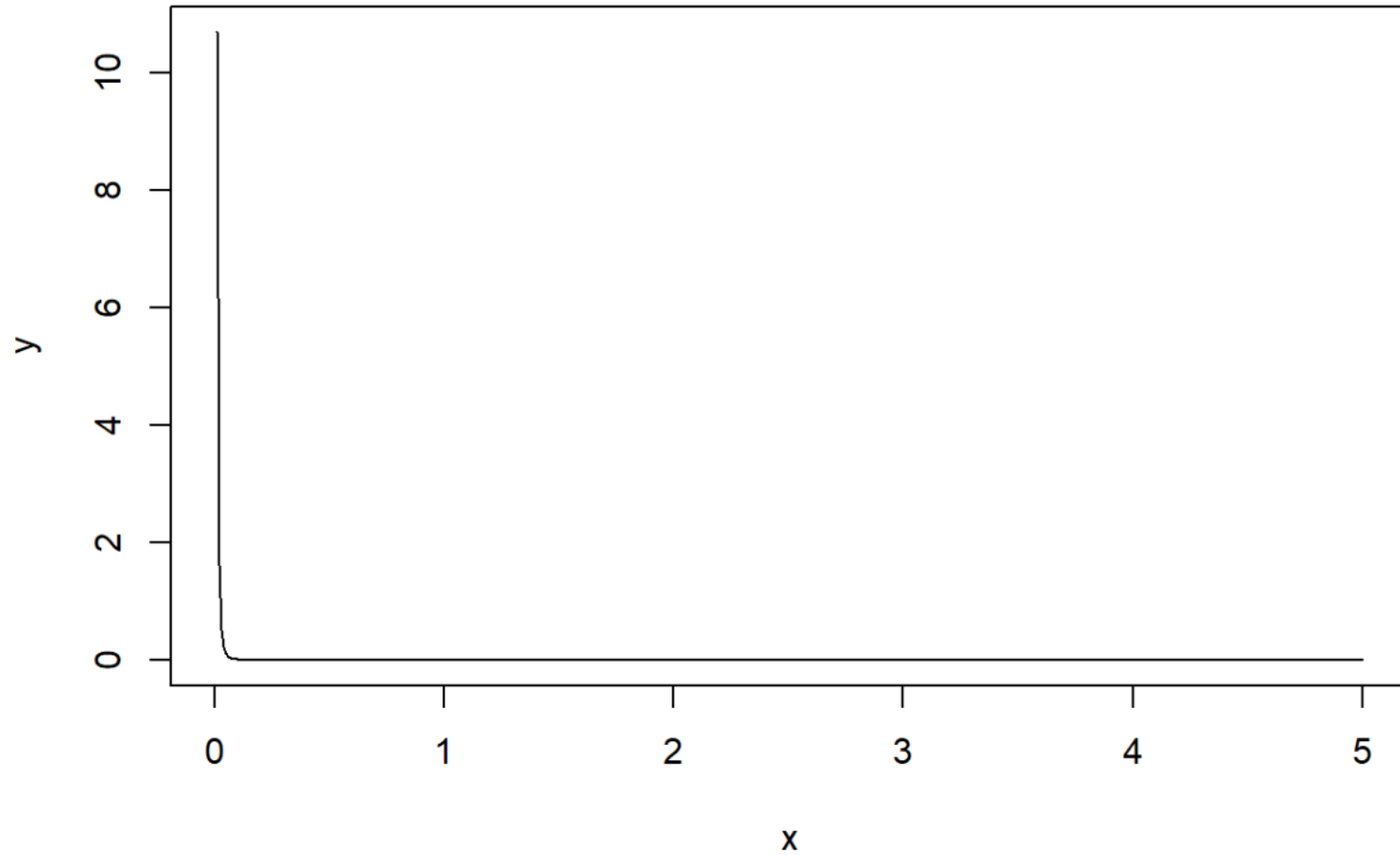
Gianola et al. (2009)

It is possible to calculate the marginal prior of the  $\beta_j$  and show that is correspond to a  $t$  distribution  $\beta_j | v, S \sim t(0, v, S)$

# Bayes A – Prior elicitation

---

$$\sigma_{\beta_j}^2 \sim \chi^{-2}(v_\beta, S_\beta) \quad v = 4.012 \text{ and } \hat{S} = 0.0020$$



# Bayes A – Prior elicitation

---

$\sigma_e^2$  is also assumed to have a “flat” uninformative prior but through a scaled inverse chi-squared distribution. In Meuwissen, Hayes, and Goddard (2001), they propose  $\nu_e = -2$  and  $S_e = 0$ . However,  $\nu_e$  and  $S_e$  must be bigger than 0.

$$\sigma_e^2 \sim \chi^{-2}(\nu_e, S_e)$$

The definition of a prior flat prior through the scaled-inverse chi-squared distribution is useful because it is a conjugate prior with the normal distribution.

# Bayes A – Posterior derivation

---

$$\sigma_{\beta_j}^2$$

The choice of a scaled inverted chi-squared distribution for the prior of  $\sigma_{\beta_j}^2$  help to derive the full conditional density because it is also in the shape of a scaled inverted chi-squared distribution (Wang, Rutledge, and Gianola (1993)).

$$p(\sigma_{\beta_j}^2 | \beta_j) = \chi^{-2}(v_{\beta} + n_j, S_{\beta} + \beta_j' \beta_j)$$

with  $S_{\beta} + \beta_j' \beta_j$  scaled parameter and  $v_{\beta} + n_j$  degrees of freedom. Where  $n_j$  is the number of SNP or haplotype segment effects. In Meuwissen, Hayes, and Goddard (2001), they express the conditional distribution of  $\sigma_{\beta_j}^2$  given  $\beta_j$ . This means that it is not possible to directly sample from this distribution and that  $\beta_j$  values must be used in Gibbs sampling.

$$\sigma_e^2$$

Similarly to  $\sigma_{\beta_j}^2$ , the choice of a scaled inverse chi-squared prior facilitate the derivation of the (full) conditional posterior because when it is combeined with the data distribution it is also a scaled inverse chi-squared distribution. In Meuwissen, Hayes, and Goddard (2001), they express the conditional distribution of  $\sigma_e^2$  given  $e$ .

$$p(\sigma_e^2 | e) = \chi^{-2}(n - 2, e' e)$$

Once again, this distribution is conditioned on parameter values ( $e$ ). We can not use it to sample directly from it.

# Bayes A – Posterior derivation

---

$$\sigma_{\beta_j}^2$$

The choice of a scaled inverted chi-squared distribution for the prior of  $\sigma_{\beta_j}^2$  help to derive the full conditional density because it is also in the shape of a scaled inverted chi-squared distribution (Wang, Rutledge, and Gianola (1993)).

$$p(\sigma_{\beta_j}^2 | \beta_j) = \chi^{-2}(v_{\beta} + n_j, S_{\beta} + \beta_j' \beta_j)$$

with  $S_{\beta} + \beta_j' \beta_j$  scaled parameter and  $v_{\beta} + n_j$  degrees of freedom. Where  $n_j$  is the number of SNP or haplotype segment effects. In Meuwissen, Hayes, and Goddard (2001), they express the conditional distribution of  $\sigma_{\beta_j}^2$  given  $\beta_j$ . This means that it is not possible to directly sample from this distribution and that  $\beta_j$  values must be used in Gibbs sampling.

$$\sigma_e^2$$

Similarly to  $\sigma_{\beta_j}^2$ , the choice of a scaled inverse chi-squared prior facilitate the derivation of the (full) conditional posterior because when it is combined with the data distribution it is also a scaled inverse chi-squared distribution. In Meuwissen, Hayes, and Goddard (2001), they express the conditional distribution of  $\sigma_e^2$  given  $e$ .

$$p(\sigma_e^2 | e) = \chi^{-2}(n - 2, e' e)$$

Once again, this distribution is conditioned on parameter values ( $e$ ). We can not use it to sample directly from it.



# Bayes A – Posterior derivation

---

$\mu$

The derivation of the conditional posterior distribution of  $\mu$  require more mathematics, according to Wang, Rutledge, and Gianola (1993), it can be shown that

$$p(\mu|\dots, y) = N(\tilde{\mu}, (X'X)^{-1}\sigma_e^2)$$

with  $\tilde{\mu} = (X'X)^{-1}X'(y - \sum Z_i u_i)$  when the model is defined as  $y = X\beta + \sum Z_i u_i + e$ .

In our case we have  $X\beta = \mathbf{1}\mu$ , and  $X\beta = \sum Z_i u_i$ . Therefore, applied to BayesA situation as defined by Meuwissen, Hayes, and Goddard (2001), it translates in the following conditional posterior distribution

$$p(\mu|\beta, \sigma_e^2, y) = N(\mathbf{1}'y - \mathbf{1}'X\beta, \frac{\sigma_e^2}{n})$$

because  $\tilde{\mu} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'(y - X\beta) = \frac{1}{n}(\mathbf{1}'y - \mathbf{1}'X\beta)$ . In Meuwissen, Hayes, and Goddard (2001), they omitted the  $\frac{1}{n}$ .

# Bayes A – Posterior derivation

---

$\beta_j$

Similarly, the derivation of the posterior full conditional distribution for  $\beta_j$  can be found in Wang, Rutledge, and Gianola (1993).

$$p(u_i | \dots, u_{j \neq i}, \dots, \mathbf{y}) = N(\tilde{u}_i, (Z_i' Z_i + I \frac{\sigma_e^2}{\sigma_{\beta_i}^2})^{-1} \sigma_e^2)$$

with

$$\tilde{u}_i = (Z_i' Z_i + I \frac{\sigma_e^2}{\sigma_{\beta_i}^2})^{-1} Z_i' (\mathbf{y} - X\beta - \sum_{j \neq i} Z_j u_j)$$

when the model is defined as  $\mathbf{y} = X\beta + \sum Z_i u_i + e$ .

If we translated that to the BayesA model such as we defined it, we get

$$\tilde{\beta}_j = (X_j' X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2})^{-1} X_j' (\mathbf{y} - \mathbf{1}_n \mu - X\beta) \quad \text{with} \quad \beta_j = 0$$

and the variance

$$V(\beta_j) = (X_j' X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2})^{-1} \sigma_e^2$$

The posterior distribution is condition on all other marker effect but the  $j$  th one  $p(\beta_i | \dots, \beta_{j \neq i}, \dots, \mathbf{y})$ . This is translated by the fact that for the sampling of  $\beta_j(it + 1)$ , its actual value is set to zero  $\beta_j(it) = 0$ .

# Bayes A – Interpretation

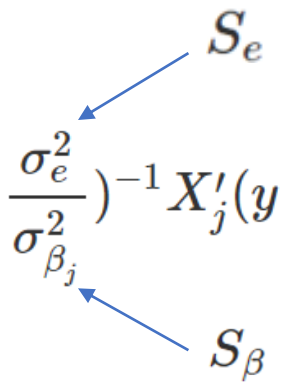
---

$$\tilde{\beta}_j = (X_j'X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2})^{-1} X_j'(y - \mathbf{1}_n\mu - X\beta)$$

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'y$$

# Bayes A – Interpretation

---

$$\tilde{\beta}_j = (X'_j X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2})^{-1} X'_j (y - \mathbf{1}_n \mu - X\beta)$$


The diagram shows two blue arrows. One arrow points from the symbol  $S_e$  to the  $\sigma_e^2$  term in the numerator of the fraction in the equation. The other arrow points from the symbol  $S_{\beta}$  to the  $\sigma_{\beta_j}^2$  term in the denominator of the fraction.

$$\hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'y$$

Similar form between beta Ridge and the mean of the conditional distribution of beta.

$S_e$  and  $S_{\beta}$ , determine the level of shrinkage applied to the beta for the estimation.

# Bayes A – Gibbs sampling

---

The Gibbs sampler is used as a method of numerical integration to get marginal distribution from posterior full conditional distribution. It goes through the following steps:

1. set initial values for the paramters:  $\mu, \beta = \beta_1, \dots, \beta_p, \sigma_\beta^2 = \sigma_{\beta_1}^2, \dots, \sigma_{\beta_p}^2, \sigma_e^2$
2. Generate new values for  $\sigma_e^2(it + 1)$ . For that you need to use the values of  $e$  because  $\sigma_e^2|e$ . Update  $\sigma_e^2$ .
3. Generate new values for  $\mu(it + 1)$ . Update  $\mu$ .
4. Generate new values for  $\sigma_\beta^2(it + 1)$ . For that you need to use the values of  $\beta$  because  $\sigma_\beta^2|\beta$ . Update  $\sigma_\beta^2$ .
5. Generate new values for  $\beta$ . Update  $\beta$ .
6. Repeat 2-5 for  $N_{iter}$ .

# Bayes A – Illustration in R

---

## R code from Ben Hayes Armidale summer school 2015

Step 2: Generate  $\sigma_e^2$  from its posterior distribution

Before sampling new values of  $\sigma_e^2$ , we need to get the value of  $e$  because the distribution of  $\sigma_e^2$  is conditioned on  $e$  ( $p(\sigma_e^2|e)$ ).

```
# First calculate the vector of residuals that is used in the posterior distribution
e <- y - x**b - mu
# Then use those to sample from the posterior distribution
vare <- (t(e)**e)/rchisq(1, n_ind-2)
```

Step 3: Generate  $\mu$  from its posterior distribution

```
m <- (t(ones)**y - t(ones)**x**b)/n_ind
v <- vare/n_ind
mu <- rnorm(1, mean = m, sd = sqrt(v))
```

Step 4: Generate  $\sigma_\beta^2$  from their posterior distribution

```
for (j in 1:n_mk) {
  # bvar[j] <- (0.002+b[j]*b[j])/rchisq(1,4.012+1) # Meuwissen et al. (2001) prior
  # bvar[j] <- (b[j]*b[j])/rchisq(1,1) # Xu (2003) prior
  bvar[j] <- (b[j]*b[j])/rchisq(1, 0.998) # Te Braak et al.(2006) prior
}
```

# Bayes B – model (Likelihood definition)

---

The Bayes B model is build on the same hierarchical type of hierarchichal regression model as Bayes A.

where,

$$y = \mu + \sum_{j=1}^p X_j \beta_j + e$$
$$\beta_j \sim N(0, \sigma_{\beta_i}^2)$$
$$\sigma_{\beta_i}^2 \sim f(\pi, v, S)$$



# Bayes B – prior elicitation

---

A possible assumption about the QTL effects distribution is that many SNPs are not in genomic regions containing QTL, and therefore have no effect. This is not accounted by Bayes A because the mass at 0 is infinitesimal. Therefore, Meuwissen, Hayes, and Goddard (2001) proposed another prior with mass at 0.

$$\sigma_{\beta_i}^2 | \pi, v, S = \begin{cases} 0 & \text{with probability } \pi \\ \chi^{-2}(v, S) & \text{with probability } 1 - \pi \end{cases}$$

This prior is a mixture distribution with a large proportion of effects ( $\pi$ ) at zero and the rest following a chi-squared inverse distribution, which gives a marginal prior as a t distribution (cf Bayes A). The marginal prior of  $\beta_j$  in the  $\pi$  case is 0.

The Bayes B model is a special case of Bayes A model with  $\pi = 0$ .  $\pi$  is assumed to be fixed a priori by the user (e.g.  $\pi = 0.95$ ).

# Bayes B – posterior derivation

---

The posterior conditional of  $\sigma_e^2$  and  $\mu$  are the same as in Bayes A and can be sampled using Gibbs sampler because the distribution are identified.

Concerning  $\sigma_{\beta_j}^2$  and  $\beta_j$ , the Gibbs sampler will not move through the entire space of  $\beta_j$  because the probability of  $\sigma_{\beta_j}^2 = 0$  is not possible if  $\beta_j' \beta_j > 0$  and the probability of  $\beta_j = 0$  is infinitesimal if  $\sigma_{\beta_j}^2 > 0$ .

To be able to sample values that cover the sampling space of  $\beta_j$  we can sample simultaneously  $\sigma_{\beta_j}^2$  and  $\beta_j$  from the joint distribution distribution

$$p(\sigma_{\beta_j}^2, \beta_j | \mathbf{y}^*) = p(\sigma_{\beta_j}^2 | \mathbf{y}^*) \times p(\beta_j | \sigma_{\beta_j}^2, \mathbf{y}^*)$$

The strategy to sample from  $p(\sigma_{\beta_j}^2, \beta_j | \mathbf{y}^*)$  and (if the candidate point is accepted) to sample from  $p(\beta_j | \sigma_{\beta_j}^2, \mathbf{y}^*)$ .

The sampling from  $p(\sigma_{\beta_j}^2 | \mathbf{y}^*)$  using the Gibbs sampling is not possible because it can not be expressed as a closed distribution. Therefore, we need to use the Metropolis-Hasting sampler. in that case we will implement the MH independence sampler.

# Bayes B – MCMC sampling – MH independance sampler

---

$p(\sigma_{\beta_j}^2 | \mathbf{y}^*)$  is the target distribution

$p(\sigma_{\beta_j}^2) \sim \chi^{-2}(v, S)$  is the proposal distribution.

Iterate over the markers. For marker  $j$

0. calculate  $\mathbf{y}^*$ , the corrected phenotype values for the mean and all marker effects except  $j$ .

$$\mathbf{y}^* = \mathbf{y} - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}$$

with  $\beta_j = 0$

1. sampling of  $r_i \sim U[0, 1]$ . if  $r_i < \pi$ ,  $\sigma_{\beta_j}^2 = 0$  and  $\beta_j = 0$ , otherwise:
2. sample a candidate point using  $\sigma_{\beta_j}^{2(new)} \sim \chi^{-2}(v, S)$
3. Use the MH independance sampler probability rule for point acceptance. (note: the connection between the likelihood and the forma definition of the independance sample is not fully explicit).

$$\alpha\left(1; \frac{p(\mathbf{y}^* | \sigma_{\beta_j}^{2(new)})}{p(\mathbf{y}^* | \sigma_{\beta_j}^{2(old)})}\right)$$

The acceptance probability depends on the ratio of the likelihood of the corrected data given the new and old parameter.

4. If  $\sigma_{\beta_j}^2 > 0$ , sample  $\beta_j$  from the same normal distribution as in Bayes A

$$p(\beta_j | \sigma_{\beta_j}^2) \sim N\left(\left(X_j'X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2}\right)^{-1}X_j'(\mathbf{y} - \mathbf{1}_n\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}); \left(X_j'X_j + \frac{\sigma_e^2}{\sigma_{\beta_j}^2}\right)^{-1}\sigma_e^2\right)$$

# Bayesian models for genomic prediction – Bayes C

---

The Bayes  $C_\pi$  proposed by Habier et al. (2011) is also based on a similar type of regression model

$$y = \mu + \sum_{j=1}^p X_j \beta_j + e$$

However, the effects of the SNP are assumed to have a “common” variance. They are also assumed to have a non zero effect with a probability  $\pi$

$$\beta_j | \pi, \sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi \\ N(0, \sigma_\beta^2) & \text{with probability } 1 - \pi \end{cases}$$
$$\pi \sim \text{Unif}[0, 1]$$

Contrary to the Bayes B model, the probability of non-zero effect  $\pi$  is considered as a variable with a uniform prior to let this probability be estimated from the model.

the “constant” SNP variance  $\sigma_\beta^2$  get a scaled inverse chi-squared distribution with  $v = 4.2$  and  $S$  determined from the data

$$\sigma_\beta^2 | v, S \sim \chi^{-2}(v, S)$$

where,

$$S = \frac{\tilde{\sigma}_a^2 (v - 2)}{v} \quad \text{and}$$
$$\sigma_a^2 = \frac{\tilde{\sigma}_s^2}{(1 - \pi) \sum_{k=1}^K 2p_k(1 - p_k)}$$

# Properties of Bayesian approach

---

- Integration of several sources of knowledge: prior and data.
- Prior knowledge encapsulated in a prior distribution can reflect biological hypotheses about the studied phenomenon.
- Take all sources of uncertainty into consideration during the estimation
- Get a probabilistic interpretation of the estimated parameter
- Natural and flexible way to realize shrinkage and variable selection