Bayesian statistics 2/4

Core concepts & parameter estimation

Oussama Abdoun (MEng, PhD) – oussama.abdoun@pm.me

Bayesian Statistics – CRNL – dec 2024

On the menu

Part 1

- Bayes law
- Parameter estimation
- Credible intervals
- Frequentist vs. Bayesian statistics

Part 2

- Prior specification
- Numerical resolution (MCMC)
- Statistical testing
- Bayesian Factor
- Application to simple models: correlation, two-sample tests

Associated code at: gitlab.com/ousabd/statscourse2023/-/tree/master/ 2.03-2.04_bayesian-statistics

Recap of frequentist statistics

- In frequentist statistics, **parameter estimation** relies on maximizing *data* likelihood
- In frequentist statistics, the **interpretation** of **CIs** and ***p*-values** is convoluted and counter-intuitive:
  - they do not state anything about population parameters/hypotheses
  - CIs can only be interpreted in relation to a *series of replications* (e.g. meta-analysis)
  - *p*-values are not absolute measures of evidence
- The modern **NHST is fundamentally broken**:
  - it combines, in an **inconsistent** way, **two distinct procedures** (Fisher's and the Neyman-Pearson) that had fundamentally different goals and only superficial resemblance
  - for example, it is used to decide between two hypotheses but it specifies only one
  - the **publication system** (favoring $p < .05$) perpetuates these inconsistencies and creates downstream issues
- **Bayesian statistics** provides inference (CIs and hypothesis testing) with **intuitive** and **straightforward interpretation**

Medical test example

A 44-year old woman got a positive test from a mammogram.

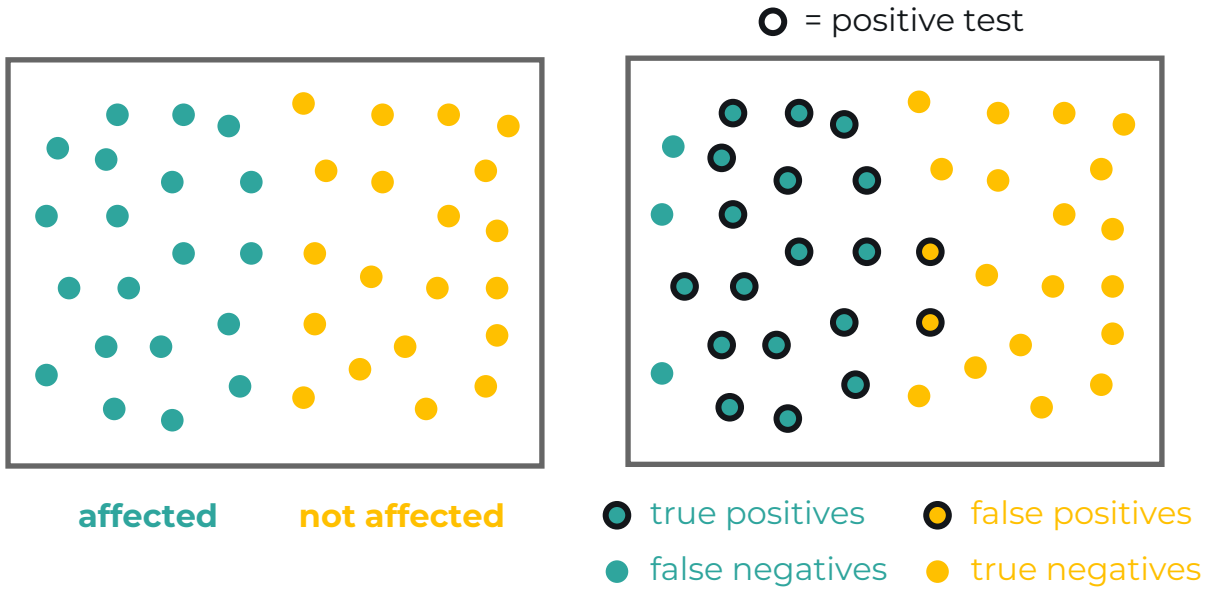
How worried should she be about having a breast cancer?

Frequentist-like approach: H_0 = no cancer.

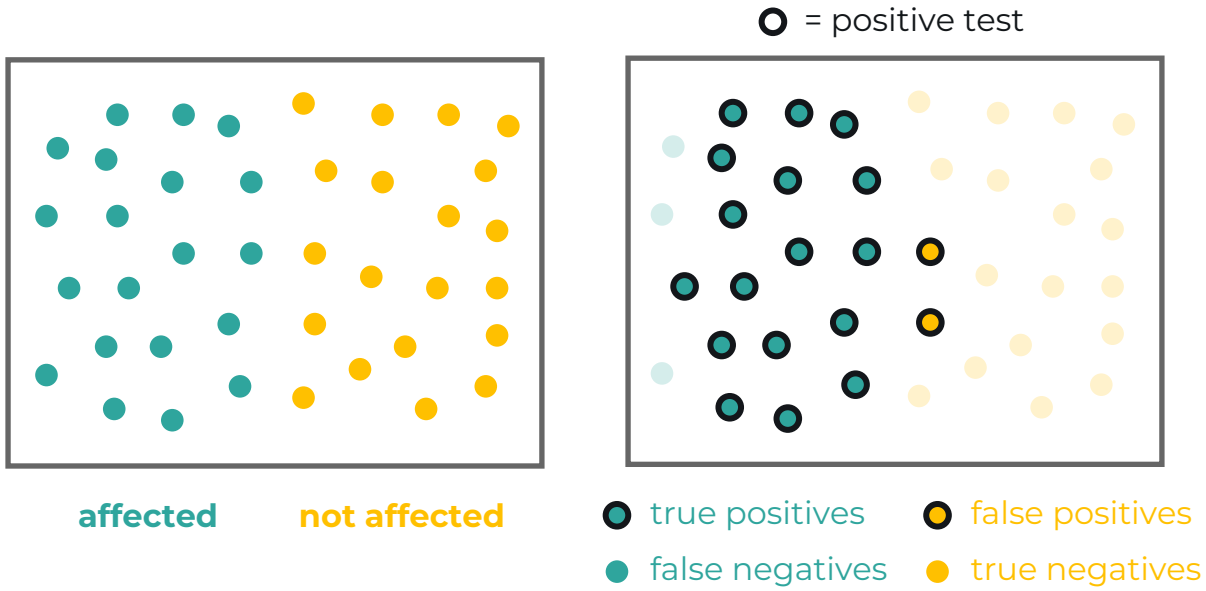
According to the medical literature, the probability of getting a *false positive* test (= observing the data assuming H_0) is .10

➔ $p > .05$ therefore we can not reject the null hypothesis

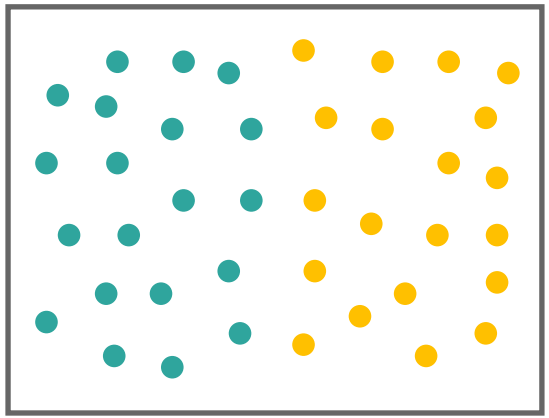
Classification errors & prevalence



Classification errors & prevalence



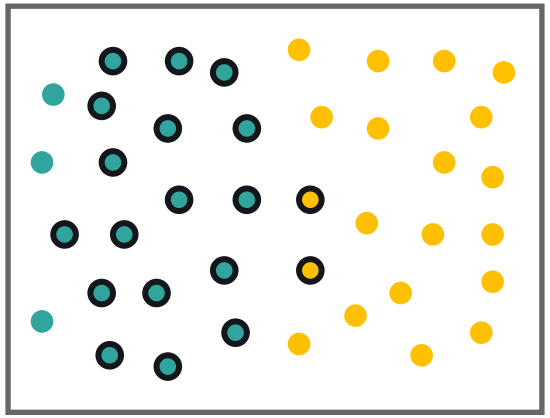
Classification errors & prevalence



affected not affected

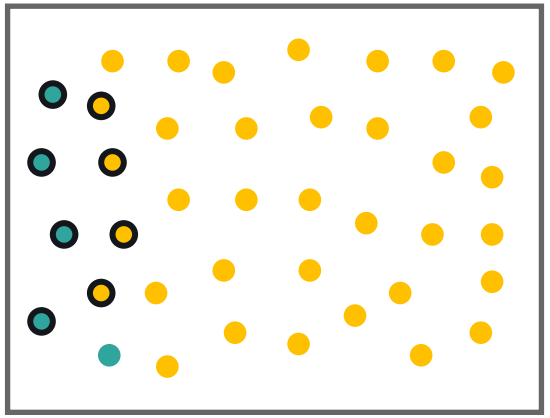
50% prevalence

= positive test



12% prevalence

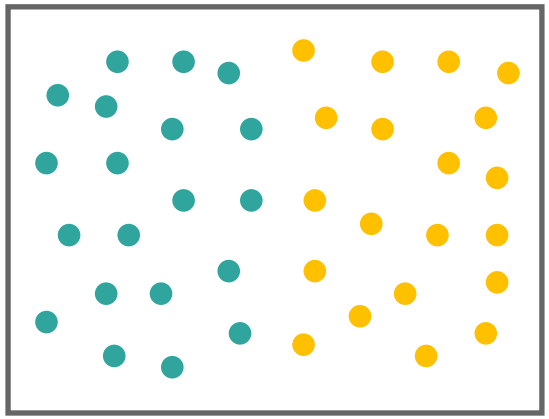
= positive test



true positives false positives

false negatives true negatives

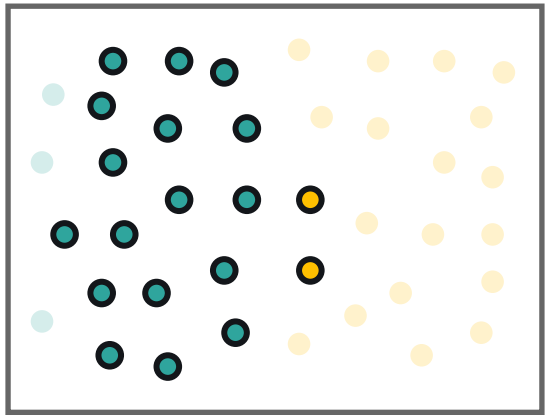
Classification errors & prevalence



affected not affected

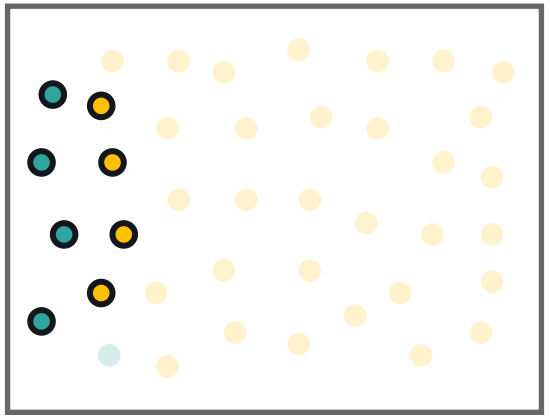
50% prevalence

= positive test



12% prevalence

= positive test



true positives false positives

false negatives true negatives

Medical test example

A 44-year old woman got a positive test from a mammogram.

How worried should she be about having a breast cancer?

Bayesian approach. Using Bayes' equation, we can write down and calculate the probability of having a cancer given the positive test:

$$p(\mathrm{C}|\mathrm{PT}) = \frac{p(\mathrm{PT}|\mathrm{C}).p(\mathrm{C})}{p(\mathrm{PT})}$$

According to the American Cancer Society, the false negative rate $p(\mathrm{NT}|\mathrm{C})$ is 12.5%, therefore the true positive rate $p(\mathrm{PT}|\mathrm{C})$ is 87.5%. In addition, the prevalence of breast cancer in women aged between 40 and 50 is about 1% = $p(\mathrm{C})$. Therefore:

$$p(\mathrm{PT}) = p(\mathrm{PT}|\mathrm{C}).p(\mathrm{C}) + p\left(\mathrm{PT}|\overline{\mathrm{C}}\right).p\left(\overline{\mathrm{C}}\right)$$
$$= .875 \times .01 + .1 \times .99 = .1078$$

$$p(\mathrm{C}|\mathrm{PT}) = \frac{.875 \times .01}{.1075} \sim .081 = 8{,}1\%$$

*(compare to the probability **prior** to observing the test result: **1%**)*

What we “discovered”

- Making a **statement about a hypothesis** (given the data) requires information on the probability of the hypothesis **prior** to seeing the data

Modern statisticians have developed extensive mathematical techniques, but for the most part have rejected the notion of the probability of a hypothesis, and thereby deprived themselves of any way of saying precisely what they mean when they decide between hypotheses.

Harold Jeffreys
Theory of probability (1961)

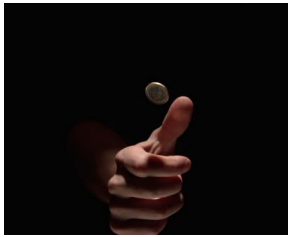
1.

Elements of probability theory

Basic concepts

Tossing a coin is a process that results in a random outcome.

Let X denote the outcome. X is a *random variable*.



X can take one of two values: heads or tails, each with a probability 0.5:

$\Pr(X = heads) = 0.5$ *and* $\Pr(X = tails) = 0.5$

Every time the coin is tossed, we obtain a *realization of* X : x_1, x_2, x_3, etc.

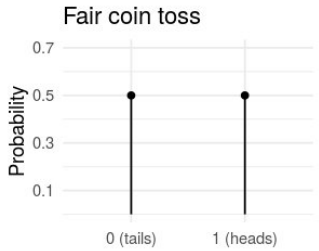
We expect about 50% of the realizations to be heads, and 50% to be tails.

If we assign the value 1 to heads and 0 to tails, then the *expected value* of X is 0.5:

$$\mathrm{E}(X) = \Pr(X = 1) \times 1 + \Pr(X = 0) \times 0 = 0.5$$

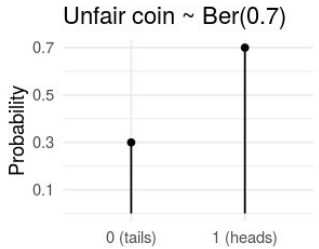
Basic concepts

The *probability distribution* is the function that gives the probabilities of occurrence for different possible outcomes of a random process.



The set of values that X can take is the *support*. Here, it is *discrete*.

The probability distribution is *uniform*: the two possible outcomes, *heads* and *tails*, are equally likely.



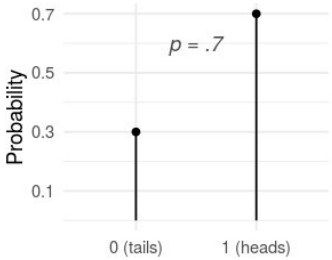
Suppose the coin is unfair: it yields *heads* with probability p ($p \neq 0.5$).

The probability distribution of X is still discrete, but not uniform anymore. X is said to follow the *Bernouilli distribution* with *parameter* p:

$$X \sim Ber(p)$$

Common discrete probability distributions

Bernouilli

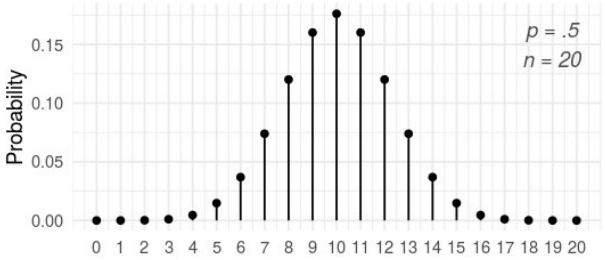


$Ber(p)$

Binary outcome (no / yes, 0 / 1, fail / success, etc.) with p = probability of "success"

- Toin coss
- Outcome of a medical treatment (cured / not cured)

Binomial

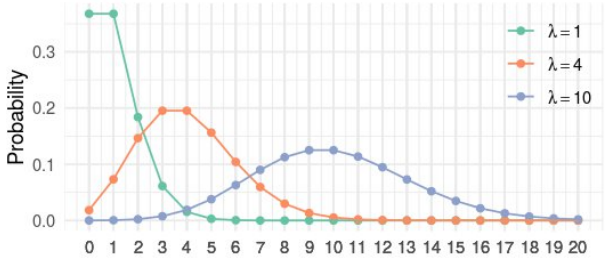


$B(n,p)$

Number of successes in a sequence of n trials of a Bernouilli process of parameter p

- Number of successful treatments in a group of patients

Poisson



$\mathbf{Pois}(\lambda)$

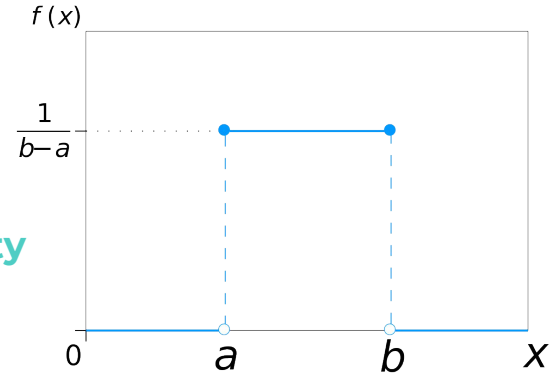
Number of events occurring in a fixed interval of time (or space), with events occurring randomly with a constant mean rate of λ

- number of spontaneous spikes in 1s bins
- number of DNA mutations per million year

Continuous distribution

Continuous distributions are distributions whose **support** is **infinite**.

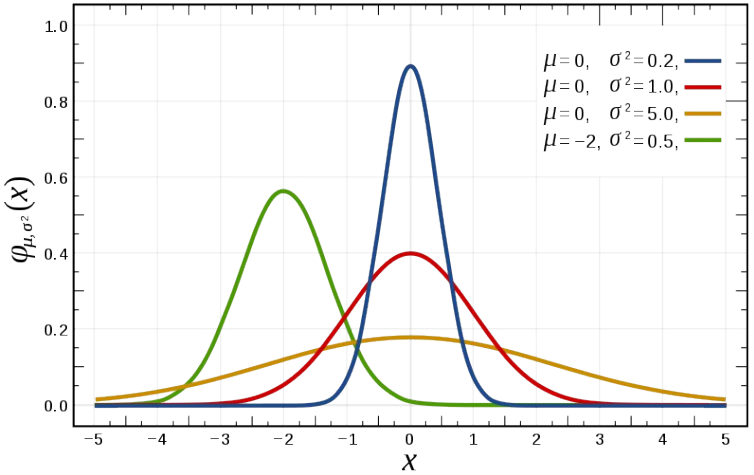
For example, the **uniform distribution** between 0 and 1, denoted $U(0,1)$ can take any real value comprised between 0 and 1 with equal probability.



⚠ The interpretation of the values of a **probability density function** (continuous distributions) is slightly different from the values of a **probability mass function** (discrete distributions): **relative vs. absolute likelihood**.

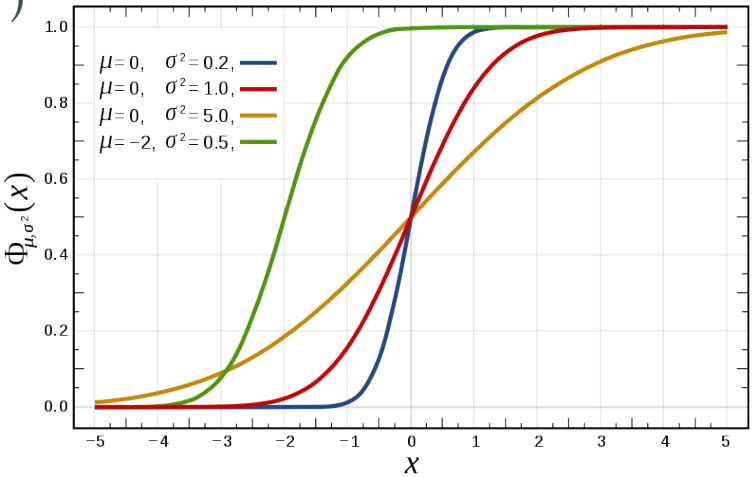
The normal distribution

$\mathcal{N}(\mu,\sigma^2)$



Probability density function

Red curve = **standard normal distribution = Z**



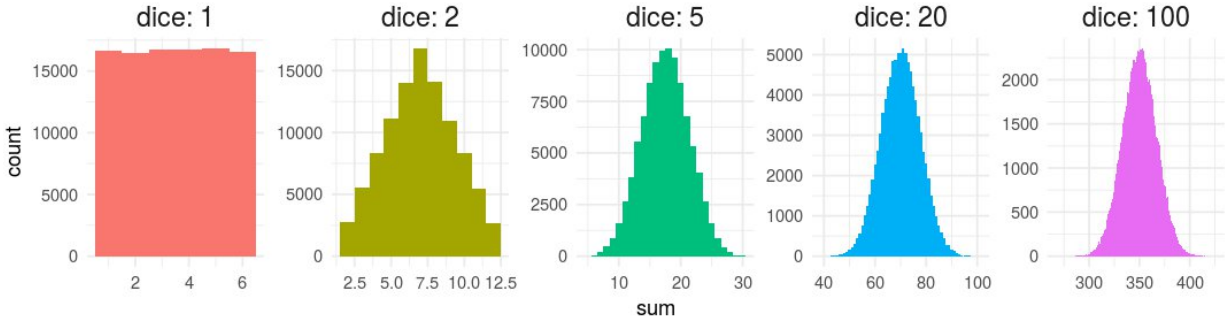
Cumulative density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

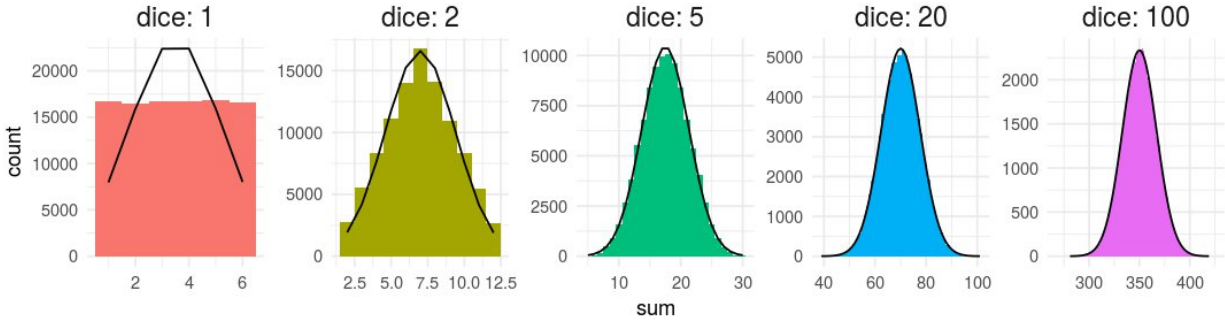
Why is the normal distribution so common?

Let's roll many dice. Dice rolls follow a discrete, uniform distribution.

Question: What is the distribution of the sum?



Result: The distribution of the sum of uniformly distributed variables looks more and more gaussian when we increase the number of variables!



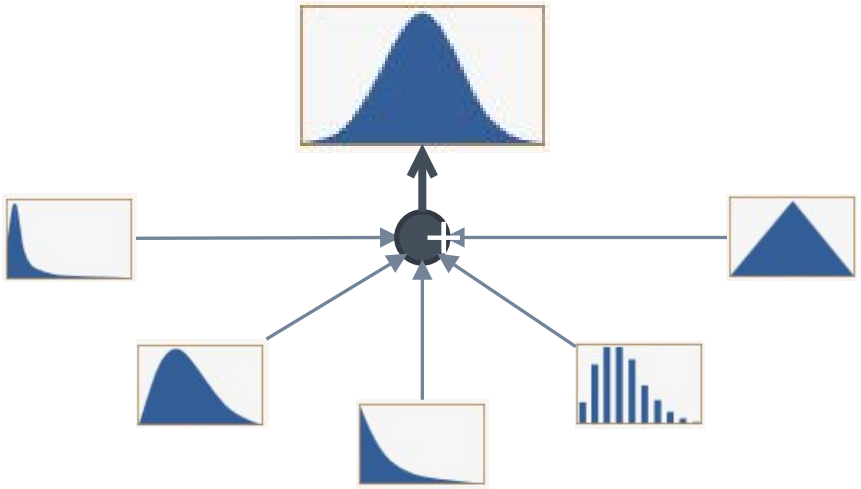
Why is the normal distribution so widespread?

Because of the **central limit theorem**: the sum of **independent** random variables tends toward a normal distribution, even if the original variables themselves are not normally distributed.

The properties of most objects, including living beings (size, weight, composition, etc.) result from the superimposed action of many sub-processes.

Whatever distributions underlie these processes, the macroscopic result will likely follow a normal distribution thanks to the central limit theorem.

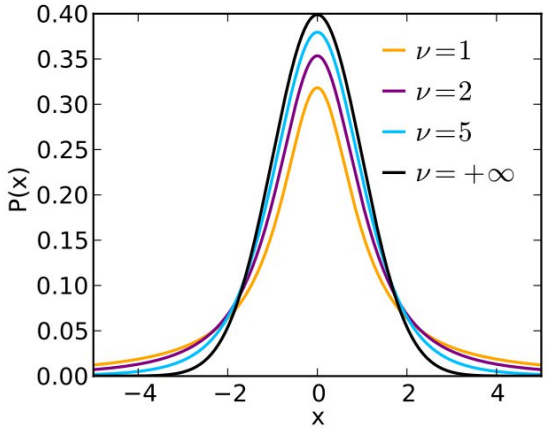
Resulting property: **normal distribution**



Sub-processes / factors with **various distributions**

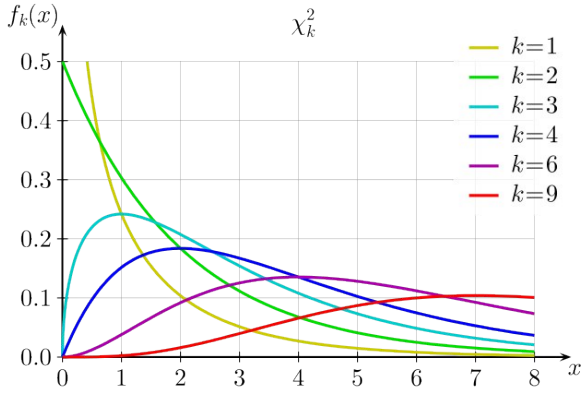
Continuous probability distributions derived from the normal distribution

Student's t



Difference between the sample mean and the population mean, divided by the sample standard deviation. ν = **degrees of freedom** (dof)

Chi square

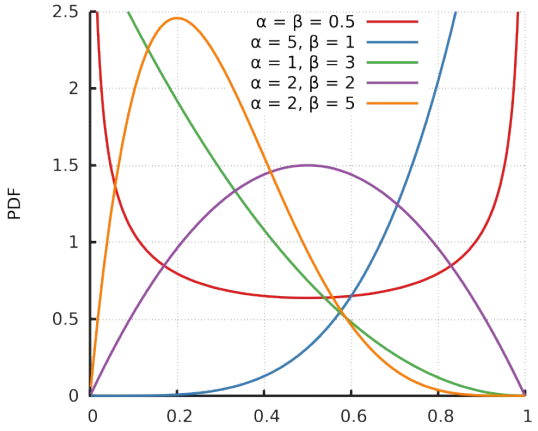


Sum of the squares of k independent, standard normal random variables.

Both are extremely useful for **hypothesis testing** and **confidence intervals**

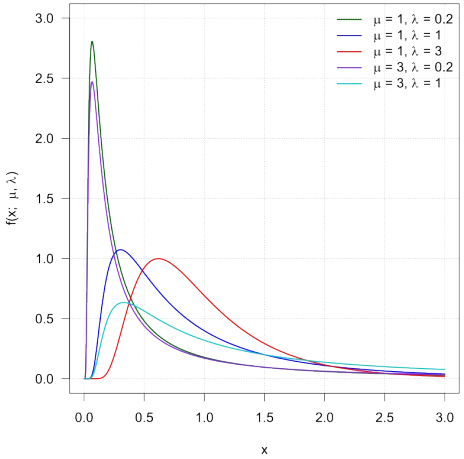
Other continuous probability distributions

Beta



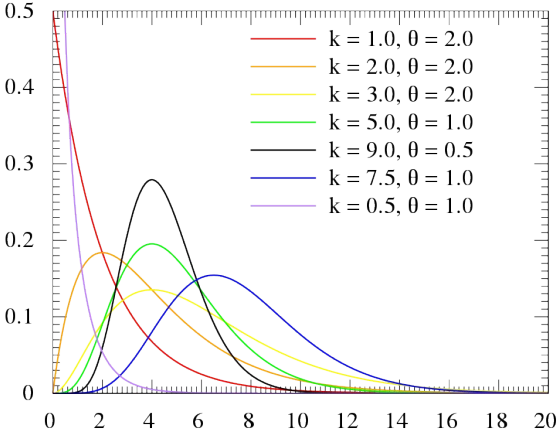
For variables constrained between 0 and 1, such as proportions and frequencies.

Inverse Gaussian



The time for a positive drift, random process to reach a certain level, as in drift diffusion models. Good for RTs.

Gamma



Time for k events to occur in a random Poisson process of rate $1/\theta$

Bayes' theorem: derivation

conditional probability of A given B $p(A|B)$

joint probability $p(A,B)$

marginal probability of A $p(A)$ and B $p(B)$

Interactive visualization

Bayes' theorem: derivation

The **conditional probability** of A given B $p(A|B)$ is the **joint probability** $p(A,B)$ divided by the **marginal probability** of B $p(B)$

1 $p(A|B) = p(A,B)/p(B)$

Symmetrically:

2 $p(B|A) = p(B,A)/p(A)$

visualize on setosa.io/conditional/

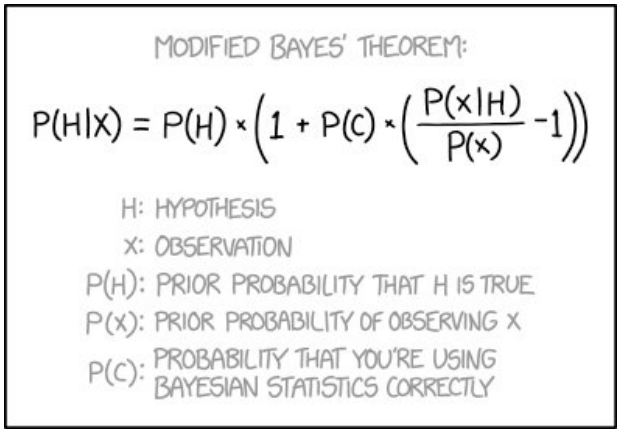
Combining 1 and 2 we obtain Bayes' law/equation/theorem:

$$p(B|A) = \frac{p(A|B).p(B)}{p(A)}$$

Reverend Thomas Bayes
1702-1761

2.

Bayesian statistics: estimation



*"Don't forget to add another term for "*probability that the Modified Bayes' Theorem is correct*"."*

xkcd.com/2059

Bayesian inference

Bayesian inference is the application of the Bayes' equation to the estimation of the parameters of a statistical model.

$$p(B|A) = \frac{p(A|B) \cdot p(B)}{p(A)}$$

(conditional) likelihood

posterior $p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)}$ prior

marginal likelihood

where θ denotes the model parameter(s) and y the observed data.

Bayesian inference: interpretations

The Bayes equation...

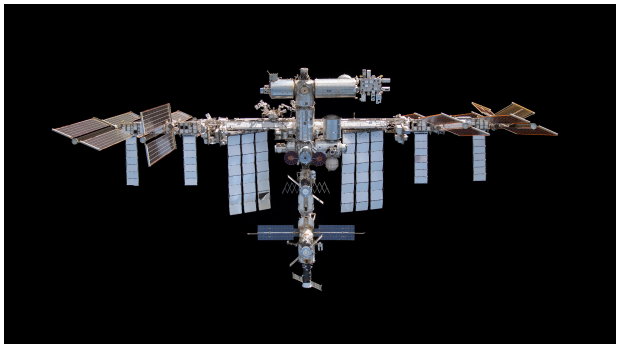
...integrates prior knowledge with new information to yield the best possible expectation

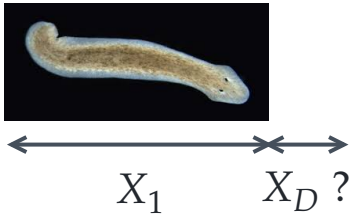
...offers the best compromise between subjective beliefs and observed data

...updates prior belief with new evidence

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

Application : the space planarians





What we “discovered”

- Making a **statement about a hypothesis** (given the data) requires information on the probability of the hypothesis **prior** to seeing the data

- Both **the prior and the data “pull” the posterior** toward their means, proportional to their **precision**

- The **precision of the posterior** is always higher than both the precision of the prior and the precision of the likelihood

  ⇒ it measures the **amount of combined information**

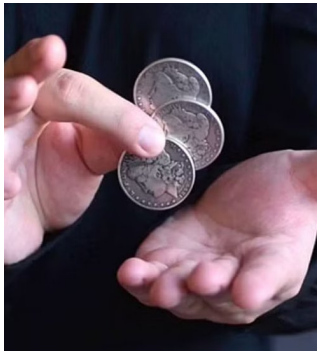
Application : the coin example

Scenario A

Standard modern coin

Scenario B

Ancient, irregular coin



Scenario C

Magic show coin

The coin example: the likelihood

$$p(q|k) = \frac{p(k|q) \cdot p(q)}{p(k)}$$

Let's apply the Bayesian equation for model parameters...

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

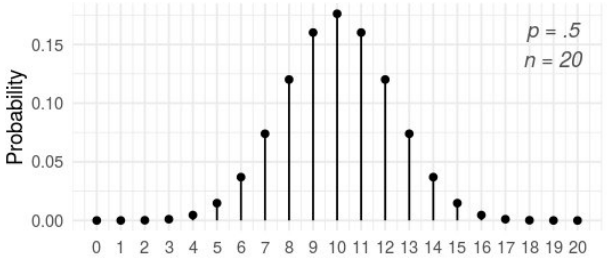
...to the specific situation of assessing whether a coin is fair:

$$p(q|k) = \frac{p(k|q) \cdot p(q)}{p(k)}$$

We model the number of heads coming out of n coin tosses as a random variable X that follows a binomial distribution. Therefore $\boldsymbol{q}$ denotes the parameter of the binomial distribution (the **probability of the coin to land on heads**) and $\boldsymbol{k}$ the **number of heads out of n coin tosses**.

$$X \sim B(n,q)$$

$$\Rightarrow p(X = k|q,n) = f(q) = \frac{n!}{k!\,(n-k)!} q^k (1-q)^{n-k}$$



probability mass function of a binomial distribution

The coin example: the prior

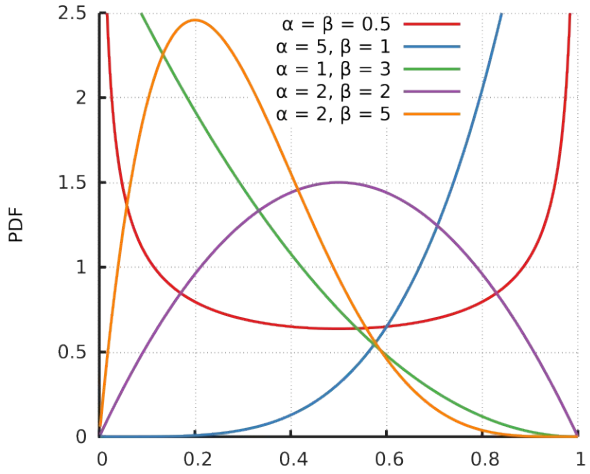
$$p(q|k) = \frac{p(k|q) \cdot p(q)}{p(k)}$$

The model parameter q is unknown. In the Bayesian paradigm, it is also modeled as a random variable. q is a continuous quantity between 0 and 1. The Beta distribution is well suited to model this kind of variable:

$$q \sim \text{Beta}$$

$$\Rightarrow p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\text{B}(\alpha,\beta)}$$

where B is the Beta *function*. α and β are parameters of the *prior* Beta distribution, and are called **hyperparameters** to avoid the confusion with parameters of the *sampling model*.



probability density functions of various Beta distributions

The coin example: the posterior

$$p(q|k) = \frac{p(k|q) \cdot p(q)}{p(k)}$$

The denominator $p(k)$ can be calculated using **marginalization**: $$p(k) = \int_0^1 p(k|q)\, p(q)\, dq = constant$$

Combining everything, and after some (omitted) maths, we finally obtain a mathematical expression for the posterior distribution:

$$p(q = x|k,n) = \frac{x^{\alpha-1+k}(1-x)^{\beta-1+n-k}}{\mathrm{B}(\alpha + k, \beta + n-k)}$$

Which is the equation of a Beta distribution too, just like the prior!

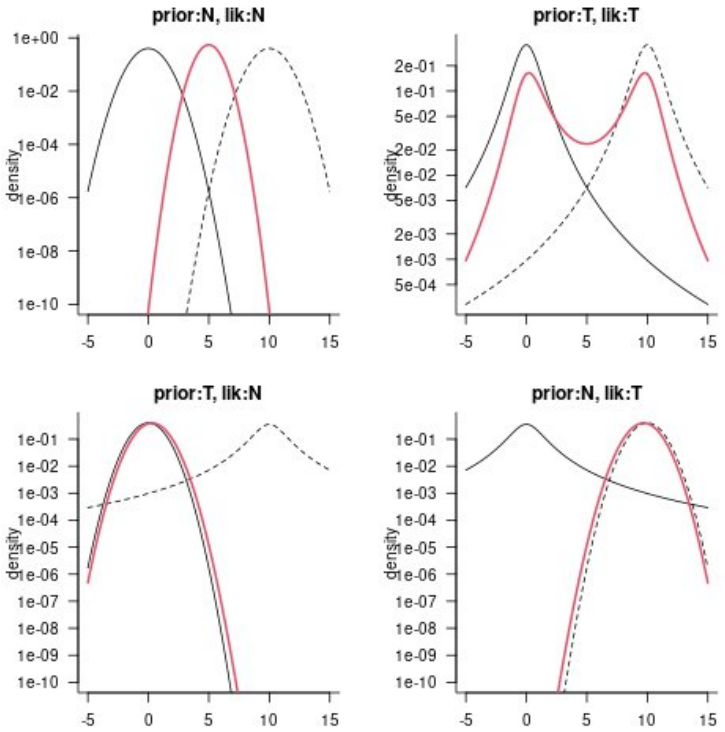
$$q|k,n \sim \mathrm{Beta}\left(\alpha + k, \beta + n-k\right)$$

What we “discovered”

- Making a **statement about a hypothesis** (given the data) requires information on the probability of the hypothesis **prior** to seeing the data
- Both **the prior and the data “pull” the posterior** toward their means, proportional to their **precision**
- The **precision of the posterior** is always higher than both the precision of the prior and the precision of the likelihood

  ⇒ it measures the **amount of combined information**
- **Well-chosen distributions make the calculation of the posterior easy**
  Given a model of how the data is generated (the likelihood function), a **conjugate distribution** can be chosen for the prior, so that the posterior will have the same distribution as the prior (see other examples on Wikipedia)
  ⇒ otherwise, **numerical simulations** (see parts 3 and 4 of this course)

Surprising combinations of prior & likelihood

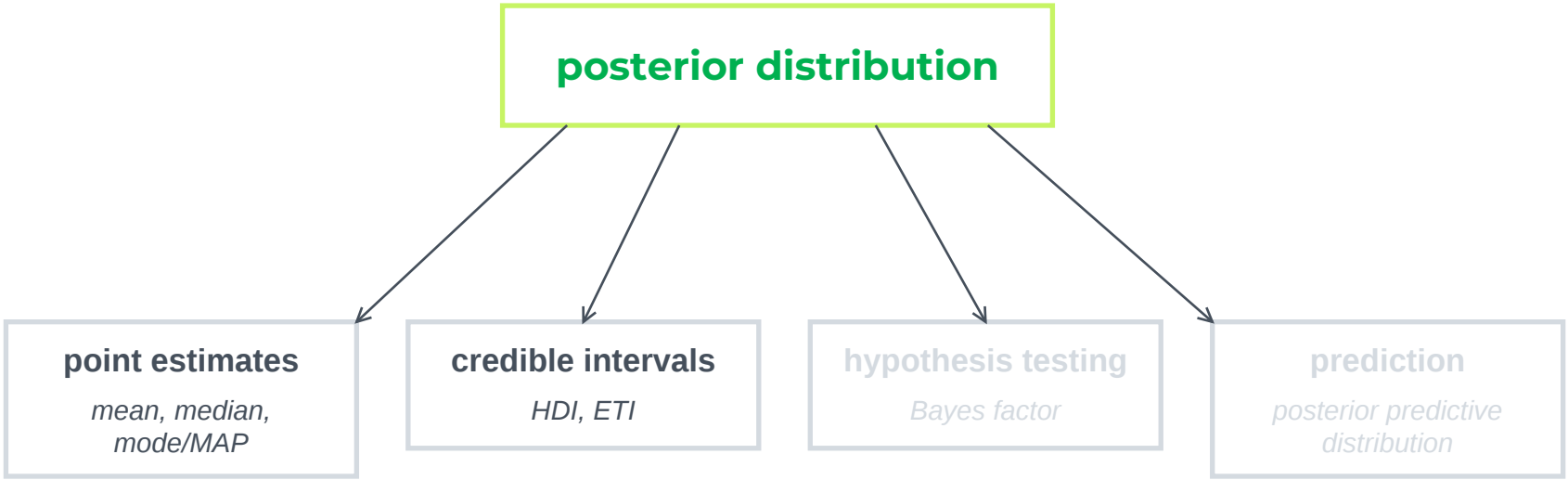


When one distribution is more heavy-tailed than the other (e.g. a t-distribution and a Gaussian), the heavy-tailed one “gives way” to the other in the posterior.

But when both are heavy-tailed and far apart, a strange posterior can arise!

The central role of the posterior distribution

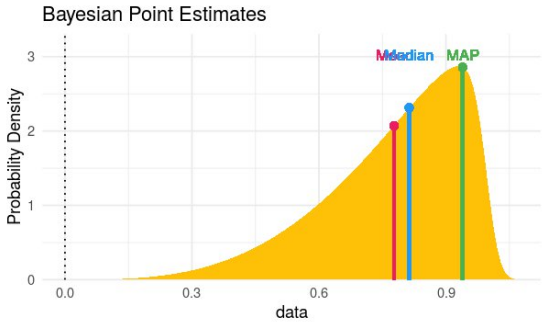
In Bayesian statistics, all results are derived from the posterior distribution



Point and interval estimates

`bayestestR` package in the `easystats` ecosystem

easystats.github.io/bayestestR/



point estimates

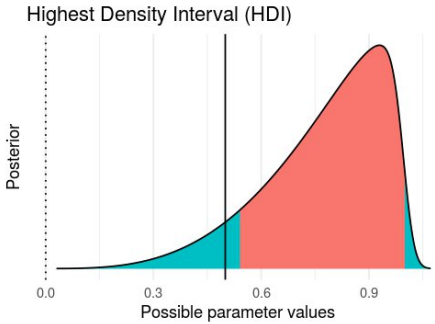
mean, median, mode/MAP

`point_estimate()`

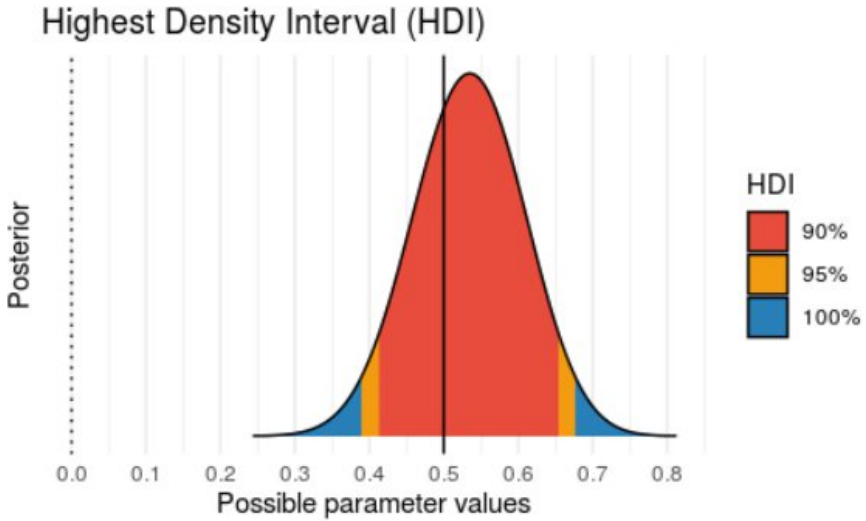
credible intervals

HDI, (ETI)

`hdi()`



Credible interval: 95% or 90%?



Compared to the 95%, the 90% credible interval is...

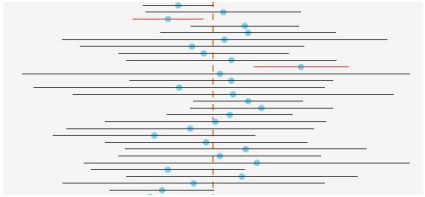
+ more stable to numerical errors
- less conservative

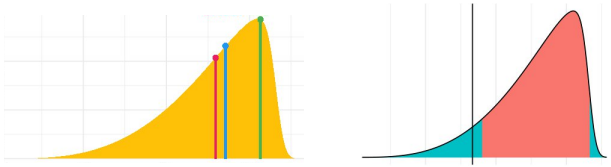
→ Use **95%** if there are more than **10.000 samples** of the posterior distribution

In `bayestestR`, the default is **89%** (!) to highlight the arbitrariness of the confidence level.

Frequentist vs. Bayesian statistics

	Frequentist	Bayesian
Definition of probability	Long-run frequency of events	Degree of belief / certainty
View on model parameters	*True value:* unknown *Estimate:* fixed	*True value:* unknown *Estimate:* probabilistic
Method of estimation	From the data only	From the posterior (data + prior)
Uncertainty interval	“Confidence interval” Confidence level is a property of the procedure, not of the interval	“Credibility intervals” Confidence level is a statement about the uncertainty around the estimate





3.

Prior specification

Choosing the prior

“I know nothing of the modeled phenomena and I want the prior to express that high uncertainty.”

diffuse priors

“There are physical constraints that dictate what values are unlikely.”

weakly informative priors

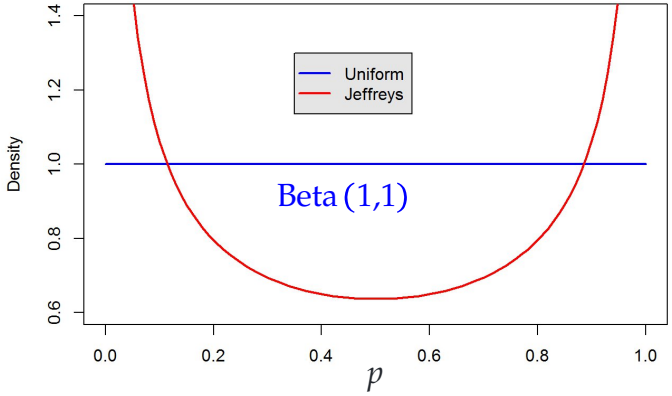
“I have expert judgment based on experience, existing literature and technical considerations.”

informative priors

Diffuse priors

Flat/uniform prior

Assign equal probability to all parameter values.



Frequentist estimate: $\hat{p}_{MLE} = \frac{k}{n}$

Bayesian estimate: $\hat{p}_{MAP} = \frac{\alpha_{posterior} - 1}{\alpha_{posterior} + \beta_{posterior} - 2} = \frac{\alpha_{prior} + k - 1}{\alpha_{prior} + k + \beta_{prior} + n - k - 2}$

$$\hat{p}_{MLE} = \hat{p}_{MAP} \iff \alpha_{prior} = \beta_{prior} = 1$$

Under uniform prior, the posterior MAP coincides with the frequentist MLE!

Diffuse priors

Jeffreys prior

= a prior that is **invariant to reparametrizations** of the sampling model

Example for a Bernoulli model

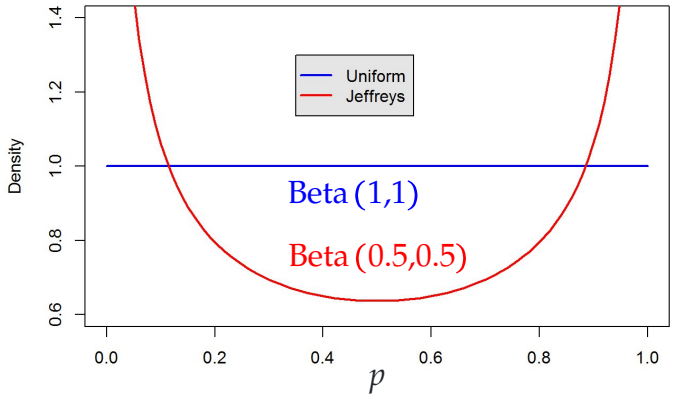
Standard parametrization q $\xrightarrow{f}$ **Odds** parametrization $\psi = f(q) = \frac{q}{1-q}$

Jeffreys prior for q $\xrightarrow{f}$ Jeffreys prior for ψ

Reference prior

= the prior that maximizes the contribution of the data relatively to the prior.

In univariate settings, the Jeffreys prior is the reference prior !



Informative / *reasonable* priors

Imagine you want to test the theory that some symptoms of schizophrenia arise from the disruption of low-level perceptual mechanisms. You design a visual discrimination task and model the ***difference in response time*** *between patients and a control group.*

upright

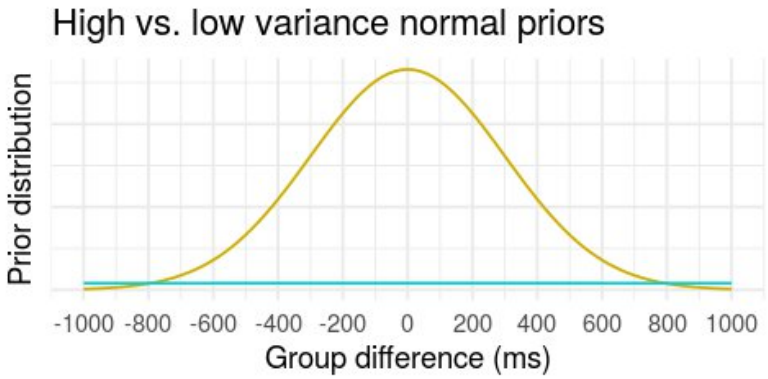
vs.

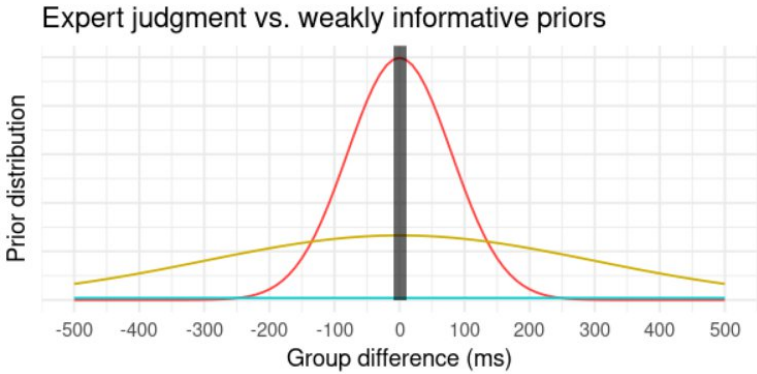
upside-down

What prior for the population difference in RT?

Informative / *reasonable* priors

Imagine you want to test the theory that some symptoms of schizophrenia arise from the disruption of low-level perceptual mechanisms. You design a visual discrimination task and model the ***difference in response time*** *between patients and a control group.*





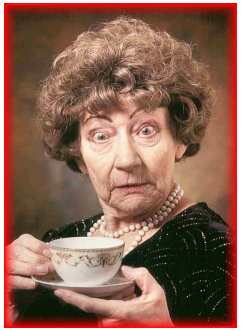
physical / logical constrains	domain knowledge	literature	past data

weakly informative → *highly informative*

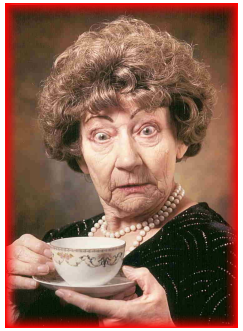
Prior as a bet

*I can **taste** which of milk or tea has been poured first!I*

*I can **tell** which of milk or tea is poured first in another room!*



or



Right 9 times out of 10 !

↓

Exceptional palate?

*Would you use the **same standard** of evidence?*

Right 9 times out of 10 !

↓

Extra-sensory perception?

Bayesian statistics offer a **formal treatment** of context and expectations

Prior as a bet

*What is the effect of **drinking coffee after dinner** compared to no coffee on the **time to fall asleep**?*

treatment / condition

dependent variable

The Sheffield Elicitation Framework

`elicit()` in `SHELF`

Prior elicitation

physical / logical constraints

What is the minimum possible value?

The maximum?

domain knowledge

What is the median value?

The quartile values?

`elicit()` in **SHELF**

literature

Posterior distribution from a Bayesian random effect meta-analysis

Or (more simply)

the distribution of effect sizes in a meta-analytic study, or informal review

past data

Posterior distribution from a previous study

On Bayesian « subjectivity »

The Bayesian framework requires the setting of a **prior belief,** which some people object to due to its **subjective** nature.

But **what is subjectivity?** Definitions from the Oxford Dictionary:

1. “the quality of being based on or influenced by personal feelings, tastes, or opinions.”

2. “the quality of existing in someone's mind rather than the external world.”

Scientific practice is full of decisions and judgments that are constantly justified, analyzed and debated between scientists:

- selection of scientific questions worthy of being investigated
- design of experiments, including the calculation of statistical power
- choice of dependent variables, preprocessing steps, statistical models, etc.

There is no loss in dispensing with the illusion of objectivity in hypothesis testing. Researchers are acclimated to elements of social negotiation and subjectivity in scientific endeavors. *Negotiating the appropriateness of various alternatives is no more troubling than of other elements, including design, operationalization, and interpretation. (...) We have the communal infrastructure to evaluate and critique the specification of alternatives. This (...) is* ***vastly preferable to the current practice, in which significance tests are mistakenly regarded as objective****. Even though inference is subjective, we can agree on the boundaries of reasonable alternatives.*

Rouder et al. (2009)