A brief introduction to Bayesian statistics

Emmanuel Grolleau
Observatoire de Paris – LESIA – Service d'Informatique Scientifique

10/2023

Summary

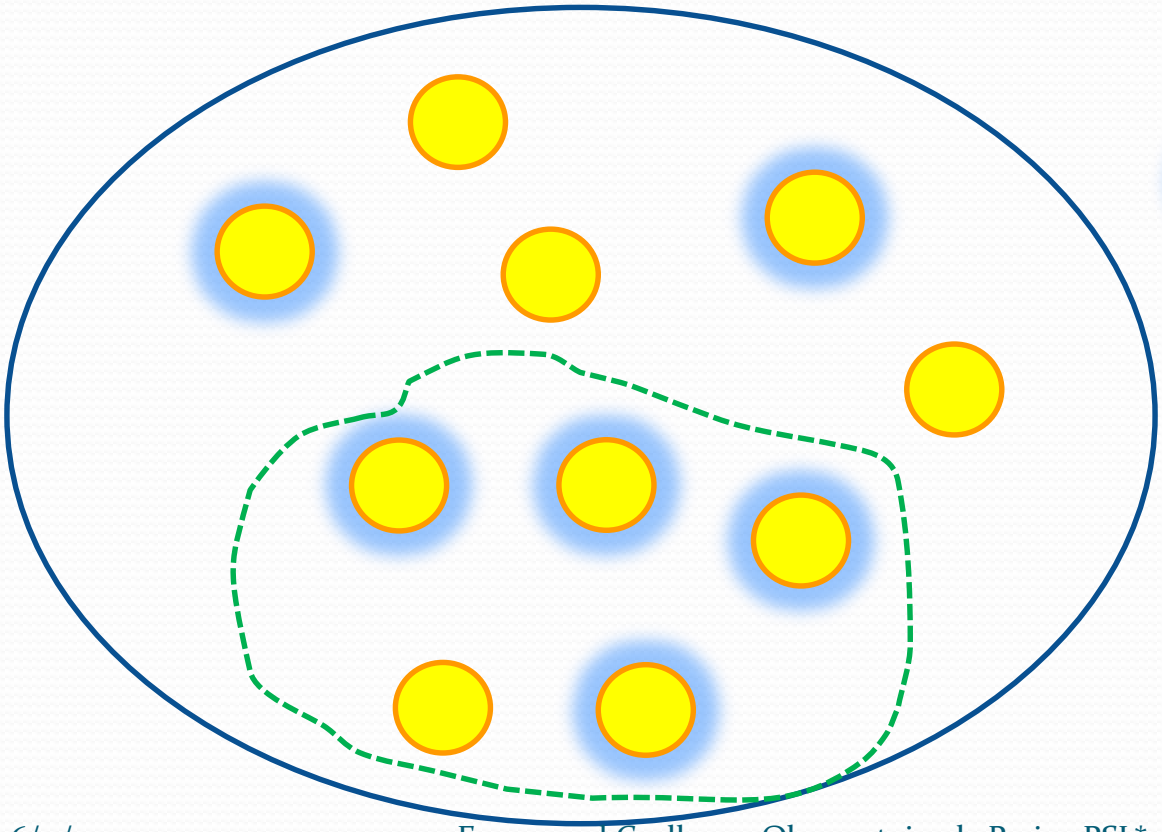
- Bayes' theorem
- Linear regression problem
- Statistical inference
- Frequentist approach
- Bayesian approach
- Bayesian Neural Networks

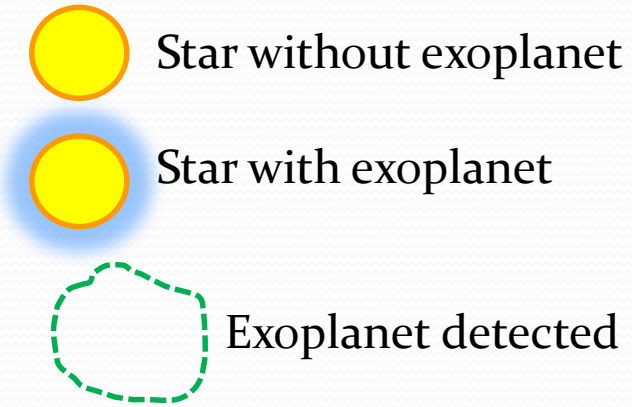
Summary

- **Bayes' theorem**
- Linear regression problem
- Statistical inference
- Frequentist approach
- Bayesian approach
- Bayesian Neural Networks

Bayes' theorem

- *Observing 10 stars, we use a method to detect stars with exoplanet.*

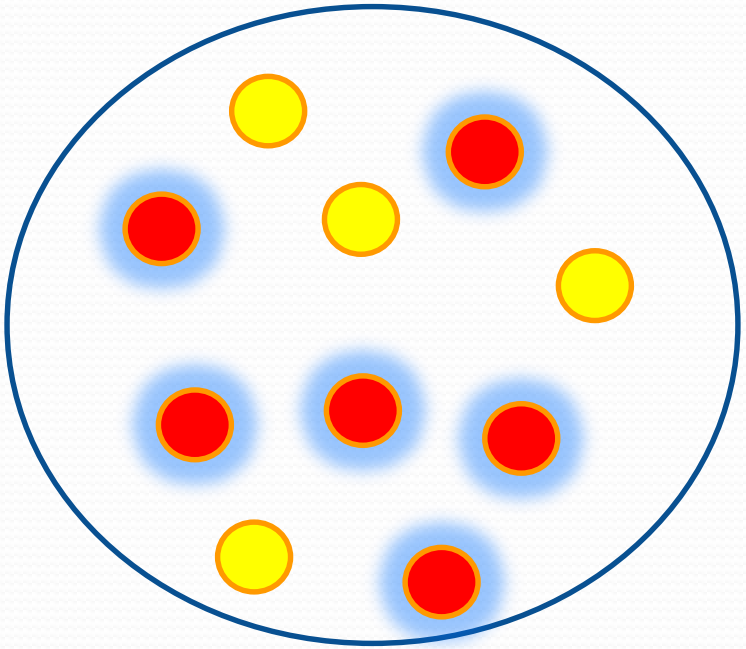




Bayes' theorem

P(Stars with exoplanet)

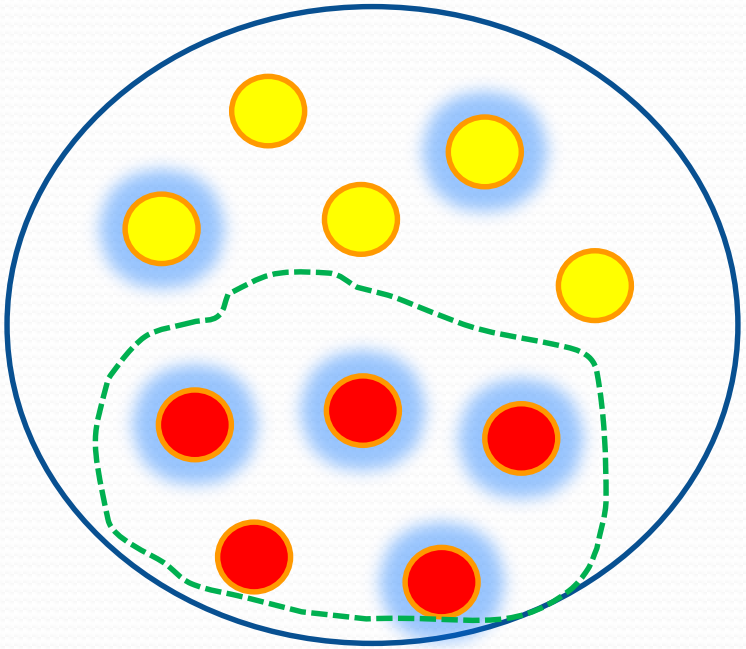
$$P(E) = \frac{6}{10}$$



Bayes' theorem

P(Method detects star with exoplanet)

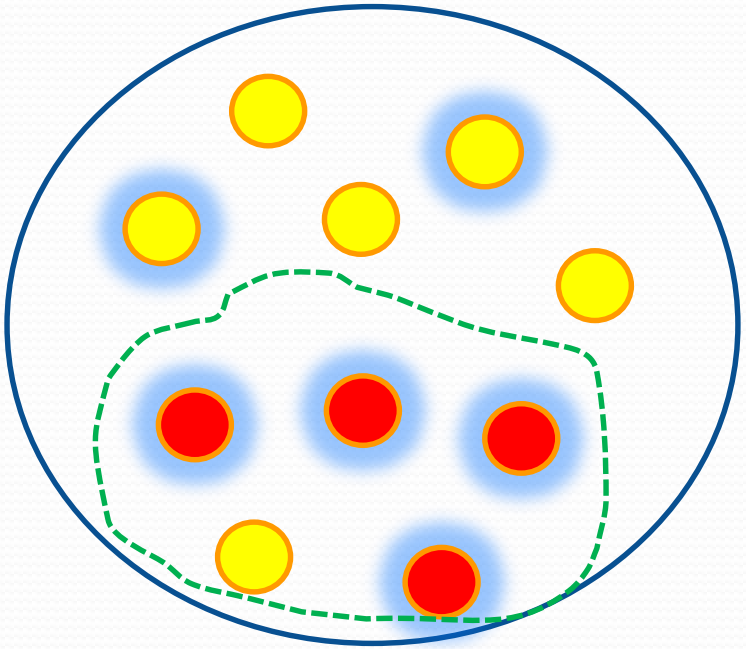
$$P(D) = \frac{5}{10}$$



Bayes' theorem

P(Method detects star with exoplanet and star having exoplanet)

$$P(D \text{ and } E) = \frac{4}{10}$$



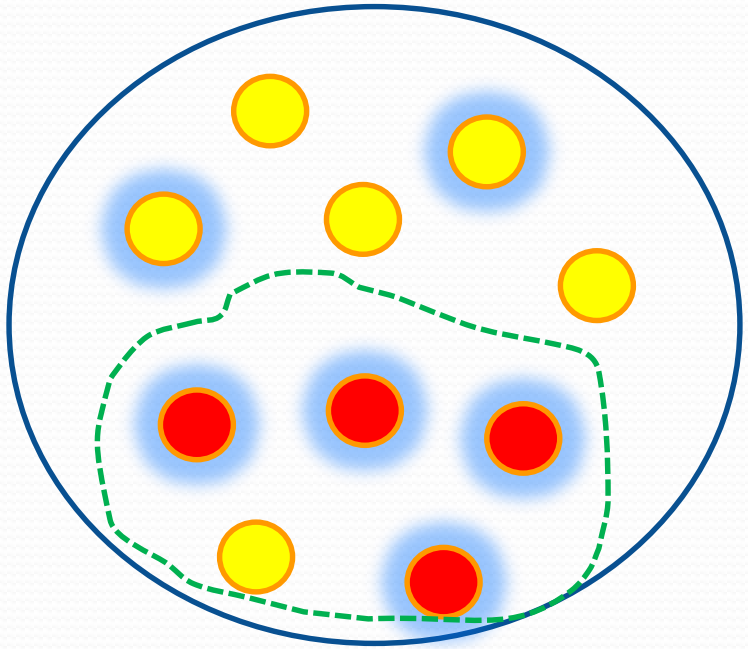
Bayes' theorem

<u>Conditional probability:</u>
P(Stars having exoplanet providing that the method detected exoplanet stars)

$$P(E|D) = \frac{P(D\ and\ E)}{P(D)}$$

$$= \frac{\frac{4}{10}}{\frac{5}{10}} = \frac{4}{5}$$

$$if\ P(D\ and\ E) \approx P(D) \Leftrightarrow P(E|D) \approx 1$$

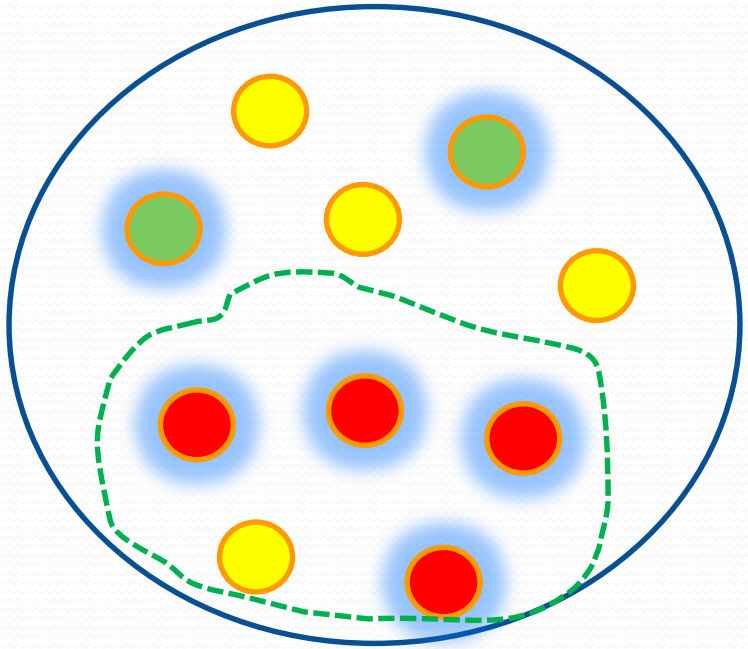


Bayes' theorem

Conditional probability:
P(Detected exoplanet stars providing that stars have exoplanet)

$$P(D|E) = \frac{P(D \text{ and } E)}{P(E)}$$

$$= \frac{\frac{4}{10}}{\frac{6}{10}} = \frac{4}{6}$$



=> ***method sensitivity***

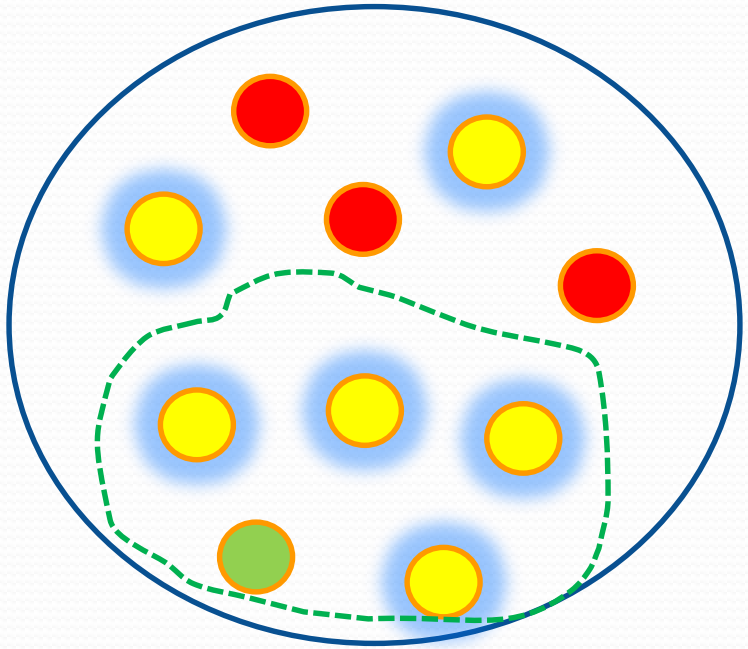
Bayes' theorem

Conditional probability: *P(**Not** detected exoplanet star providing that star does not have exoplanet)*

$$P(\bar{D}|\bar{E}) = \frac{P(\bar{D} and \bar{E})}{P(\bar{E})}$$

$$= \frac{\frac{3}{10}}{\frac{4}{10}} = \frac{3}{4}$$

=> ***method specificity***



Bayes' theorem

$$P(D|E) = \frac{P(D\ and\ E)}{P(E)} \Leftrightarrow P(D\ and\ E) = P(D|E) * P(E)$$

$$P(E|D) = \frac{P(D\ and\ E)}{P(D)} \Leftrightarrow P(D\ and\ E) = P(E|D) * P(D)$$

$$\Leftrightarrow P(E|D) * P(D) = P(D|E) * P(E)$$

$$\boldsymbol{P(E|D) = \frac{P(D|E) * P(E)}{P(D)}}$$

Bayes' theorem : application

- *Observing 10 stars, we use a method to detect stars with exoplanet.*
- *This method has the following performance:*
  - *Sensitivity: When a star has an exoplanet, the method designates this star as an exoplanet star with a probability of 90%, $P(D|E)$*
  - *When a star does not have an exoplanet, the method designate this star as an exoplanet star with a probability of 1%, $P(D|\bar{E})$*
- *What is the probability that a star has really an exoplanet if the method designates this star as having an exoplanet $P(E|D)$?*

Bayes' theorem : application

$$P(D|E) = 0{,}9$$
$$P(D|\bar{E}) = 0{,}01$$

$$P(E|D) = \frac{P(D|E) * P(E)}{P(D)} = \frac{0{,}9 * P(E)}{P(D)}$$

$$P(D) = P(D|E) * P(E) + P(D|\bar{E}) * P(\bar{E})$$

$$P(D) = P(D|E) * P(E) + P(D|\bar{E}) * \left(1 - P(E)\right)$$

$$P(D) = \left(P(D|E) - P(D|\bar{E})\right) * P(E) + P(D|\bar{E})$$

$$P(E|D) = \frac{P(D|E) * P(E)}{\left(P(D|E) - P(D|\bar{E})\right) * P(E) + P(D|\bar{E})}$$

Bayes' theorem : application

$$P(D|E) = 0{,}9$$
$$P(D|\bar{E}) = 0{,}01$$

$$P(E|D) = \frac{P(D|E) * P(E)}{(P(D|E) - P(D|\bar{E})) * P(E) + P(D|\bar{E})}$$

$$P(E|D) = \frac{0{,}9 * P(E)}{0{,}89 * P(E) + 0{,}01}$$

What is the probability that a star has really an exoplanet if the method designates this star as having an exoplanet ?

The answer totally depends on the probability of exoplanets in the universe !!!

and we don't know this probability…

Bayes' theorem : application

$$P(E|D) = \frac{0{,}9 * P(E)}{0{,}89 * P(E) + 0{,}01}$$

What is the probability that a star has really an exoplanet if the method designates this star as having an exoplanet ?

The answer totally depends on the probability of exoplanets in the universe !!!

and we don't know this probability…

If $P(E) = 1e^{-5}$ *(one star over 100 000 has exoplanet) then*

$$P(E|D) = \frac{0{,}9 * P(E)}{0{,}89 * P(E) + 0{,}01} = 0{,}000899 \Leftrightarrow \mathbf{0{,}0899\ \%}$$

Bayes' theorem : application

- Let's say that we look at 1 000 000 stars. If $P(E) = 1e^{-5}$, one star over 100 000 has exoplanet, then:
  - 10 stars have exoplanet
  - 9 stars with exoplanet will be detected by our method
  - Over 999 990 star with no exoplanet, 1 % , approximatively 10 000 will be false positives.
  - So we will have 10 009 star detected with exoplanet but only 9 true positives !

Summary

- Bayes' theorem
- **Linear regression problem**
- Statistical inference
- Frequentist approach
- Bayesian approach
- Bayesian Neural Networks

Linear regression problem

Hubble-Lemaître's law

$$H_0 * d = v$$

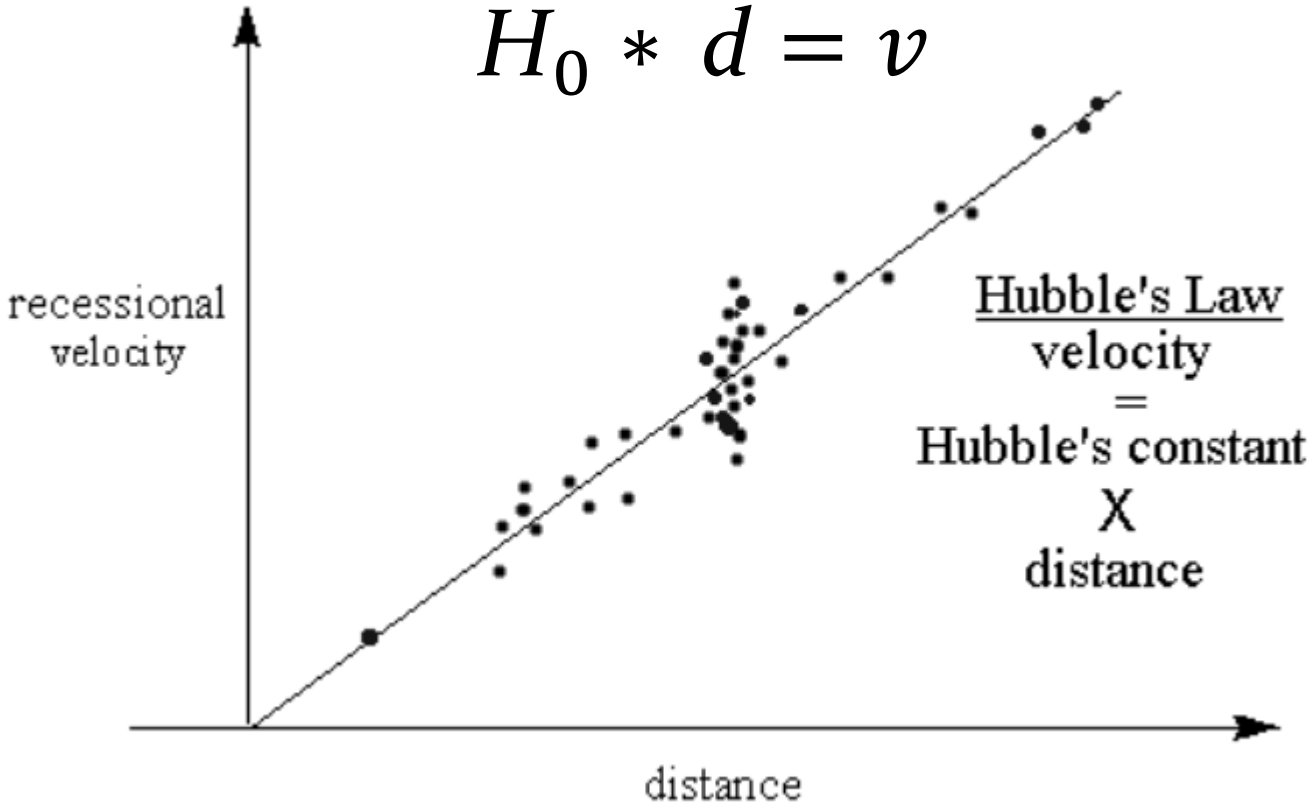
- H: Hubble constant
- d: distance of the galaxy from the Earth
- v: velocity of the galaxy

We want to estimate H_0, we have observed a lot of pair (d, v)

Linear regression problem

Hubble-Lemaître's law

$$H_0 * d = v$$



Summary

- Bayes' theorem
- Linear regression problem
- **Statistical inference**
- Frequentist approach
- Bayesian approach
- Bayesian Neural Networks

Statistical inference

- *Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.*

- Two schools from 18th century:
  - **Frequentist** inference (classical): repeated sampling
  - **Bayesian** inference: Bayesian inference uses the available posterior beliefs as the basis for making statistical propositions

Summary

- Bayes' theorem
- Linear regression problem
- Statistical inference
- **Frequentist approach**
- Bayesian approach
- Bayesian Neural Networks

Statistical inference

Let's toss a coin 10 times.

Frequentist approach: If, for example, we get tails 6 times out of 10, then the probability of getting tails from the results of this experiment is equal to 6/10 = 0.6.

Linear regression problem
Frequentist approach

- **Hubble-Lemaître's law** $H_0 * d = v$

We want to estimate H_0

<u>Frequentist approach:</u>

Linear regression model: we use the pairs (d_i, v_i) to find the estimate of the value H_0

$$H_0 * d_i + \varepsilon_i = v_i$$

We call the unobserved deviations from the above equation the errors ε_i

Linear regression problem
Frequentist approach

- $\widehat{\varepsilon_i}$: estimated errors, i.e. differences between actual and predicted values of the dependent variable v_i

$$\widehat{\varepsilon_i} = v_i - H_0 * d_i$$

- We use a least square approach to minimize the sum of estimated squared residuals

$$\mathrm{MIN}\left(\sum_{i=1}^{n} \widehat{\varepsilon_i}^2\right)$$

$$= MIN\left(\sum_{i=1}^{n} (v_i - H_0 * d_i)^2\right)$$

Linear regression problem
Frequentist approach

Frequentist approach:

$$\widehat{H_0} = \frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)\left(v_i - \bar{v}\right)}{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}$$

So, what is the problem with frequentist approach with a linear regression model ?

Linear regression problem
Frequentist approach

Frequentist approach:

So, what is the problem with pure frequentist approach with a linear regression model ?

- How to proceed if we have only few samples (couple velocity/distance) ?
- How to proceed if we need to have an estimation of our parameter as soon as possible without waiting for a lot of sample ? (e.g. SPAM detection).

Statistical inference

Let's toss a coin 10 times.

Frequentist approach: If, for example, we get tails 6 times out of 10, then the probability of getting tails from the results of this experiment is equal to 6/10 = 0.6.

Bayesian approach, we're not interested in this probability, but rather in its a priori distribution. Indeed, if the coin is balanced, then a priori the probability of getting heads is the same as that of getting tails, i.e. 1/2 = 0.5. This a priori probability is obtained from the results of other experiments carried out in the past.

It's obvious that the probability calculated by the frequentist method will converge towards 0.5 if the coin is tossed a significant number of times.

Summary

- Bayes' theorem
- Linear regression problem
- Statistical inference
- Frequentist approach
- **Bayesian approach**
- Bayesian Neural Networks

Linear regression problem
Bayesian approach

$$P(E|D) = \frac{P(D|E) * P(E)}{P(D)}$$

For our problem, the goal of the bayesian statistics is to find the best value of Ho (in terms of probability) to explain the relationship between the pairs of distance and velocity observed.

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

e.g. $\theta = H_0$

Linear regression problem
Bayesian approach

$$\boldsymbol{P(H_0|Y) = \frac{P(Y|H_0) * P(H_0)}{P(Y)}}$$

$\boldsymbol{P(Y)}$ is useless, it doesn't provide any information but simply normalizes the result $P(H_0|Y)$, therefore we can reformulate as:

$$\boldsymbol{P(H_0|Y) \propto P(Y|H_0) * P(H_0)}$$

Bayesian approach Vocabulary

$$\boldsymbol{P(H_0|Y) \propto P(Y|H_0) * P(H_0)}$$

- $P(H_0|Y)$: posterior probability density
- $P(Y|H_0)$: likelihood
- $P(H_0)$: prior probability density

Bayesian approach
Maximum likelihood

$$\boldsymbol{P(H_0|Y) \propto P(Y|H_0) * P(H_0)}$$

$$\mathbf{P(}\boldsymbol{Y}\mathbf{|}\boldsymbol{H_0}\mathbf{)}\textbf{: Maximum likelihood}$$

- This is the frequentist part of the equation

  $H_0 * d = v$

- Proving that we have:

v_0, d_0

v_1, d_1

…

v_i, d_i

- Ordinary least squares (most popular)
- Method of moments $\widehat{H_0} = m1 = E(\frac{v}{d})$
- **Maximum Likelihood**

Bayesian approach
Maximum likelihood

- Let's say that we have an estimated value of $\widehat{H_0}$, what is the probability to get the pairs v_i, d_i:
  - For the first pair $P(v_1|\widehat{H_0}, d_1) = 0{,}02$ (2%), with this $\widehat{H_0}$ value, it is higly unlikely to get the pair v_1, d_1
  - For the second pair $P(v_2|\widehat{H_0}, d_2) = 0{,}95$ (95%), with this $\widehat{H_0}$ value, it is higly probable to get the pair v_2, d_2
  - And so on…
  - For the last pair $P(v_n|\widehat{H_0}, d_n) = 0{,}5$ (50%), with this $\widehat{H_0}$ value, it is possible to get the pair v_n, d_n
- The join probability for this value of $\widehat{H_0}$ is:
  - $P(v|\widehat{H_0}, d) = \prod P(v_i|\widehat{H_0}, d_i) = 0{,}02 * 0{,}95 * \cdots * 0{,}5 = 0{,}0095$

Bayesian approach
Maximum likelihood

- Our goal is to maximise this probability, that is the maximum likelihood, MAX($P(v|\widehat{H_0}, d) = \prod P(v_i|\widehat{H_0}, d_i)$)

$$H_0 * d_i + \varepsilon_i = v_i$$

If ε_i follows a Normal law, then the equation is:

$$P(v_i|\widehat{H_0}, d_i) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(v_i - \widehat{H_0} d_i)^2}{2\sigma^2}\right]$$

Bayesian approach
Prior probability density

$$\boldsymbol{P(H_0|Y) \propto P(Y|H_0) * P(H_0)}$$

$P(H_0)$: Prior probability density

We no longer look at our data. We rely on our experience of the problem, our expertise in the subject.

We have an idea of the value of our parameter $\widehat{H_0}$ and also the uncertainty about this value: σ^2

That is the prior probability, whatever the data, we propose as expert of this field : $P(\underline{H_0}, \underline{{\sigma_{H0}}^2})$

Bayesian approach
Prior probability density

We should resolve $P(Y|H_0) * P(H_0)$ and it will be easier if we choose some special density probability for our prior beliefs, providing that our maximum likelihood probability density follows a Normal law.

Bayesian approach
Posterior probability density

$$\boldsymbol{P(H_0|Y)} \propto P(Y|H_0) * P(H_0)$$

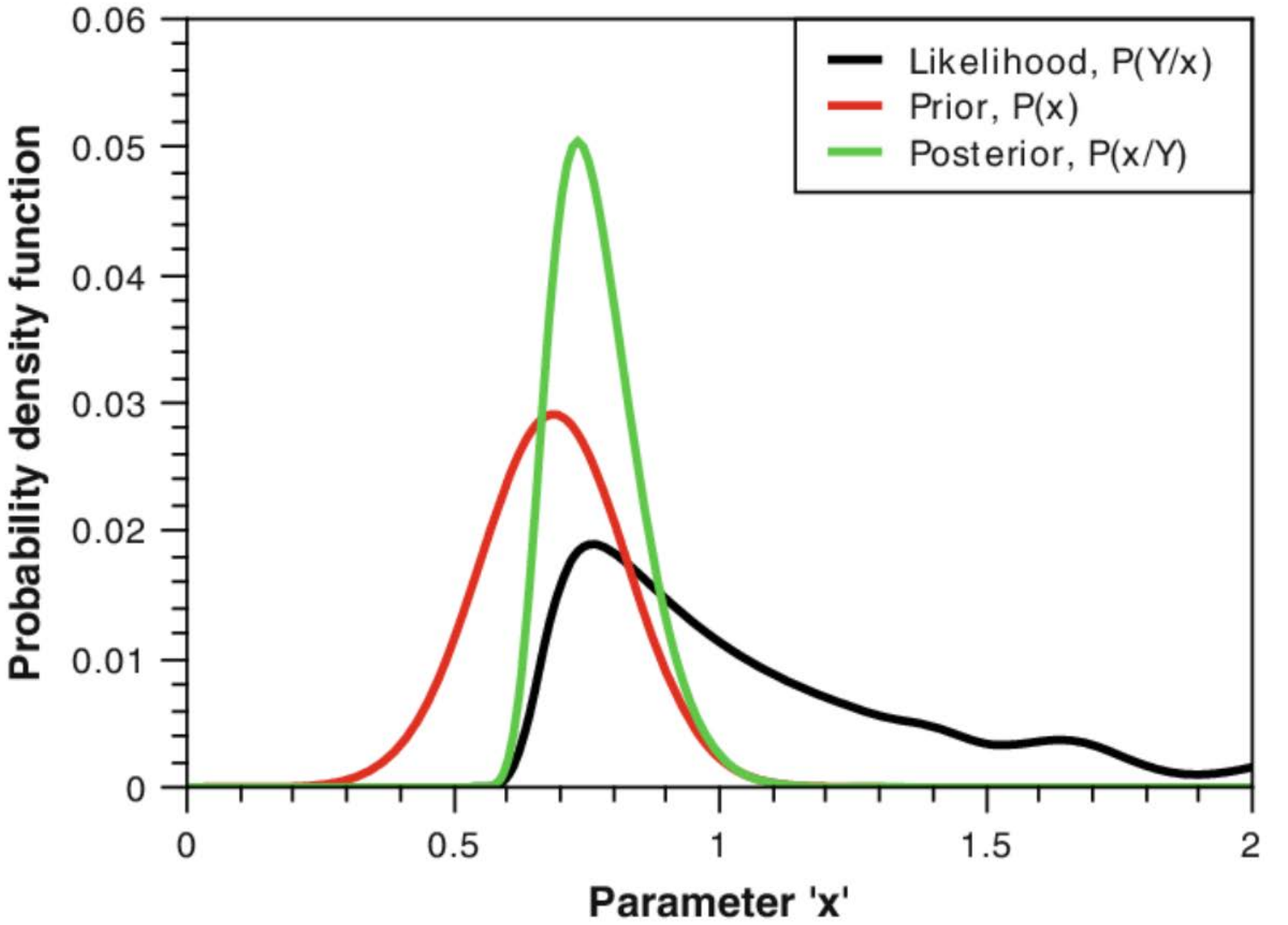
$P(H_0)$: Posterior probability density

The posterior probability density is a combination of our a priori beliefs and our likelihoods derived from the data.

Estimate the posterior probability density is the goal of the Bayesian approach.

The result of our Bayesian approach is then a probability law (a density of probability) not only a simple mean like in frequentist approach. We will be able to determine easily confidence interval for our estimator.

Bayesian approach
Posterior probability density



Bayesian approach
Posterior probability density

$$\boldsymbol{P(H_0|Y)} \propto P(Y|H_0) * P(H_0)$$

$P(H_0)$: Posterior probability density

Thus we can estimate the Esperance of our posterior density probability

$$\boldsymbol{E(H_0|Y) = \int H_0 * P(H_0|Y) dH_0}$$

If we choose the a priori probability function wisely, then we can calculate this expectation analytically, otherwise calculating this integral will be difficult.

Bayesian approach
Conjugate distributions

$$P(H_0|Y) \propto P(Y|H_0) * P(H_0)$$

For certain likelihood functions, selecting a specific prior results in the posterior sharing the same distribution as the prior. This type of prior is then called a conjugate prior.

Bayesian approach
Conjugate distributions

- [Beta posterior]
- Beta prior * Bernoulli likelihood → Beta posterior
- Beta prior * Binomial likelihood → Beta posterior
- Beta prior * Negative Binomial likelihood → Beta posterior
- Beta prior * Geometric likelihood → Beta posterior

- [Gamma posterior]
- Gamma prior * Poisson likelihood → Gamma posterior
- Gamma prior * Exponential likelihood → Gamma posterior

- [Normal posterior]
- Normal prior * Normal likelihood (mean) → Normal posterior

Bayesian approach
Conjugate distributions

- **Hubble-Lemaître's law** $H_0 * d = v$

Observations: H_0(70 km/s/Mpc),d(Mpc),v(km/s)

We want to estimate H_0 on our posterior (i.e. $\overline{H_0}$)

$$P(\overline{H_0}|Y) \propto P\left(Y\middle|\widehat{H_0}\right) * P(\underline{H_0})$$

	Ho Mean	Ho Variance	Density function
Maximum likelihood $\widehat{H_0}$	70 km/s/Mpc	8	$\widehat{H_0} \sim N(70, 2.8)$
Prior probability $\underline{H_0}$	70,5	0,5	$\underline{H_0} \sim N(70.5, 0{,}7)$

Bayesian approach
Conjugate distributions

	Ho Mean	Ho Variance	Density function
Maximum likelihood $\widehat{H_0}$	70 km/s/Mpc	8	$\widehat{H_0} \sim N(70, 2.8)$
Prior probability $\underline{H_0}$	70,5	0,5	$\underline{H_0} \sim N(70.5, 0,7)$

If $prior \propto N\left(\underline{\mu}, \underline{\sigma}^2\right)$ $and\ likelihood \propto N(\hat{\mu}, \hat{\sigma}^2)$ $then\ posterior \propto N(\bar{\mu}, \bar{\sigma}^2)$

$$\bar{\mu} = \frac{n\underline{\sigma}^2}{n\underline{\sigma}^2 + \hat{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{\hat{\sigma}^2}{n\underline{\sigma}^2 + \hat{\sigma}^2}\underline{\mu}$$

$$\bar{\sigma}^2 = \left(\frac{1}{\underline{\sigma}^2} + \frac{n}{\hat{\sigma}^2}\right)^{-1}$$

n: number of measurements

Conjugate Bayesian analysis of the Gaussian distribution. Kevin P. Murphy

Bayesian approach
Conjugate distributions

	Ho Mean	Ho Variance	Density function
Maximum likelihood $\widehat{H_0}$	70 km/s/Mpc	8	$\widehat{H_0} \sim N(70, 2.8)$
Prior probability $\underline{H_0}$	70,5	0,5	$\underline{H_0} \sim N(70.5, 0{,}7)$

If $prior \propto N\left(\underline{H_0}, {\underline{\sigma_{H_0}}}^2\right)$ *and* $likelihood \propto N\left(\widehat{H_0}, {\widehat{\sigma_{H_0}}}^2\right)$ *then* $posterior \propto N\left(\overline{H_0}, {\overline{\sigma_{H_0}}}^2\right)$

$$\overline{H_0} = \frac{n{\underline{\sigma_{H_0}}}^2}{n{\underline{\sigma_{H_0}}}^2 + {\widehat{\sigma_{H_0}}}^2} \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{d_i} + \frac{{\widehat{\sigma_{H_0}}}^2}{n{\underline{\sigma_{H_0}}}^2 + {\widehat{\sigma_{H_0}}}^2} \underline{H_0}$$

$$\overline{\sigma_{H_0}}^2 = \left(\frac{1}{{\underline{\sigma_{H_0}}}^2} + \frac{n}{{\widehat{\sigma_{H_0}}}^2} \right)^{-1}$$

Bayesian approach
Conjugate distributions

$$P(\overline{H_0}|Y) \propto N\left(\overline{H_0}, \overline{\sigma_{H_0}}^2\right)$$

If n is very small, just a few observations, we use mainly the prior

$$\overline{H_0} \sim 0 * \frac{1}{n}\sum_{i=1}^{n}\frac{v_i}{d_i} + \underline{H_0}$$

$$\overline{\sigma_{Ho}}^2 \sim \underline{\sigma_{H_0}}^2$$

Bayesian approach
Conjugate distributions

$$P(\overline{H_0}|Y) \propto N\left(\overline{H_0}, \overline{\sigma_{H_0}}^2\right)$$

If n is very big, a lot of observations, we trust the observations

$$\overline{H_0} \sim \frac{1}{n}\sum_{i=1}^{n}\frac{v_i}{d_i} + 0 * \underline{H_0}$$

$$\overline{\sigma_{Ho}}^2 \sim \frac{\widehat{\sigma_{H_0}}^2}{n}$$

Bayesian approach
Conjugate distributions

$$P(\overline{H_0}|Y) \propto N\left(\overline{H_0}, \overline{\sigma_{H_0}}^2\right)$$

If $\widehat{\sigma_{H_0}}^2 >> \underline{\sigma_{H_0}}^2$, we do not trust the observations, we use mainly the prior

$$\overline{H_0} \sim \frac{small}{n\underline{\sigma_{H_0}}^2 + \widehat{\sigma_{H_0}}^2} \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{d_i} + \frac{big}{n\underline{\sigma_{H_0}}^2 + \widehat{\sigma_{H_0}}^2} \underline{H_0}$$

$$\overline{\sigma_{H_0}}^2 \sim \underline{\sigma_{H_0}}^2$$

Bayesian approach
Conjugate distributions

$$P(\overline{H_0}|Y) \propto N\left(\overline{H_0}, \overline{\sigma_{H_0}}^2\right)$$

If $\underline{\sigma_{H_0}}^2 \gg \widehat{\sigma_{H_0}}^2$, we trust the observations,

$$\overline{H_0} = \frac{big}{n\underline{\sigma_{H_0}}^2 + \widehat{\sigma_{H_0}}^2} \frac{1}{n} \sum_{i=1}^{n} \frac{v_i}{d_i} + \frac{small}{n\underline{\sigma_{H_0}}^2 + \sigma_{H_0}^2} \underline{H_0}$$

$$\overline{\sigma_{H_0}}^2 \sim \frac{\widehat{\sigma_{H_0}}^2}{n}$$

Bayesian approach

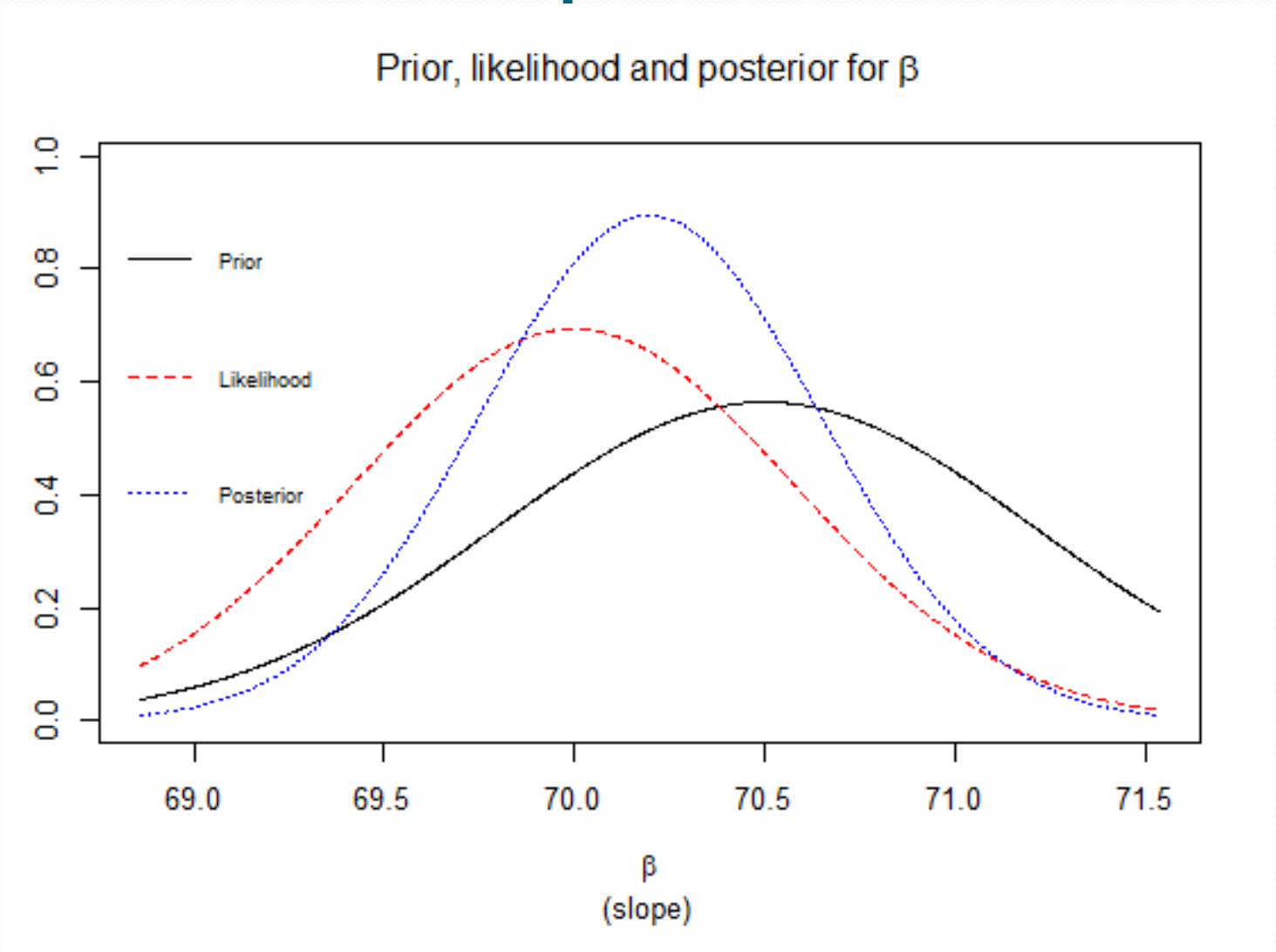
- If we do not choose a conjugate distribution between the prior and the likelihood, then the computation may be very difficult and we should have to do **Monte-Carlo simulation**.

Posterior computation with R

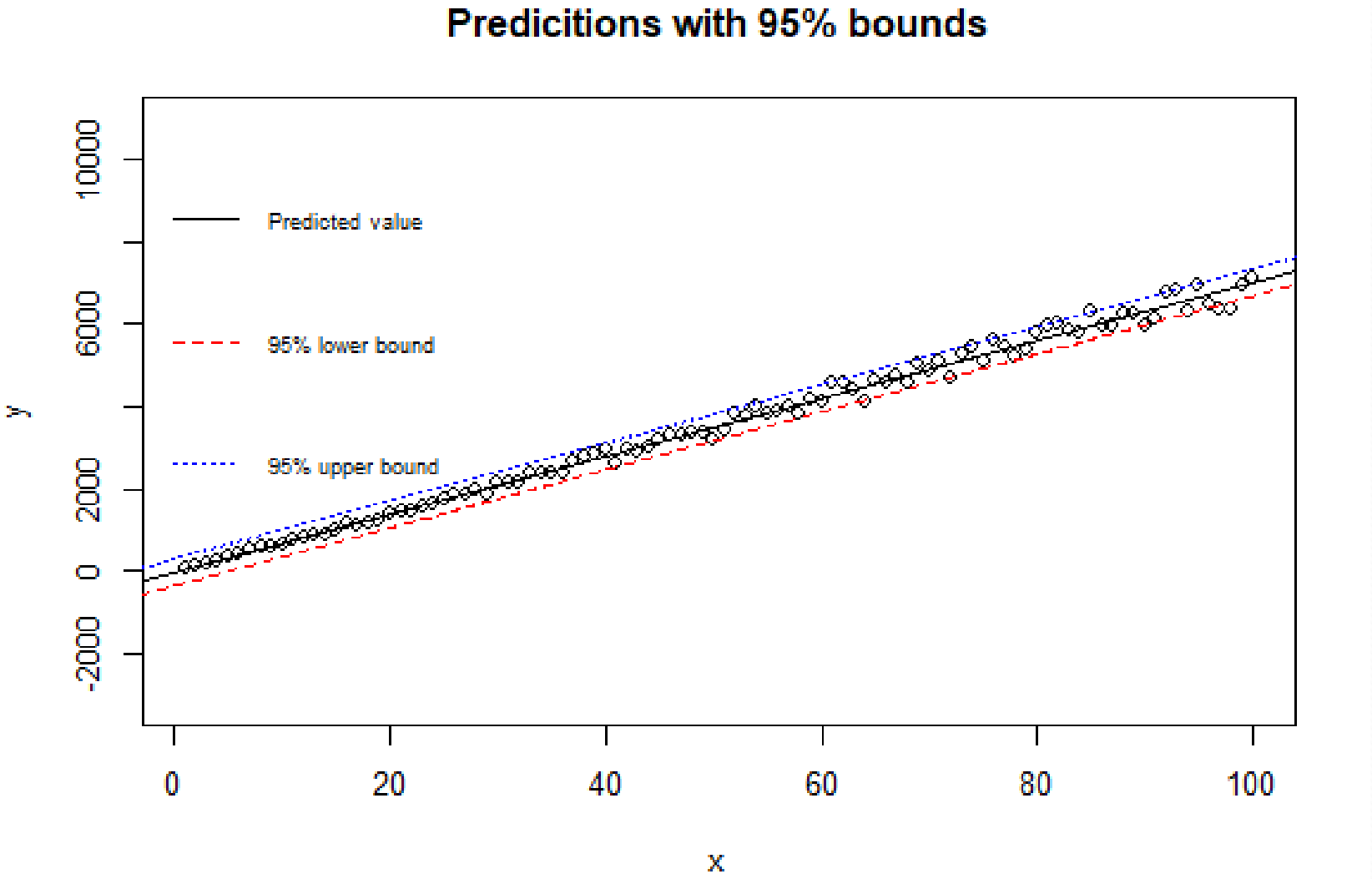
R language, bayesian model, library(Bolstad)
https://www.rdocumentation.org/packages/Bolstad/versions/0.2-41/topics/bayes.lin.reg

```
# Linear regression
# We choose n=100
obs100_df <- build_observations(100, TRUE)
ggplot(data = obs100_df, aes(x = d, y = v)) + geom_point()

# Explain v as a function of d
linear_model <- lm(obs100_df$v ~ obs100_df$d)
summary(linear_model)

# Bayes model
# https://www.rdocumentation.org/packages/Bolstad/versions/0.2-41/topics/bayes.lin.reg
bayes_model <- bayes.lin.reg(obs100_df$v,
                             obs100_df$d,
                             slope.prior = "normal", # use a ``flat'' prior or a ``normal'' prior for Beta.
                             intcpt.prior  = "flat", #  use a ``flat'' prior or a ``normal'' prior. for alpha
                             mb0 = 70.5,  # the prior mean of the simple linear regression slope variable. This
                             sb0 = sqrt(0.5), # the prior std. deviation of the simple linear regression slope
                             ma0 = 0,  # the prior mean of the simple linear regression intercept variable alpl
                             sa0 = 0,  # the prior std. deviation of the simple linear regression variable alpl
                             sigma = NULL,  # the value of the std. deviation of the residuals. By default, thi
                             alpha = 0.05,  # controls the width of the credible interval.
                             plot.data = TRUE, # if true the data are plotted, and the posterior regression lir
                             pred.x = NULL # a vector of x values for which the predicted y values are obtainec
                             )
```

Posterior computation with R

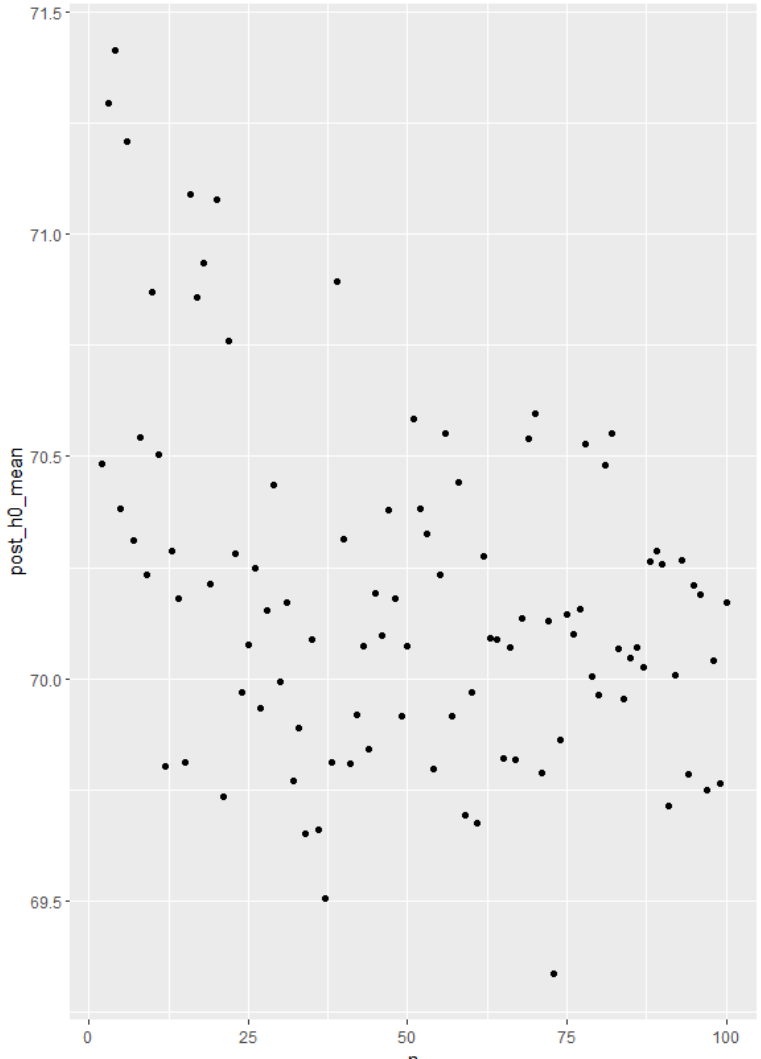


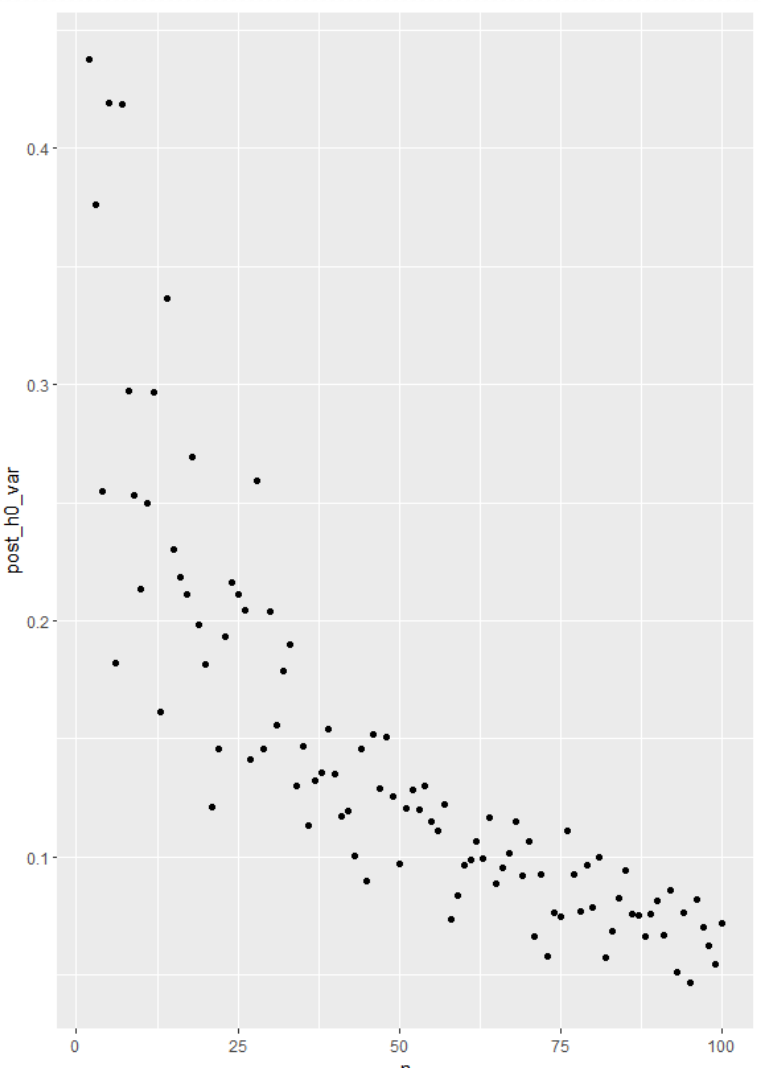
Posterior computation with R



Posterior mean and variance

Posterior mean and variance of Ho as function of number of observations





Bayesian approach

- During the 19th and 20th centuries, frequentist methods largely supplanted Bayesian methods.
- Since the early 1980s, there has been a major resurgence in research and applications of Bayesian methods.
- One might ask why it took so long for Bayesian statistics to return to the forefront ?
- The reason is simple: Bayesian statistics is often computationally complex or unfeasible when applied to simple examples, so we had to wait until numerical resolution methods were sufficiently powerful to enable us to obtain numerical approximations in reasonable times.

Pro/cons

https://documentation.sas.com

Pro

- It provides a natural and principled way of combining prior information with data, within a solid decision theoretical framework. You can incorporate past information about a parameter and form a prior distribution for future analysis. When new observations become available, the previous posterior distribution can be used as a prior. All inferences logically follow from Bayes' theorem.

- It provides interpretable answers, such as "the true parameter has a probability of 0.95 of falling in a 95% credible interval."

Pro/cons

Cons

- It does not tell you how to select a prior. There is no correct way to choose a prior. Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior. If you do not proceed with caution, you can generate misleading results.

- It can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might sometimes be difficult to convince subject matter experts who do not agree with the validity of the chosen prior.

- It often comes with a high computational cost, especially in models with a large number of parameters. In addition, simulations provide slightly different answers unless the same random seed is used. Note that slight variations in simulation results do not contradict the early claim that Bayesian inferences are exact. The posterior distribution of a parameter is exact, given the likelihood function and the priors, while simulation-based estimates of posterior quantities can vary due to the random number generator used in the procedures.

Summary

- Bayes' theorem
- Linear regression problem
- Statistical inference
- Frequentist approach
- Bayesian approach
- **Bayesian Neural Networks**

Bayesian Neural Networks

Based on https://www.datagenius.fr/post/bayesian-deep-learning-soyez-sur-de-vos-incertitudes

Neural networks are:

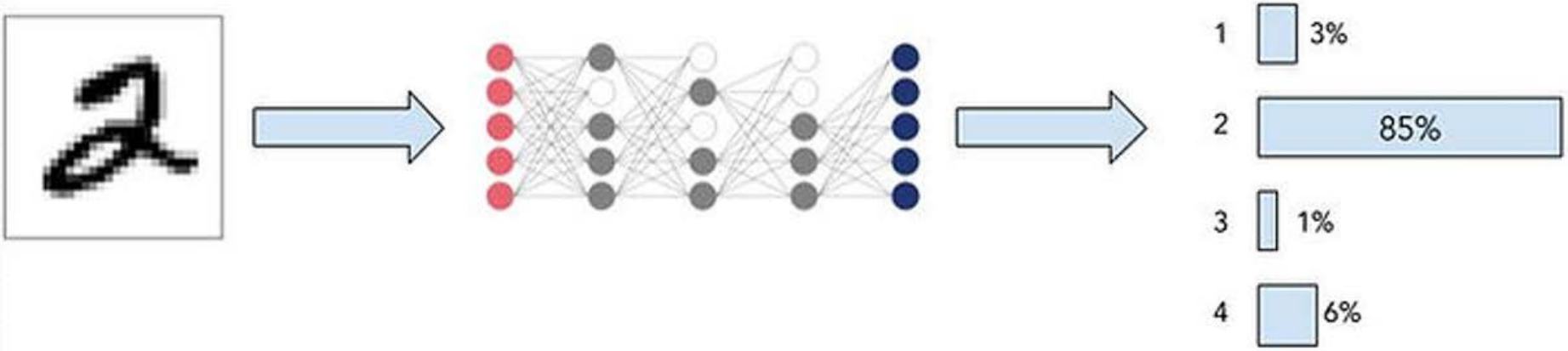
- opaque algorithms, giving little insight into their inner workings;
- predictions are given without any information on the uncertainties of their results.
  - **Bayesian deep learning provides an answer to this second difficulty, by providing more complete information at network output.**

Bayesian Neural Networks

MNIST handwritten digit database

Consider that we have trained a classical neural network on the first 5 classes (0-4) of the MNIST database.

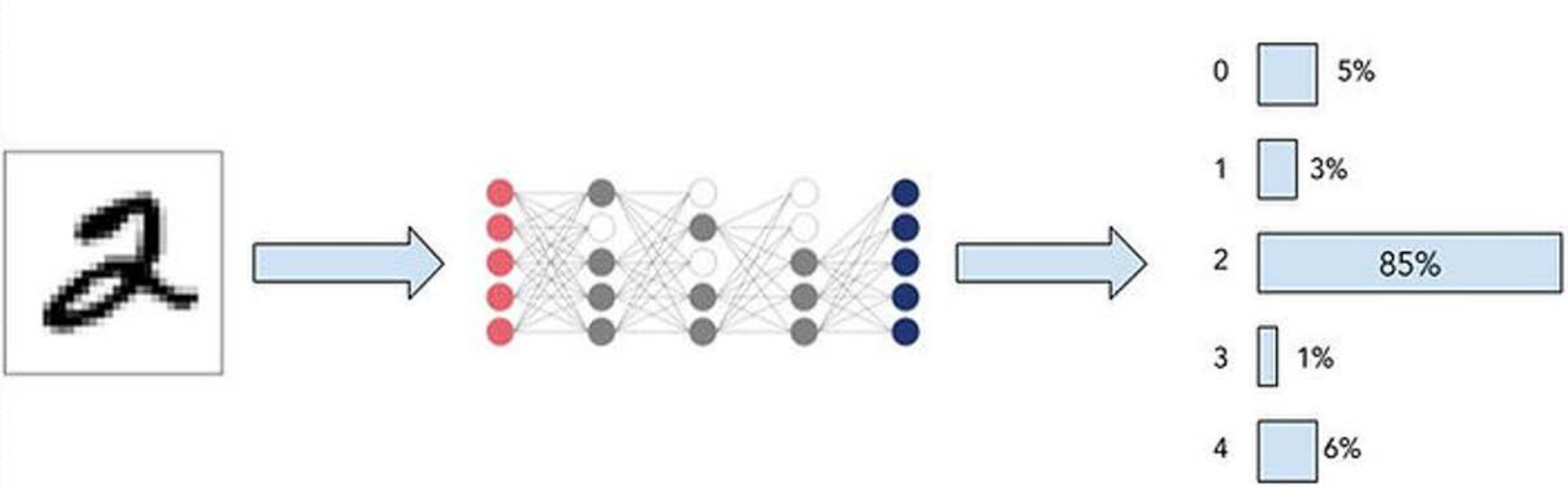
Bayesian Neural Networks



https://www.datagenius.fr/

What does the result obtained by the classical neural network mean?

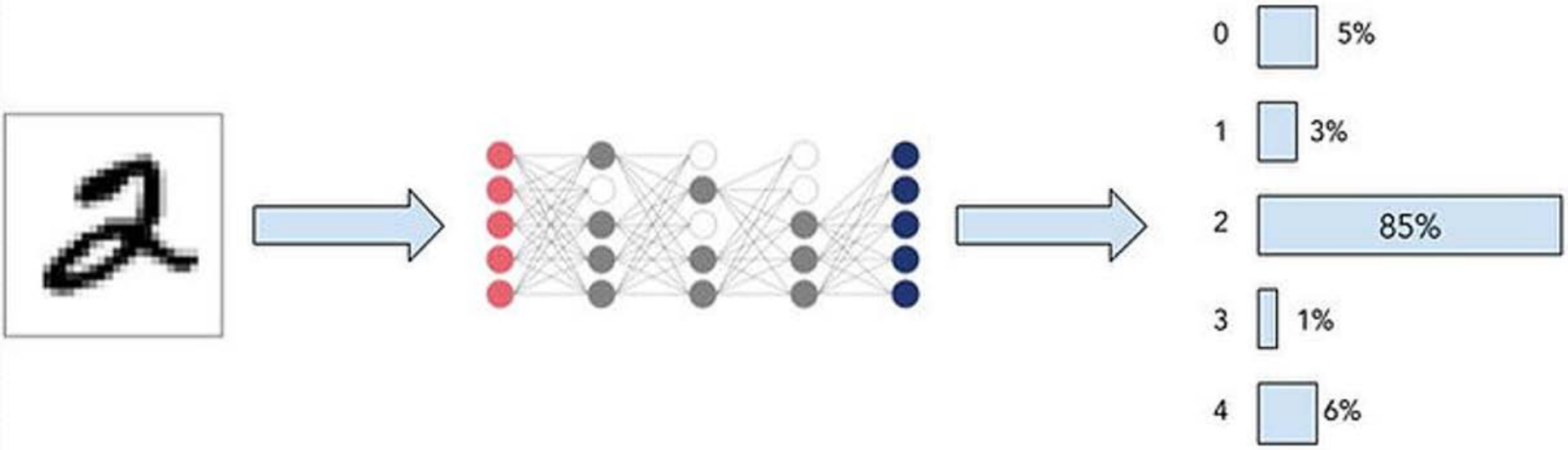
Bayesian Neural Networks



https://www.datagenius.fr/

The image presented is a 2 with 85% certainty

Bayesian Neural Networks



https://www.datagenius.fr/

~~The image presented is a 2 with 85% certainty~~

The image has an 85% chance of being a 2 rather than a 0, 1, 3 or 4

Bayesian Neural Networks

- Classical network gives a measure of the likelihood of the result that **depends** on the classes on which the network has been trained

- We want a measure of the degree of confidence of the network that is **independent** of the training classes.

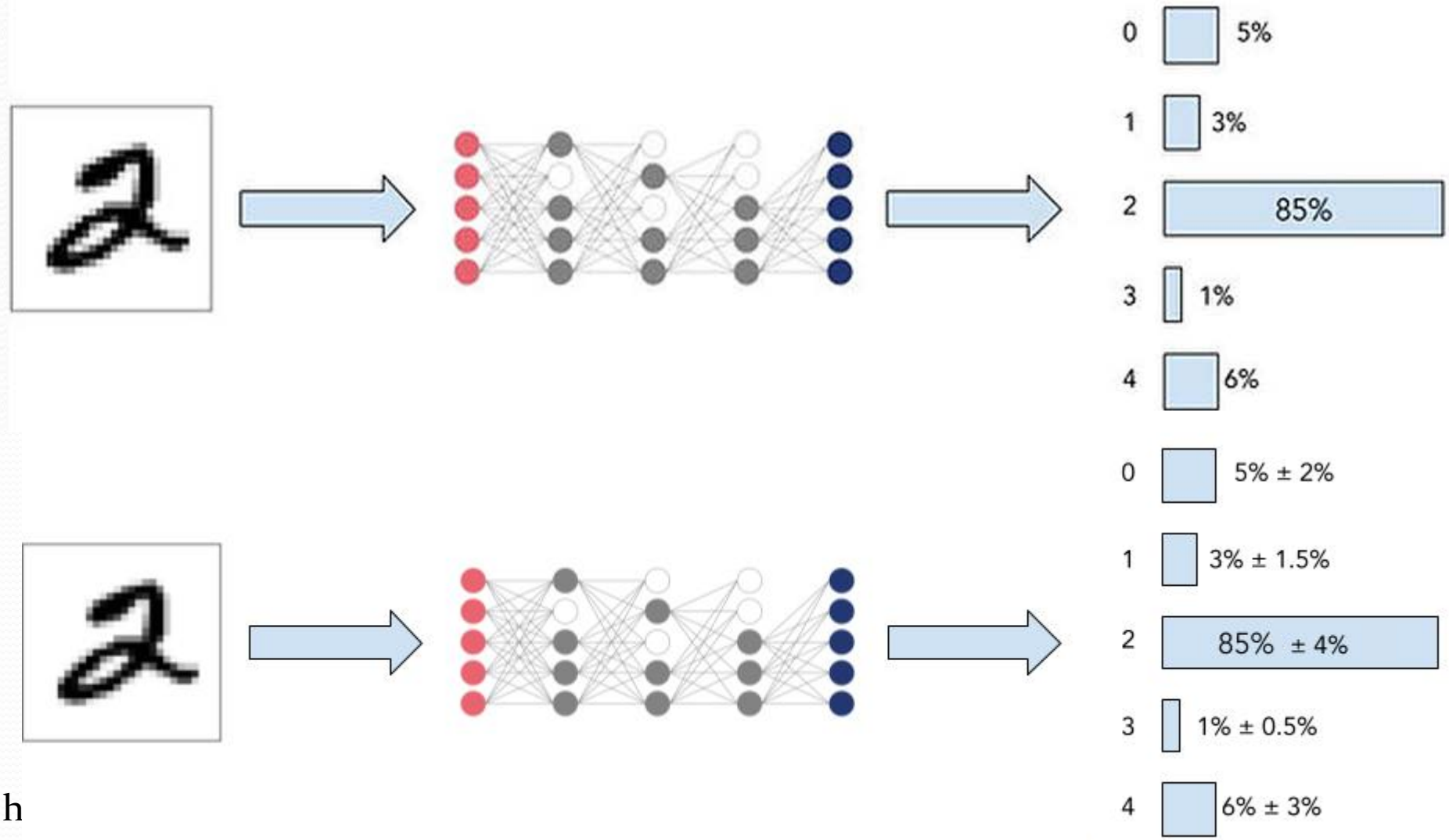
Bayesian Neural Networks

- Uncertainty
  - Random uncertainties:
    - uncertainties associated with **data quality**.
    - e.g. uncertainties due to the level of blur or lack of sharpness of the image.
  - Epistemic uncertainties
    - uncertainties due to **model quality**.
    - It is these uncertainties that provide the most information on the validity of the model's prediction

Bayesian Neural Networks

- Bayesian deep learning introduce the notion of probability during the training and prediction phases of the network.
- The instances manipulated are probability distributions rather than scalars.
- The network weights and the neurons' contained values follow normal distributions characterized by a mean value and a standard deviation
- Finally, each neuron in the network's output layer returns a normal distribution whose mean value and standard deviation have a value determined by the network's input image.

Bayesian Neural Networks



h

Bayesian Neural Networks

CONS

- Compared to classical neural network twice as many values have to be stored and modified (the mean value and the standard deviation).
  - Need more storage space

- the algorithm for updating the weights, which is a modification of the gradient backpropagation algorithm, requires a complex calculation.
  - Need more CPU

Summary

- Bayes' theorem
- Linear regression problem
- Statistical inference
- Frequentist approach
- Bayesian approach
- Bayesian Neural Networks
- **Entertainment**

The Two Children Problem

- *Mr. Smith has two children.*
- *At least one of them is a boy.*

- *What is the probability that both children are boys?*

The Two Children Problem

Older child	Younger child
~~Girl~~	~~Girl~~
Girl	Boy
Boy	Girl
Boy	Boy

$$\mathrm{P}(\mathrm{BB} \mid \mathrm{B}) = \mathrm{P}(\mathrm{B} \mid \mathrm{BB}) \times \frac{\mathrm{P}(\mathrm{BB})}{\mathrm{P}(\mathrm{B})} = 1 \times \frac{\left(\frac{1}{4}\right)}{\left(\frac{3}{4}\right)} = \frac{1}{3}.$$

Pitfalls of relying solely on a fitted model (Anscombe's quartet)