UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di
Scienze Economiche, Aziendali,
Matematiche e Statistiche "Bruno de Finetti"

Statistica

A quick introduction to Bayesian statistics

Leonardo Egidi

legidi@units.it

Dipt.to di Scienze Economiche, Aziendali, Matematiche e Statistiche, Univ. di Trieste

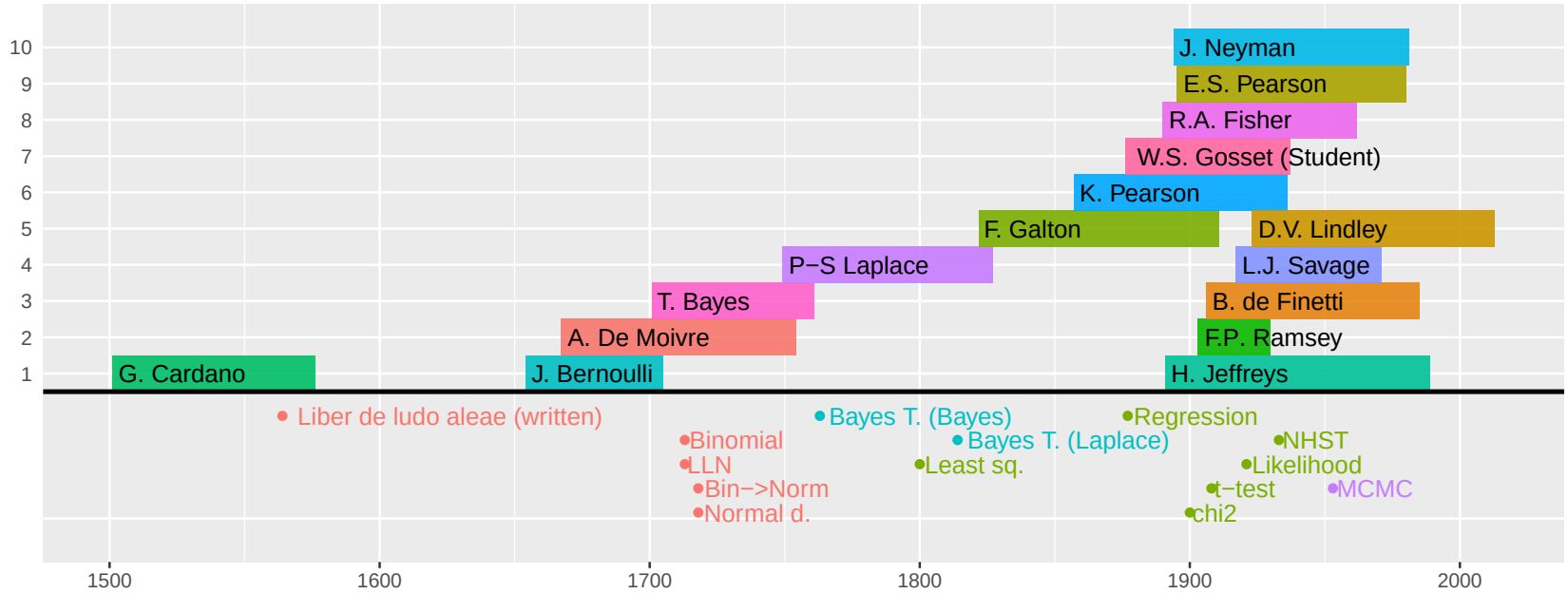
A.A. 2025/26

Indice

Why starting from history

We will start with an overview of the history of Bayesian statistics: its development from probability calculus and its relationship with the so called classical statistics.

This is useful to better understand the differences between Bayesian and classical statistics.



Games of chance are the cradle of probability

Probability calculus is initially developed to study games of chance: developing strategies to win in games was of interest to nobles, who were willing to pay scholars for them.

For example Galileo in 1620 wrote a note offering the solution of this issue:

> *suppose three dices are thrown and the three numbers obtained added. What is the probability that the total equals 9?*

This circumstances help not only because of the money, but also because of the simple structure of the problems involved.

Elementary probability

The first examples of probability problems are concerned with simple random mechanisms whose symmetry offered the solution.

Q: A marble is randomly drawn from an urn containing R red marbles and W white marbles, what is the probability that the marble is red?

A: By symmetry

$$P(red) = \frac{R}{R+W}$$

Games of chances are easily tackled using the first definition of probability, based on symmetry

$$\text{prob. of event} = \frac{\#\text{ favourable outcomes}}{\#\text{ possible outcomes}}$$

Elementary probability and combinatorics

For more "complicated" questions tools were developed to count favourable and unfavourable outcomes.

Q: We draw a marble from an urn containing R red marbles and W white marbles m times (putting it back in the urn after each draw), what is the probability that r out of m are red?

A: Still by symmetry

$$P(r \text{ red out of } m) = \binom{m}{r} \left(\frac{R}{R+W}\right)^r \left(1 - \frac{R}{R+W}\right)^{m-r}$$

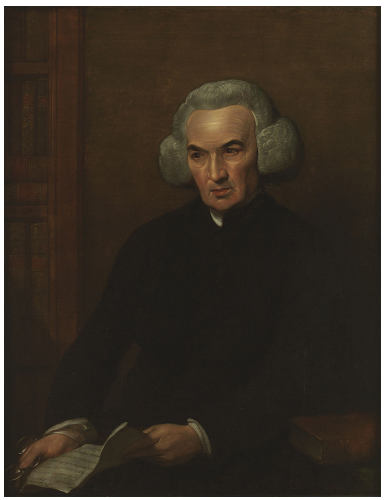
Girolamo Cardano (1501-1576), wrote the first systematic treatment of probability in 1576: *Liber de ludo aleae*; this, however was not published until 1663. He was a polymath with interests ranging from mathematics to biology. He was also a gambler (and a rumor exists that he did not publish his book on probability because his knowledge gave him an advantage in betting).

Indirect problems: the probability of causes

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

> *Having observed $X = x$, what is the probability that the urn contains R red marbles?*

This is solved by **Bayes theorem**.



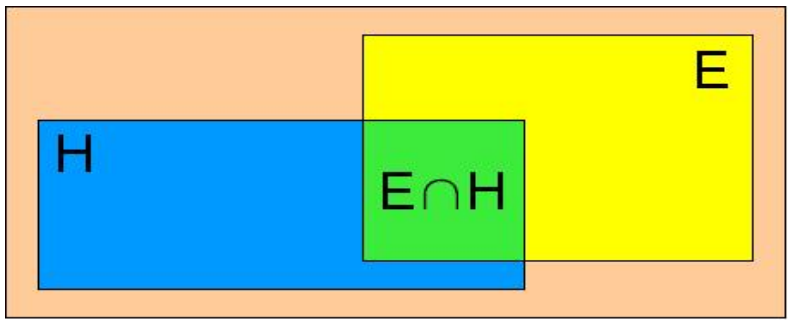
Thomas Bayes (c. 1702-1761) was a Presbyterian minister. In *Essay Towards Solving a Problem in the Doctrine of Chances (1763)* he considers the inverse probability problem for which he formalizes a solution.
His work was published posthumously by his friend **Richard Price (1723-1791)**.

Bayes theorem

Theorem (Bayes theorem)

Let E and H be two events, assume $P(E) \neq 0$, then

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

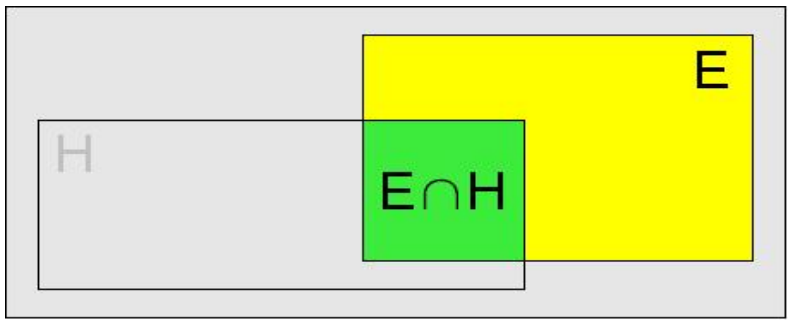


Bayes theorem

Theorem (Bayes theorem)

Let E and H be two events, assume $P(E) \neq 0$, then

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



Probability of a female birth, Laplace

Laplace was the first to formulate a statistical problem and solve it with Bayesian statistics.

The question he poses was whether the probability of a female birth (θ) is or is not lower than 0.5.

The problem is analogous to that of the urn above, but for the fact that "there is a continuum of possible urn compositions".

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

The statistical problem

Back to

> *An urn contains 10 marbles, R of which are red ($R \geq 1$), we draw a marble from the urn 5 times ... we observe X red marbles. What can we say about R?*

the point here is that $R\ (\theta)$ is not random in the sense of being generated through a random experiment (such as the die).

Rather, $R\ (\theta)$ is *unknown* to us.

How are we then to interpret the probability we attach to θ: $P(\theta|X = x)$?

Can it represent our beliefs on the value of θ?

According to some it could, according to other, this was nonsense.

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)

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 percent confidence interval:
$[0.4889, 0.4917]$,

p-value for the hypotheses $H_0 : \theta \geq 0.5$:
≈ 0.

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 : $\hat{\theta} = \frac{241,945}{493,472} = 0.4903$,	The best guess for θ is 0.4903,
a 95 percent confidence interval: $[0.4889, 0.4917]$,	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,
p-value for the hypotheses $H_0 : \theta \geq 0.5$: ≈ 0.	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.

Classical and Bayesian statistical inference, differences

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.

Problems with probability

Although employed in special contexts, Bayesian Statistics did not achieve general acceptance (far from it).

One of the reasons why Bayesian statistics was difficult to accept is related to the frequentist definition of probability.

It is conceptually difficult to frame a distribution on the model (parameter) as a frequentist probability.

We need another probability!

Let us take a step back and discuss about this.

Subjective probability

For 'frequentist-friendly' events ('tail is observed when a coin is thrown')

- everyone (presumably) would agree on the value of the probability;
- the frequentist definition is intuitively applied;
- $\rightarrow$ this is an 'objective' probability.

For more general events such as 'Italy wins the next world cup',

- it is still possible to state a probability;
- everyone would assign a different probability;
- the probability given by someone will change in time.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) proposes the subjective definition of probability and the coherence framework, based on the bet interpretation (see *Theory of probability (1970)*).
In *Theory of probability* he wrote

> *Probability does not exist*

Subjective probability

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Definition (Subjective probability)

The probability of an event is, for an individual, his degree of belief on the event.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) proposes the subjective definition of probability and the coherence framework, based on the bet interpretation (see *Theory of probability (1970)*).
In *Theory of probability* he wrote

> *Probability does not exist*

Subjective probability

One then accepts that the probability is not an objective property of a phenomenon but rather the opinion of a person and one defines

Definition (Subjective probability)

The probability of an event is, for an individual, his degree of belief on the event.

If the probability is a subjective degree of belief, it depends on the information which is subjectively available, and it is also clear that by random we mean not known for lack of information.

Bruno de Finetti (c. 1906-1985), Italian probabilist and actuary (for Generali) proposes the subjective definition of probability and the coherence framework, based on the bet interpretation (see *Theory of probability (1970)*).
In *Theory of probability* he wrote

> *Probability does not exist*

Bayesian statistics and subjective probability

The subjective definition of probability is most compatible with the Bayesian paradigm, stated as follows

- the parameter to be estimated is a well specified quantity but is not known for lack of information
- a probability distribution is (subjectively) specified for the parameter to be estimated, this is called the *prior distribution*
- after seeing experimental results the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (likelihood) and prior distribution to obtain the *posterior distribution*.

Subjective probability and Bayesian update rule (which is actually a consequence of probability rules) establish a system to describe inference whose input are the prior beliefs and the data and the output is updated (posterior) beliefs.

Need for prior

A critical issue in Bayesian inference (and one of the reasons why it did not get acceptance at the beginning) is the need for prior information.

While classical statistics is only concerned with the information coming from the data, Bayesian statistics is a rule to update information based on the data: we must start somewhere.

This was seen as a major issue since it introduces an element of subjectivity in the analysis.

This will be discussed later, we make now two preliminary notes concerning

- where the prior comes from;
- the subjectivity (in the sense of arbitrariness) of results.

Source of prior information

Think of the female birth example again, but with the following sample:

> *In 2010 in Muggia (small city near Triest) 38 males and 47 females were born.*

According to likelihood inference (for θ, pr. of a female birth)

- the ML estimate is $\hat{\theta} = 0.553$
- the 95% c.i. is $[0.441, 0.659]$
- the p-value for $H_0 : \theta \geq 0.5$ is 0.8

What do you think of this information?

Source of prior information

Think of the female birth example again, but with the following sample:

> *In 2010 in Muggia (small city near Triest) 38 males and 47 females were born.*

According to likelihood inference (for θ, pr. of a female birth)

- the ML estimate is $\hat{\theta} = 0.553$
- the 95% c.i. is $[0.441, 0.659]$
- the p-value for $H_0 : \theta \geq 0.5$ is 0.8

What do you think of this information?

You probably think something along the lines of 'This sample tells me nothing'.

Why is that? Well, because you have, in fact, prior information.

We usually have prior information

In fact, it would be rare that we model a situation were we have no prior information at all.

Prior information may come from

- substantive knowledge about the process generating the data (we may be unsure about the exact mechanism but we usually know something),
- observations made in the past.

With this in mind, the prior distribution should not look so strange.

Subjectivity of results: make the prior irrelevant

The idea is that the prior does not need to convey information, rather it is regarded as a technical component of the model.

This idea lies behind the so called

- non informative priors
- reference priors

whose name tells it all, although maybe too optimistically:

- 'informativeness' is not a well defined concept, beware of attaching a precise meaning to the intuitive idea
- the posterior still depends on the prior

With these caveats let's say that particular distributions can be defined to avoid the subjective interpretation of the prior distribution.
This approach is sometimes called objective Bayes (or automatic Bayes).

Role of Bayesian reasoning

These circumstances lead some to think that Bayesian reasoning could (should) be used as the paradigm of inductive logic, that is, beyond its statistical scope: a recipe for human reasoning in general.

Let us then consider contexts were beliefs are important (central) and discuss to what extent Bayesian reasoning fits practice:

- diagnostic: where interest lies on whether a tested person is ill
- law: where interest lies in the belief on guilt or innocence of a defendant.
- science (epistemology): where interest lies in the truth of a theory,

The question is whether (to what extent, under which conditions) Bayesian reasoning can describe (model) the reasoning process of a scientist (judge/juror, clinician) who accept/rejects theories (decides over guilt/innocence, diagnose patients).

Diagnostic: Hanahaki disease

Hanahaki diseases has prevalence of 1% in a population.

A test is available such that the probability that it comes out positive is

- 90% if the tested individual is affected by Hanahaki: $P(T|H) = 0.9$
- 10% if the tested individual is not affected by Hanahaki: $P(T|\bar{H}) = 0.1$

John Smith is tested and is positive, should he worry seriously?

How likely is he to be affected?

$$P(H|T) = 0.08.$$

Indice

Regression foundations

The problem we will consider in this section is that of using the values of one variable to explain or predict values of another. We shall refer to an *explanatory* (covariate) and a *dependent* (outcome, response) variable.

A common question is: how does one quantity, y, vary as a function of another quantity (or vector of quantities), x?

In general, we are interested in the conditional distribution of y, given x, parametrized as $p(y|\theta, x)$, where n observations (x_i, y_i) are available. The term *regression* dates back to Francis Galton, namely *regression toward the mean*.

Typical examples arising in epidemiology/biostatistics:

- causal explanations for some treatments/drugs (causal inference)
- predict and estimate levels of specific antigens (observational studies)
- model the probability to die/survive (randomized clinical trials)
- ...

Regression foundations

The distribution of y given x is typically studied in the context of a set of units or experimental subjects. Regression models the statistical relationship between the variates and the covariates. There are two slightly different situations:

- in the first one the experimenters are free to set the values of x_i: in such a case, x is *deterministic*. We will typically assume that the covariates are independent of the model parameters: $p(x|\theta) = p(x)$.
- in the second one both values are random, although one is thought of as having a causal or explanatory relationship with the other.

The analysis, however, turns out to be the same in both cases.

Regression foundations

The two approaches differ in the evaluation of the joint distribution of y and x:

- in the first approach the joint distribution of x and y is proportional to the conditional distribution of y given x:

$$p(y, x|\theta) = p(y|x, \theta)p(x)$$

$$p(y, x|\theta) \propto p(y|x, \theta)$$

Ex: studying the height(y) vs the weight (x) of n individuals

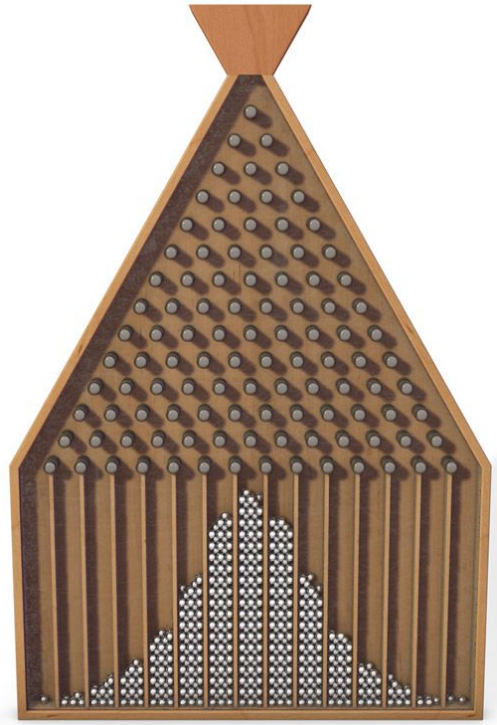
- in the second approach there is a model for the x as well:

$$p(y, x|\theta) = p(y|x, \theta)p(x|\theta)$$

Ex: studying the effect of a noisy protein measurements x on some outcome response.

Regression foundations

Galton developed the following model: pellets fall through a quincunx to form a normal distribution centered directly under their entrance point. These pellets might then be released down into a second gallery corresponding to a second measurement. Galton then asked the reverse question: "From where did these pellets come?"



The answer was not 'on average directly above'. Rather it was 'on average, more towards the middle', for the simple reason that there were more pellets above it towards the middle that could wander left than there were in the left extreme that could wander to the right, inwards.

Regression foundations

Goals of a regression analysis

1. understanding the behavior of y, given x
2. predicting y, given x, for future observations
3. causal inference, or predicting how y would change if x were changed in a specified way.

Regression foundations

Ingredients of Bayesian regression

- $p(y|\theta, x)$: model likelihood, or even said the data-generating mechanism distribution (examples: Gaussian, Bernoulli, Poisson, Gamma, etc.)
- $p(\theta)$: prior for the parameter θ, possibly incorporating prior information about the parameters.

- $p(\theta|y)$: posterior distribution for the parameter, used for making inferential conclusions.
  Due to Bayes theorem, we have:

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

Linear model

Linear model

- response measurement $y_i, i = 1, \ldots, n$ continuous
- predictor matrix X, with dimensions $n \times p$, containing all the observed information for the units/individuals (examples: age, sex, comorbidities, etc.)
- model the expected value for the single unit, conditioned on his/her characteristics (covariates)

$$E(y_i|\beta, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

and the variance for each y_i is σ^2. The parameter vector is then $\theta = (\beta_1, \ldots, \beta_p, \sigma)$.

- likelihood of the model: $p(y|X, \theta) = \mathcal{N}(X\beta, \sigma^2)$.
- prior distributions for the β's.

Linear model

A Gaussian linear model is suitable when:

- the outcome of interest, the end-point in epidemiology, is continuous, bounded between $-\infty$ and $+\infty$.
- when the relationship between the covariates and the mean of the outcome is linear, or approximately linear.
- the distribution for the response variable given the covariates is (approximately) Gaussian.

Example: prostate data

Prostate data

- dataset `Prostate`: used to examine the correlation between the level of prostate-specific antigen and other covariates for men who were about to undergo a radical prostatectomy.

VARIABLE NAME	DESCRIPTION
`lpsa`	level of prostate-specific antigen
`lcavol`	log(cancer volume)
`lweight`	log(prostate weight)
`age`	age
`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion
`lcp`	log(capsular penetration)
`gleason`	Gleason score
`pgg45`	percentage Gleason scores 4 or 5

- Linear model for the response variable `lpsa`: which variables are influential for explaining the response variable?
- Check the R code

Extend linear models

The purpose of generalized linear models is to extend the idea of linear modelling to cases for which the linear relationship between X and $\mathrm{E}(y|X)$ or the normal distribution for each y is not appropriate, even after any transformation of the data.

Examples:

- when y is discrete (e.g. the number of phone calls received by a person in one hour)
- when the mean of y may be linearly related to X, but the variation term cannot be described by the normal distribution.
- when Y is a binary variable, such as die/not die, disease/not disease, etc.
- ...

We quickly deal with generalized linear models from a Bayesian perspective, although this class of models may be usefully applied from a classical perspective too.

Logistic regression

Logistic regression

- response measurements y_i, $i = 1, \ldots, n$, binary
- predictor matrix X, with dimensions $n \times p$, containing all the observed information for the units/individuals (examples: age, sex, comorbidities, etc.)
- model binary outcomes through the logit function. Being p_i the probability $\Pr(y_i = 1)$, we have

$$\text{logit}(p_i) \equiv \log \frac{p_i}{1 - p_i} = x_i \beta,$$

  where $\eta_i = x_i \beta$ is the linear predictor. It is easy to check we can invert the relationship above to obtain the probability: $\text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$.
- likelihood of the model: $p(y|X, \beta) = \text{Bernoulli}(p)$.
- prior distributions for the β's.

Logistic regression

A logistic regression model is suitable when:

- the outcome of interest, the end-point in epidemiology, is binary/dichotomous, taking then values 0 or 1, such as die/survive, disease/not disease, cured/not cured.
- when we can express the relationship between the covariates (such as age, sex, and others) and the probability that the event occurs. Pay attention: this relationship is not linear!
- the distribution for the response variable given the covariates is a Bernoulli.

Example: prostate data

Melanoma data

- dataset `Melanoma`: melanoma cancer clinical trials. Model the probability to observe relapse-free survival ($y_i = 0$) or survival with at least one relapse ($y_i = 1$) for some patients.
- CT patients randomly divided in two groups: placebo and interferon treatment. Is the treatment influential for relapse free-survival?

VARIABLE NAME	DESCRIPTION
`rfscens`	0 if relapse-free, 1 if with relapse
`age`	age of the patients
`trt`	interferon treatment, yes =1, no=0
`sex`	patients' sex
`perform`	walking ability, completely active =0, otherwise = 1
`nodes`	categorical: presence of nodes
`Breslow`	tumor's depth in mm

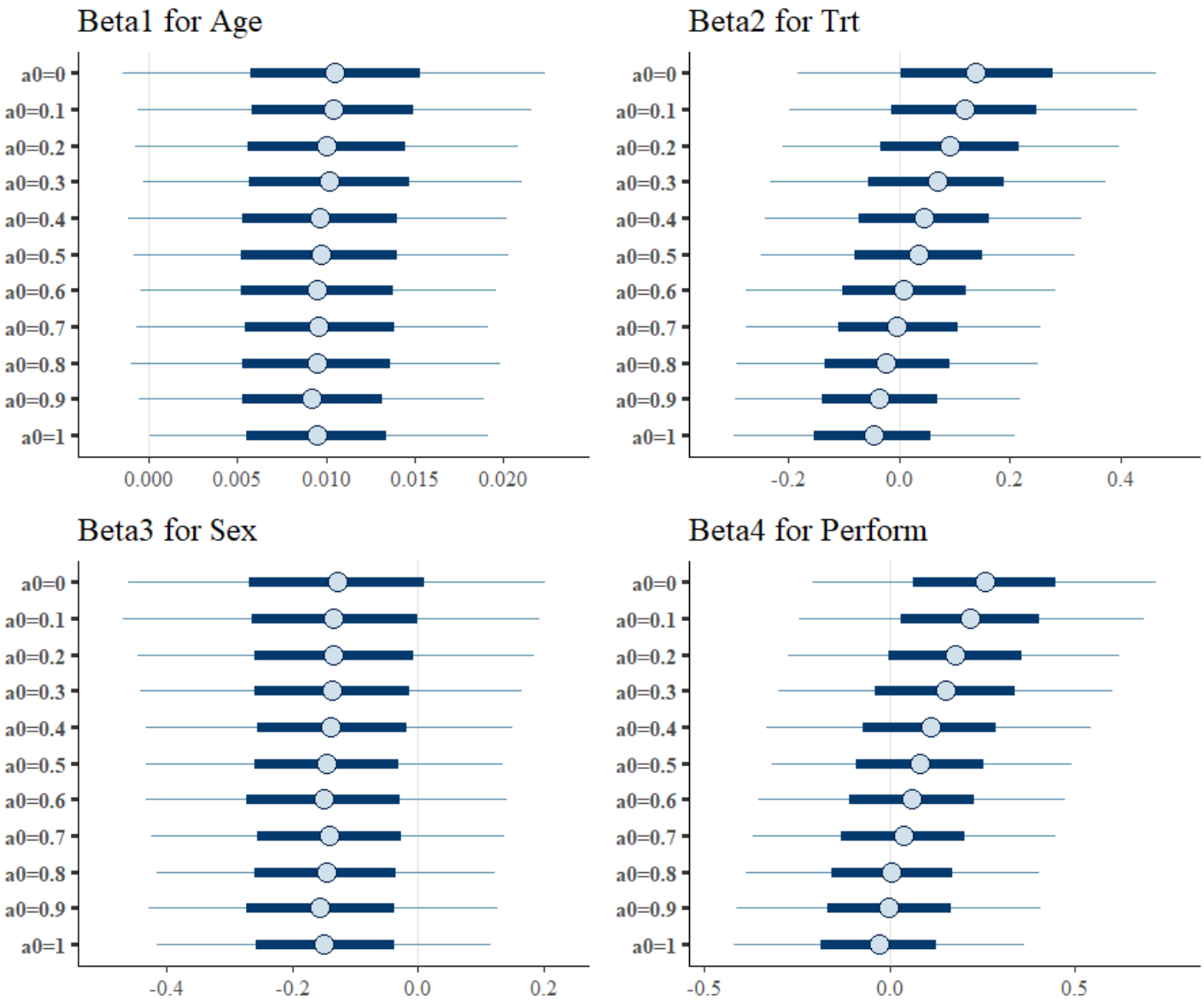
- Logistic regression for the response variable `rfscens`: which variables are influential for explaining the probability of a survival with at least one relapse?
- Check the R code

Power priors for melanoma data

To complete the melanoma application:

- A rather important question that can arise in Bayesian statistics is the following: how much past information to consider, especially if part of this information is in contrast with the current one?
- A widely used method in the Bayesian analysis of clinical trials is represented by the use of power-type prior distributions, also called power priors, which allow you to enter none, a fraction or all of the historical information of a previous trial to make inferential conclusions about the current data. The fraction of information is managed by a parameter a_0.
- Graphic to show how inferential conclusions (for example those on treatment) vary if we use different levels of historical information.

Power priors for melanoma data



Indice

Classical hypothesis testing

In the classical framework we compare two alternatives:

$$H_0 : \theta \in \Theta_0; H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 form a partition of the parameter space.

- In the classical approach, data are used to verify whether they are compatible with the null hypothesis through the calculation of the *p*-value, that is the probability that, under the null hypothesis, we may observe a sample which would give a result which is even *less convincing* under the null hypothesis (compared with the one we observe).

Classical hypothesis testing

Problems from classical hypothesis testing:

- dichotomization or tricothomization of the p-value according to a subjective scale, such as significant, weakly significant, not significant: however, the p-value is a continuous probabilistic measure, in [0,1]!
- interpretation of the p-value is tricky: this does not represent the probability that the null hypothesis is true nor the probability that the alternative hypothesis is true!! Rather, the p-value is a conditional probability...hard to interpret for clinicians.
- to use p-value, we need to pose ourselves in the perspective that the hypothesis H_0 holds, which could be a bit at odds from an interpretative point of view.
- when we evaluate a p-value, we never consider the alternative: the whole p-value finding is based on the null hypothesis...weird!
- Statisticians can cheat with p-values: this is the so-called p-hacking.
- Rather, we could want to get the posterior probability that a given hypothesis is true $\Rightarrow$ Bayesian hypothesis testing.

Bayesian hypothesis testing

- In a Bayesian framework, the most natural way to proceed is to quantify the **posterior weights** of the two competing hypootheses, that is, $\Pr(H_0|y)$, defined as:

$$\Pr(H_0|y) = \Pr(\Theta_0|y) = \int_{\Theta_0} \pi(\theta|y)d\theta,$$

where we assume that the random variable Θ is continuous.

The simplest case: Two simple hypotheses

This is the most elementary case (enough to present the irreconcilability)

$$H_0 : \theta = \theta_0, \quad H_1 : \theta = \theta_1$$

We may elicit the following prior distributions:

$$\pi_0 = \Pr(H_0), \quad \pi_1 = \Pr(H_1) = 1 - \pi_0.$$

Likelihood:

$$p(y|\theta_0), \quad p(y|\theta_1)$$

The relative weights of the two hypotheses is then given by the ratio:

$$\frac{\Pr(H_0|y)}{\Pr(H_1|y)} = \frac{\pi(\theta_0|y)}{\pi(\theta_1|y)} = \frac{\pi_0}{\pi_1}\frac{p(y|\theta_0)}{p(y|\theta_1)} \tag{1}$$

The simplest case: Two simple hypotheses

The posterior odds are the product of two terms:

- π_0/π_1 is the relative weight of the two hypotheses before observing the data.
- the second factor is usually denoted as:

$$BF_{01} = \frac{p(y|\theta_0)}{p(y|\theta_1)} \tag{2}$$

and is called Bayes factor: it represents the multiplicative factor which transforms the prior odds into the posterior odds. BF_{01} represents a measure of evidence in favour of H_0, where $BF_{01} > 1 (< 1)$ indicates that data favour H_0 (H_1).

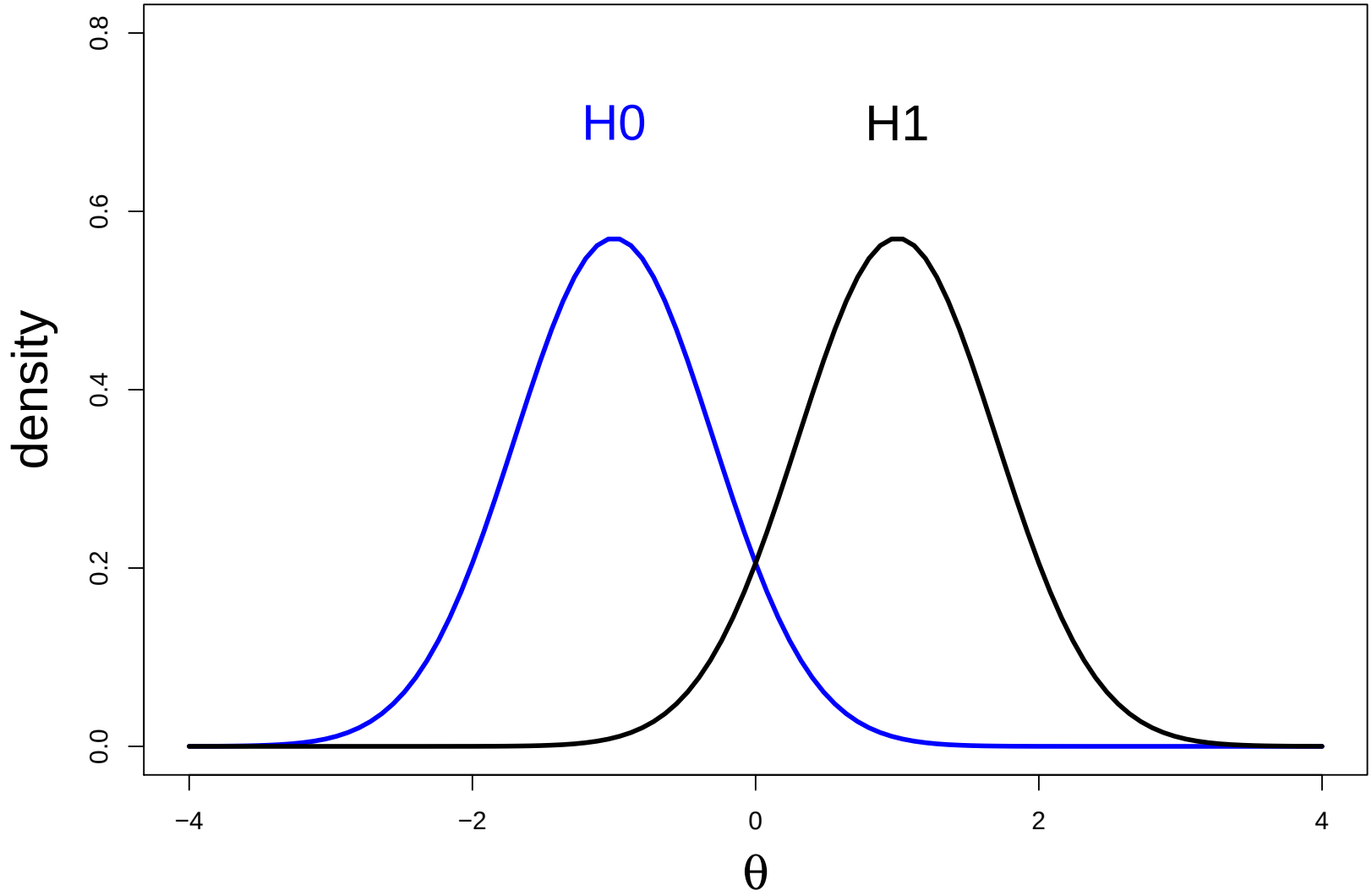
Toy example

Observe $\bar{y} \sim \mathcal{N}(\theta, \sigma^2)$, with $\sigma = 0.7$, we want to test:

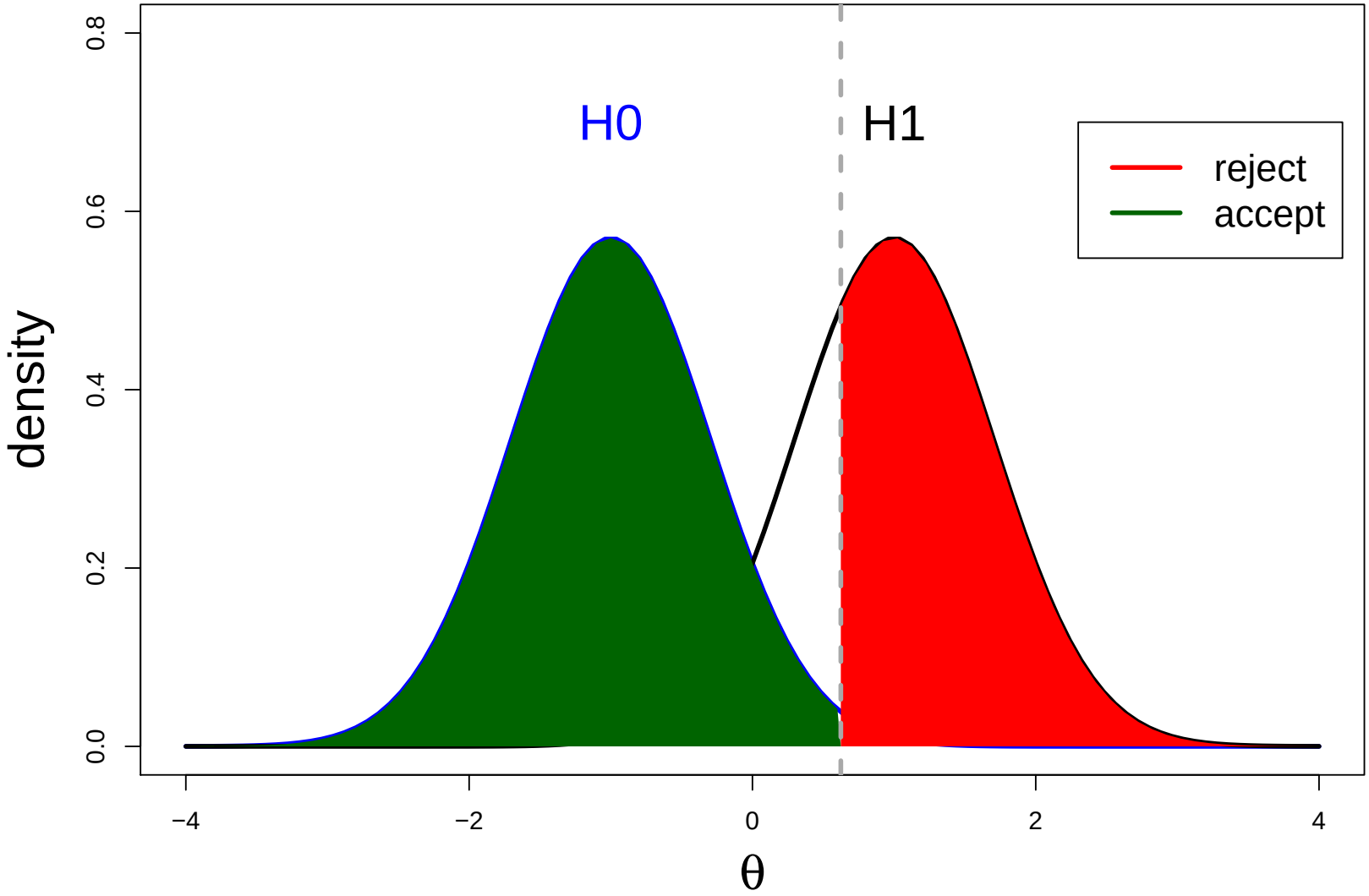
$$H_0 : \theta = \theta_0 = -1, \quad \text{vs.} \quad H_1 : \theta = \theta_1 = 1.$$

- Frequentist approach: fix $\alpha = 0.01$ and calculate the rejection region as $\frac{\bar{y}-\theta_0}{\sigma} > 2.32$, that is $\bar{y} > 0.624$.
- Bayesian approach: the data information is summarized by the Bayes factor $BF_{01} = p(y|\theta_0)/p(y|\theta_1)$.

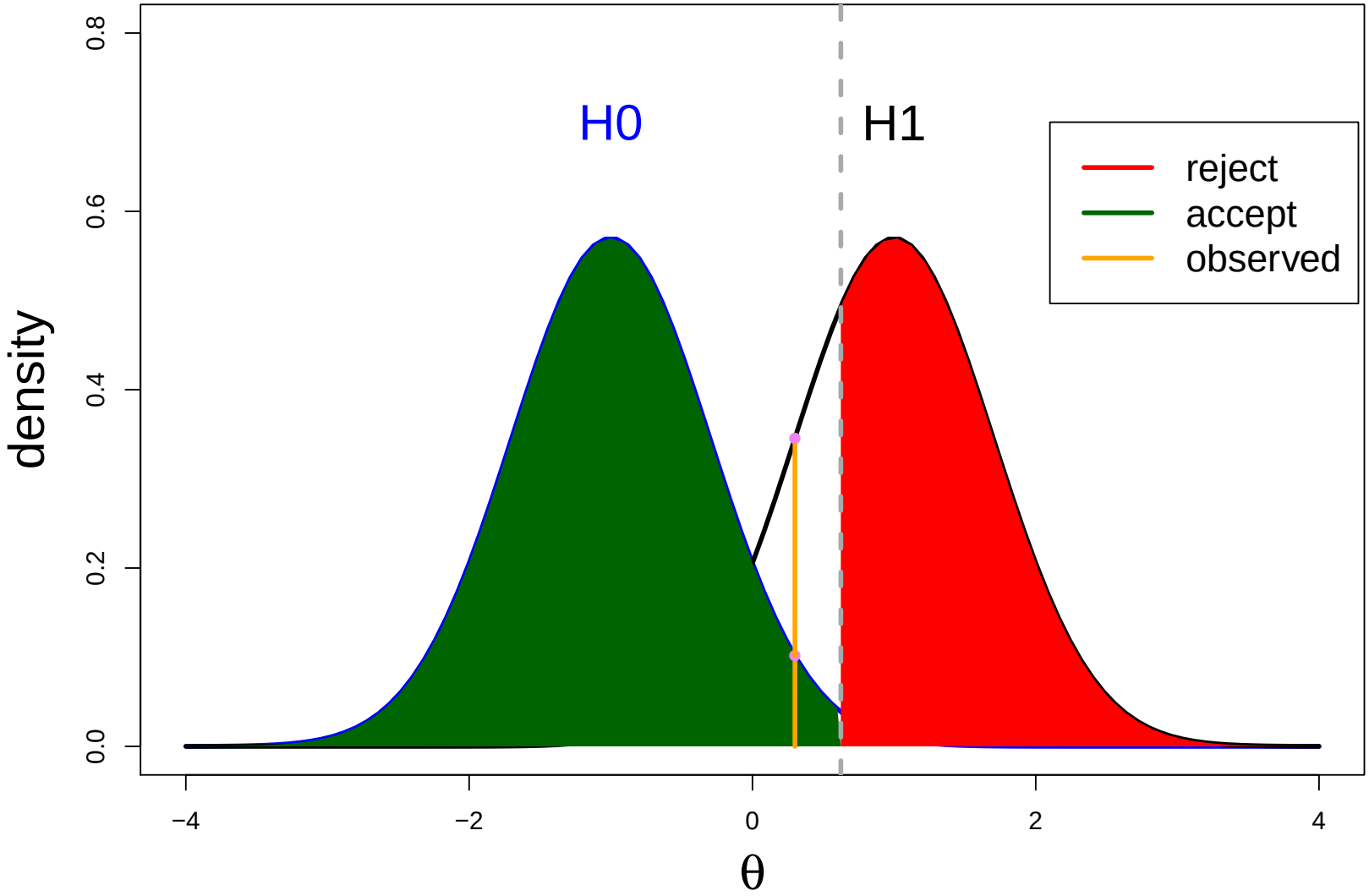
Toy example



Toy example: classical testing



Toy example: classical testing. If $\bar{y} = 0.3$...



Toy example

Comments:

- Frequentist approach tends to accept H_0 even for values which are more likely under H_1 (see what happens between 0 and 0.624...).
- Bayesian approach tends to accept H_0 only if the posterior probability of H_0 is higher than the posterior probability of H_1. If $\pi_0 = \pi_1 = 1/2$, H_0 is 'accepted' if the marginal likelihood under H_0 is higher than the marginal likelihood under H_1.

The general case

$$H_0 : \theta \in \Theta_0; \quad \text{vs.} \quad H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 form a partition of the parameter space.
Prior: two steps

1. Prior probabilities:

$$\pi_0 = \Pr(\Theta_0), \quad \pi_1 = \Pr(\Theta_1).$$

2. For H_i, $i = 0, 1$, $g_i(\theta)$ is the prior density of θ. Then the global prior is:

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \theta \in \Theta_0 \\ (1 - \pi_0) g_1(\theta) & \theta \in \Theta_1 \end{cases}$$

The general case

The beliefs about the two hypotheses are summarized by the posterior odds ratio:

$$\frac{\Pr(\theta \in \Theta_0|y)}{\Pr(\theta \in \Theta_1|y)} = \frac{\int_{\Theta_0} \pi(\theta|y)d\theta}{\int_{\Theta_1} \pi(\theta|y)d\theta} = \frac{\pi_0}{1-\pi_0} \frac{\int_{\Theta_0} p(y|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(y|\theta)g_1(\theta)d\theta}.$$

- The Bayes factor $BF_{01} = \frac{\int_{\Theta_0} p(y|\theta)g_0(\theta)d\theta}{\int_{\Theta_1} p(y|\theta)g_1(\theta)d\theta}$ is a ratio between marginal likelihoods. Their evaluation is central in testing hypotheses and model selection!
- Prior information plays a minor role through the densities g_0 and g_1.