Introduction to Bayesian Statistics

Direct and Inverse problems, Bayes Theorem, Classical Statistics, Bayesian Statistics (parte 2)

Teacher: Matilde Trevisani

DEAMS

A.A. 2024/2025

(aggiornato: 2025-03-13)

Agenda (about 3 lectures)

Introduction to Bayesian Statistics

- Context
- Machinery of Probability
- Interpretations of Probability
- Direct and Inverse problems
- Bayes Theorem
- Modern (classical) Statistics
- Bayesian Statistics
- Prior Distributions
- Modern Approaches
- Final Notes

Direct and Inverse Problems

From Direct Problems to $\longrightarrow$

The calculation of probability, from its origins to the formulation of Bayes' Theorem, was developed to solve **direct** problems: I know the random mechanism that generates the observations and can calculate the probability of the various outcomes.

In a pack of jelly bears, there are R red ones and G green ones. If we draw m (with replacement), what is the probability that x of them are red?

If $\theta = \frac{R}{R+G}$,

$$\Pr(X = x) = \binom{m}{x} \theta^x \, (1-\theta)^{m-x}$$

Suppose we do not know R. We can still calculate the probability above by assuming a probability distribution for (the random variable) R and applying the theorem of total probability (see next slide).

For example, suppose R is determined by rolling an n-sided die (where $n = R + G$).

$$\Pr(X = x) = \sum_{i=1}^{n} \Pr(R = i \cap X = x) = \sum_{i=1}^{n} \Pr(R = i) \Pr(X = x | R = i)$$

where $x = 1, \ldots, m$ and $\theta = \frac{R}{n}$.

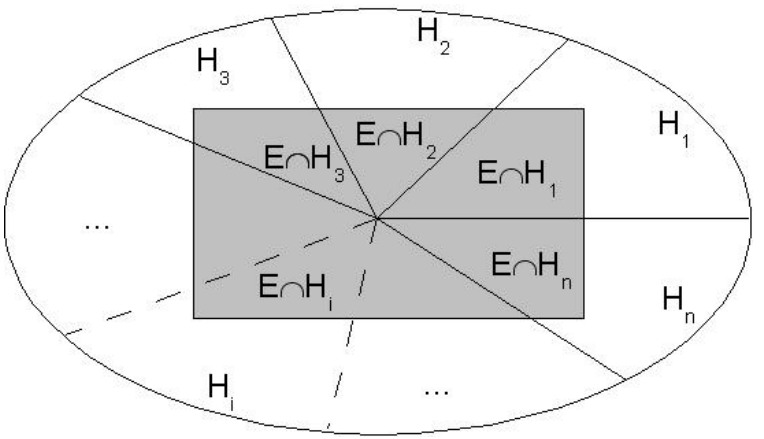
Law of total probability

Let $\{H_i | i = 1, \ldots, n\}$ be a partition of Ω, i.e.,

1. $\bigcup_{i=1}^{n} H_i = \Omega$ (exhaustive),
2. $H_i \cap H_j = \phi$ if $i \neq j$ (pairwise incompatible),

then

$$P(E) = P(E \cap \Omega) = \sum_{i=1}^{n} P(H_i \cap E) = \sum_{i=1}^{n} P(H_i) P(E|H_i)$$



Law of total probability in tabular form

E.g., if $n = 10, m = 5$, then $x = 0, 1, \ldots, 5, \theta = 0.1, 0.2, \ldots, 0.9, 1$

Conditional probabilities:

$$\Pr(X = x|R = i) \text{ or } \Pr(X = x|R = 10\theta)$$

		Urn composition (θ, proportion of red marbles)									
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Sample (x)	0	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	0
	1	.3280	.4096	.3601	.2592	.1562	.0768	.0284	.0064	.0004	0
	2	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	0
	3	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	0
	4	.0005	.0064	.0284	.0768	.1562	.2592	.3601	.4096	.3280	0
	5	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	1

Joint probabilities:

$$\Pr(R = i \cap X = x) = \Pr(R = i)\Pr(X = x|R = i))$$

x	Urn composition (θ, proportion of red marbles)										$P(X = x)$
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
0	.05905	.03277	.01681	.00778	.00313	.00102	.00024	.00003	.00000	0	.12083
1	.03280	.04096	.03601	.02592	.01562	.00768	.00283	.00064	.00004	0	.16252
2	.00729	.02048	.03087	.03456	.03125	.02304	.01323	.00512	.00081	0	.16665
3	.00081	.00512	.01323	.02304	.03125	.03456	.03087	.02048	.00729	0	.16665
4	.00005	.00064	.00284	.00768	.01562	.02592	.03602	.04096	.03281	0	.16253
5	.00000	.00003	.00024	.00102	.00313	.00778	.01681	.03277	.05905	.1	.22083

which, when summed, give: $\sum_{i=1}^{10} \Pr(R = i \cap X = x) = \Pr(X = x)$

Indirect or Inverse Problems (probability of causes)

Within the above experiment, we can also ask the following question

Having drawn $X = x$ bears, what is the probability that there are R red ones in the package?

This problem is solved by **Bayes' Theorem**.

Thomas Bayes (c. 1702-1761)} was a Presbyterian minister. In *Essay Towards Solving a Problem in the Doctrine of Chances* (1763) he considers the inverse probability problem for which he formalizes a solution. His work was published posthumously by his friend Richard Price (1723-1791).

Teorema di Bayes

Bayes theorem: original formulation

In *Essay Towards Solving a Problem in the Doctrine of Chances* (1763) we find

Theorem (PROP. 3)

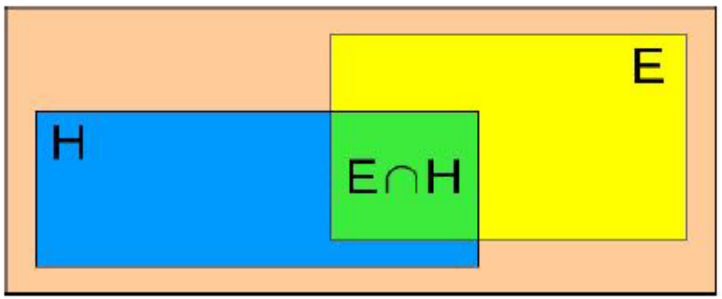
> The probability that two subsequent events will both happen is a ratio compounded of the probability of the 1st, and the probability of the 2d on supposition the 1st happens.

Corollary (PROP. 3)

> Hence if of two subsequent events the probability of the 1st be a/N , and the probability of both together be P/N , then the probability of the 2d on supposition the 1st happens is P/a.

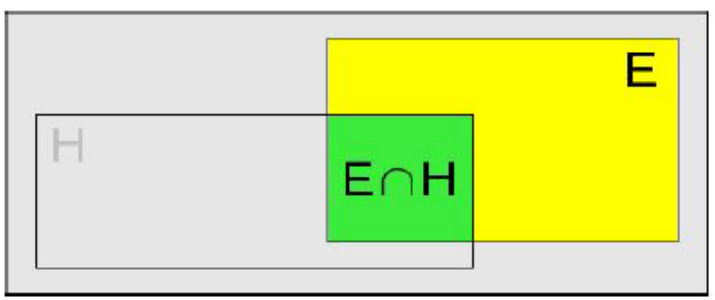
Bayes' Theorem

Let E and H be two events, with $\mathrm{P}(E) \neq 0$,



then

$$\mathrm{P}(H|E) = \frac{\mathrm{P}(H \cap E)}{\mathrm{P}(E)} = \frac{\mathrm{P}(H)\mathrm{P}(E|H)}{\mathrm{P}(E)}$$



Bayes' Theorem for the *jelly bears* example

Given that $X = x$ jelly bears have been extracted, what is the probability that there are R red ones in the package?

The answer from Bayes' theorem is:

$$\mathrm{P}(R = \theta n | X = x) = \frac{\mathrm{P}(R = \theta n \cap X = x)}{\mathrm{P}(X = x)} = \frac{\mathrm{P}(R = \theta n)\mathrm{P}(X = x | R = \theta n)}{\mathrm{P}(X = x)}$$

Assume $X = 2$, consider the joint probabilities

$$\Pr(R = 10\theta \cap X = 2)$$

x	Urn composition (θ, share of red marbles)										$P(X = x)$
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	
0	.05905	.03277	.01681	.00778	.00313	.00102	.00024	.00003	.00000	0	.12083
1	.03280	.04096	.03601	.02592	.01562	.00768	.00283	.00064	.00004	0	.16252
2	.00729	.02048	.03087	.03456	.03125	.02304	.01323	.00512	.00081	0	.16665
3	.00081	.00512	.01323	.02304	.03125	.03456	.03087	.02048	.00729	0	.16665
4	.00005	.00064	.00284	.00768	.01562	.02592	.03602	.04096	.03281	0	.16253
5	.00000	.00003	.00024	.00102	.00313	.00778	.01681	.03277	.05905	.1	.22083

then

$$\Pr(R = 10\theta \mid X = 2) = \frac{\Pr(R = 10\theta \cap X = 2)}{\Pr(X = 2)}$$

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$P(R = 10\theta \mid X = 2)$	.0437	.1229	.1852	.2074	.1875	.1383	.0794	.0307	.0049	0

A statistical problem in the search for *causes*

What we have obtained, the probability of each urn composition, is uncontroversial and straightforward.

Let us make this problem more interesting.

In a pack of n jelly bears, there are R red ones ($R \geq 1$). We draw m of them (with replacement) and observe x red ones. What can we say about R?

In the previous presentation of the problem, R was assumed to be generated by a random mechanism, which made the solution standard (non-controversial).

Now, R ($R = R(\theta)$) is simply *unknown*.

Now the problem is presented as a **general statistical problem**, where it is known that the observation is generated by a random mechanism whose **characteristics** are not fully known.

- How do we interpret the probability of a probability θ (of the causes), $P(\theta|X = x)$?
- Can it represent our beliefs about θ?
- For some, Yes; for others, No.

Two Types of Uncertainty

- **Aleatory**, where uncertainty is due to randomness
  - We are unable to obtain observations that could reduce this uncertainty
- **Epistemic**, where uncertainty is due to a lack of knowledge
  - We are able to obtain observations that can reduce this uncertainty
  - Two observers may have different epistemic uncertainty

Updating Uncertainty (Probability)

Probability of red $\frac{\mathrm{R}}{\mathrm{R+G}} = \theta$

- $p(y = \text{red} \,|\, \theta) = \theta$ **aleatory** uncertainty
- $p(\theta)$ **epistemic** uncertainty

Repeated draws of jelly bears update our uncertainty about the proportion

- $p(\theta \,|\, \mathrm{y} = \text{red}, \text{yellow}, \text{red}, \text{red}, \ldots) = ?$
- According to ***Bayes' rule***

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

(*Anticipation*) Model *vs* Likelihood, a priori *vs* a posteriori

- Bayes' rule
  $p(\theta|y) \propto p(y|\theta)p(\theta)$
- Model: $p(y|\theta)$ as a function of y, given a fixed θ, describes the aleatory uncertainty
- Likelihood: $p(y|\theta)$ ($= L(\theta|y)$)
  as a function of θ, given a fixed y, **provides information about the epistemic uncertainty**,
  but it is not a probability distribution
- *Bayes' rule* combines the likelihood with the a priori uncertainty $p(\theta)$ and transforms them into the updated a posteriori uncertainty.

Bayes: inference for a probability

This is stated and more or less solved in Bayes essay as follows.

> *Given* the number of times on which an unknown event has happened and failed:
>
> *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

If nothing is known of an event but that it has happened p times and failed q in $p+q$ or n trials, and from hence I judge that the probability of it's happening in a single trial lies between $\frac{p}{n}+z$ and $\frac{p}{n}-z$ my chance to be right is *greater* than $\frac{\sqrt{\mathrm{K}pq}\times h}{2\sqrt{\mathrm{K}pq}+hn^{\frac{1}{2}}+hn^{-\frac{1}{2}}} \times 2\mathrm{H}-\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n+1}{n+2}\times\frac{1}{mz}\times\sqrt{1-\frac{2m^2z^2}{n}|^{\frac{n}{2}+1}}$ and less than $\frac{\sqrt{\mathrm{K}pq}\times h}{2\sqrt{\mathrm{K}pq}-hn^{\frac{1}{2}}-hn^{-\frac{1}{2}}}$ multiplied by the 3 terms $2\mathrm{H}-\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n+1}{n+2}\times\frac{1}{mz}\times\sqrt{1-\frac{2m^2z^2}{n}|^{\frac{n}{2}+1}}+\frac{\sqrt{2}}{\sqrt{\mathrm{K}}}\times\frac{n}{n+2}\times\frac{n+1}{n+4}\times\frac{1}{m^3z^3}\times\sqrt{1-\frac{2m^2z^2}{n}|^{\frac{n}{2}+2}}$ where m^2, K, h and H stand for the quantities already explained.

Bayes solution was not actually very clear, the one from Laplace was better.

Laplace and the (epistemic) probability of a female birth

Laplace (1749-1827) was the first to formulate a statistical problem and solve it using Bayesian statistics.

The question he posed was whether the probability of a female birth (θ) was less than 0.5.

The problem is entirely analogous to the one of drawing from the *urn* (pack) above, except that "there exists a continuum of possible compositions of the urn".

He observed that in Paris, from 1745 to 1770,
there had been 493,472 births, of which 241,945 were females, from which he derived that

$$P(\theta \geq 0.5 \,|\, \text{data}) \approx 1.15 \times 10^{-42}$$

leading to the "moral certainty" that $\theta < 0.5$.

Pierre-Simon Laplace (1749-1827) in *Essai philosophique sur les probabilites (1814)* gives a systematic treatment to the approach which we call Bayesian today.

Laplace and the Extension of Bayes' Theorem

Laplace extended Bayes' theorem to n possible causes, H_i, $i = 1, \ldots, n$, of an event E.

Given:

- $H_1, \ldots, H_n$, a set of hypotheses
- $\mathrm{P}(H_i)$, $i = 1, \ldots, n$, prior probabilities, $\sum_i \mathrm{P}(H_i) = 1$
- $\mathrm{P}(E|H_i)$, $i = 1, \ldots, n$, likelihood of E when H_i is true

$$\mathrm{P}(H_i|E) = \frac{\mathrm{P}(E|H_i)\mathrm{P}(H_i)}{\sum_{i=1}^{n} \mathrm{P}(E|H_i)\mathrm{P}(H_i)}$$

The posterior probability of H_i given E is proportional to the product of the prior probability of H_i and the likelihood of E when H_i is true.

Corollary: Comparison Between Two Hypotheses

Suppose two particular hypotheses, H_i and H_j, are compared; the posterior ratio is given by:

$$\frac{\mathrm{P}(H_i|E)}{\mathrm{P}(H_j|E)} = \frac{\mathrm{P}(H_i)}{\mathrm{P}(H_j)} \times \frac{\mathrm{P}(E|H_i)}{\mathrm{P}(E|H_j)}$$

i.e., the ratio of the prior probabilities (*prior odds*) to the ratio of the likelihoods (*likelihood ratio*).

Example: Medical Diagnosis

Objective: Evaluate if a person with a positive test result is ill.

- A patient can either be affected by a specific disease - being in state H_1 - or not - being in state H_2
- $P(H_1)$ is the prevalence of the disease in the population to which the patient is assumed to belong - the **prior** probability of H_1; $P(H_2) = 1 - P(H_1)$
- The patient undergoes a test that provides **information** on whether he is likely ill: positive test result, E, or negative test result, $\bar{E}$
- $P(E|H_1)$, $P(\bar{E}|H_2)$ are the *true positive* and *true negative* rates, also called sensitivity and specificity of the clinical test.

> The optimal situation is
>
> $$\begin{aligned} P(E|H_1) = 1 \quad &\rightarrow \quad P(\bar{E}|H_1) = 0 \quad \text{false negative} \\ P(\bar{E}|H_2) = 1 \quad &\rightarrow \quad P(E|H_2) = 0 \quad \text{false positive} \end{aligned}$$

- Bayes' theorem helps us understand, by combining the test characteristics with the disease prevalence, the **discriminative power** of the diagnostic test.

Covid-19 and Nasopharyngeal Swabs

Recent studies suggest that nasopharyngeal swabs used for diagnosing COVID-19 have a sensitivity $P(E|H_1)$ = 0.777 and specificity $P(\bar{E}|H_2)$= 0.988, and that the prevalence of COVID-19 in Italy (during the pandemic phase) is about $P(H_1)$ = 0.13.
We apply Bayes' theorem to calculate the probability of being ill after a positive diagnostic test result:

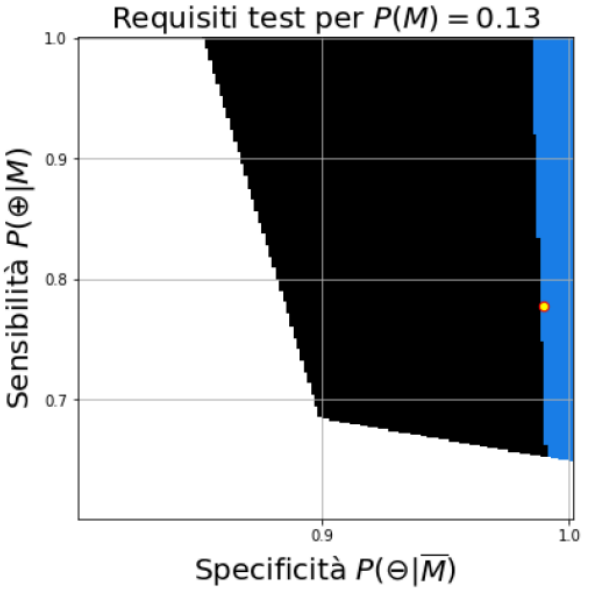
$$\begin{aligned} P(H_1|E) =& \frac{\mathrm{P}(E|H_1)\mathrm{P}(H_1)}{\mathrm{P}(E|H_1)\mathrm{P}(H_1) + \mathrm{P}(E|H_2)\mathrm{P}(H_2)} = \\ =& \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (\text{1-specificity}) \cdot (\text{1-prevalence})} = \\ =& \frac{0.777 \cdot 0.13}{0.777 \cdot 0.13 + (1 - 0.988) \cdot (1 - 0.13)} = 0.9063257 \end{aligned}$$

The probability of not being ill after a negative result:

$$\begin{aligned} P(H_2|\bar{E}) =& \frac{\text{specificity} \cdot \text{1-prevalence}}{\text{specificity} \cdot \text{1-prevalence} + (\text{1-sensitivity}) \cdot (\text{prevalence})} = \\ =& \frac{0.988 \cdot (1 - 0.13)}{0.988 \cdot (1 - 0.13) + (1 - 0.777) \cdot 0.13} = 0.9673738 \end{aligned}$$

Ideal Test

Given the prevalence of a disease, M, the ideal characteristics of a test for the most accurate possible diagnosis are shown in the graph.



Test requirements for COVID-19

The **black** area indicates the **necessary requirements**

$$P(M|T) > .5 \quad \text{e} \quad P(M|\bar{T}) < .05$$

for a test with $P(M) = .13$.

The **blue** area indicates the **optimal requirements**: $P(M|T) > .9$.

Il **yellow** point indicate sthe sensitivity and specificity parameters of the RT-PCR SARS-CoV-2 RNA test for COVID-19, respectively: $P(T|M) = 0.777$ and $P(\bar{T}|M) = 0.988$.

A statistical problem

Returning to the urn example, the crucial point is that $R(\theta)$ is not random in the sense of being generated through a random experiment (aleatory uncertainty). Rather, $R(\theta)$ is unknown (epistemic uncertainty).

How should we then interpret the probability we assign to θ: $P(\theta|x)$?

Can it represent our opinion about the value of θ?

According to some, it could; according to others, this interpretation is meaningless.

Bayesian approach put aside

Since Laplace, and for a relatively long time, the Bayesian approach was put aside because it was deemed unscientific.

- The idea that the probability could be used to model ignorance/beliefs was ridiculed.
- Also, in order to get $P(\theta|X = x)$ we need to start from $P(\theta)$, a *prior* belief about θ (which comes before observations), this amounted to introducing an element of subjectivity in the analysis, which was, again, deemed unscientific.
- Moreover, there were practical problems: even for relatively simple problems, the Bayesian approach easily leads to intractable computations (Laplace used clever approximations to get his inference about θ)

Modern (classical) statistics

New questions, new answers

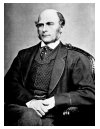
Between XIX and XX-th centuries new fields of application of statistical techniques rise

- quality control
- heredity and genetics

One has a, possibly incomplete, theory, i.e., a $Model$, on how some outcome is generated and wants to use $Data$ to confirm/clarify the said model.

New approaches are developed in which the **parameter θ is a fixed number**.

William Gosset (1876-1937) Working for Guiness, he developed the *Student-t* distribution to evaluate quality of barley.



sir Francis Galton (1822-1911) Founded the Eugenics Record Office in London, later the Galton labooratory. Develops *linear regression*.

Karl Pearson (1857-1936) Introduces the concept of correlation and of goodness of fit.

Likelihood and Repetead Sampling

- Inference is based on the **likelihood**: we compare $P(Data|Model)$ (in urn example $L(\theta) \propto P(X = x|R = n\theta)$) for the different models (In Bayesian statistics we compare $P(Model|Data)$),
- Performance is evaluated according to the principle of **repeated sampling**: procedures are evaluated based on fictitious repetitions of the experiment (*Mean Square Error*, *coverage probability*, *significance level*, etc.).

sir Ronald Fisher (1890-1932) introduces, among other things, the concepts of *likelihood*, *analysis of variance*, *experimental design*. Also, he originates the ideas of *sufficiency*, *ancillarity*, and *information*. His main works: *Statistical Methods for Research Workers* (1925), *The design of experiments* (1935), *Contributions to mathematical statistics (1950)*, *Statistical methods and statistical inference* (1956).

Egon Pearson (1895-1980) with Jerzy Neyman develops the *theory of hypotheses testing*.

Likelihood

The likelihood summarizes information on θ coming from $X = x$

$$L(\theta) \propto P(X = x | R = 10\theta)$$

x	Urn composition (θ, share of red marbles)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
0	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	0
1	.3280	.4096	.3601	.2592	.1562	.0768	.0284	.0064	.0004	0
2	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	0
3	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	0
4	.0005	.0064	.0284	.0768	.1562	.2592	.3601	.4096	.3280	0
5	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	1

Suppose we observe $x = 3$.

From the likelihood alone we get answers in the form of

- maximum likelihood estimator:
  - $\hat{\theta} = X/5 = 0.6$
- p-values:
  - e.g., the p-value for the hypotheses $\theta \leq 0.2$ is 0.0579 (*How do we compute it?*)

Repeated sampling principle

According to the repeated sampling principle, we evaluate our procedures based on how they would behave in the long run with new sets of data.

Using the repeated sampling principle we can evaluate the performance of

- estimators $\rightarrow$ *Mean Square Error*
- confidence intervals $\rightarrow$ *coverage probability*

Neyman-Pearson hypotheses testing has the most evident link with repeated sampling:

- *significance level* (α) is the relative frequencies with which we expect to reject a null hypotheses if we were to perform the test on a number of samples coming from a population for which the null is true;
- *power* is ...

Classical inference for female births

As far as the probability θ of a female birth is concerned, Laplace observations that in Paris, from 1745 to 1770 there were $493, 472$ births, of which $241, 945$ were girls would lead to

- ML estimate $\hat{\theta} = \frac{241,945}{493,472} = 0.4903$
- a 95% confidence interval $[0.4889, 0.4917]$
- *p-value* for the hypotheses $H_0 : \theta \geq 0.5$ è ≈ 0
- The best guess for θ is 0.4903
- we obtained an interval $[0.4889, 0.4917]$ as a realization of a random interval which has probability 95% of covering the true value of θ
- if $H_0 : \theta \geq 0.5$ were true, the probability of observing a sample as extreme as the one we saw would be ≈ 0.

What these tell us about θ is not obvious, where by this I mean that we need to make a further step to translate it in information on θ. (see slides on frequentist inference)

Classical approach or approaches?

Note that within the classical approach different views can be distinguished, this is particularly evident in hypotheses testing.

A Fisherian approach is to view the likelihood as central as a measure of evidence brought by the data. As such, a p-value is a measure of evidence against a given hypotheses.

The Neyman-Pearson view is behavioural, they devise a decision rule which controls the probability of error (not the overall one, but at least the conditional ones).

The above is a very simplistic summary, however it is true that the two approaches are incompatible and there have been harsh debates between the proponents.

Interpretation of results in Bayesian Statististics and Classical Statistics

> A primary motivation for Bayesian thinking is that it facilitates a ***common sense interpretation*** of statistical conclusions.
>
> -- *Gelman*

Contrast interval estimation or hypotheses testing, BS tells us what we want to know, classical statistics does not, and it is likely that many users'' would incorrectly interpret classical statistics results the Bayesian way (luckily in many cases this is ok).

> Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a ***probability distribution on the parameters of the model and on unobserved quantities*** such as predictions for new observations.
>
> -- *Gelman*

Classical approach vs Bayesian approach to statistical inference

In CLASSICAL INFERENCE

- the parameter is a constant.
- the conclusion is not derived within probability calculus rules (these are used in fact, but the conclusion is not a direct consequence)
- the likelihood and the probability distribution of the sample are used

Framework to extract evidence from data.

In BAYESIAN INFERENCE

- the parameter is a r.v.
- the reasoning and the conclusion is an immediate consequence of probability calculus rules (of Bayes' theorem in particular);
- the likelihood and the prior distribution are used

Framework to update information.

Bayesian Statistics

> Il teorema di Bayes costituisce la chiave di volta e il concetto informatore di ogni attività costruttiva del pensiero.
>
> -- *Bruno de Finetti, 1970*

Bayes' theorem is the tool par excellence capable of updating the probabilities of the hypotheses H_j in light of new facts.

The following example highlights the error that is made when we ignore available information, in our case the prior probabilities, and draw inferences (and conclusions) using only the likelihoods. (See slides Introduction to Bayesian inderence)