Bayesian inference and Monte Carlo methods

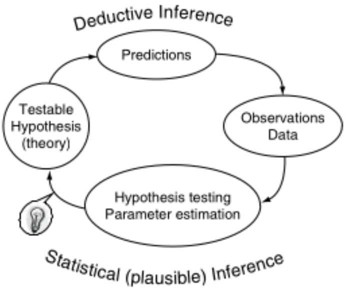
Lecture 1: Basic Statistical Conecpts and Introduction to Bayesian Statistics

Susana J. Landau

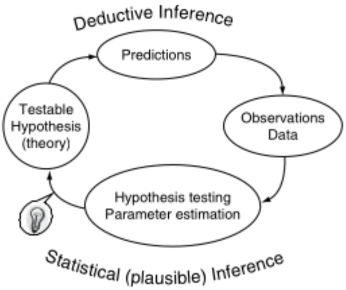
6 de octubre de 2025

Outline

The problem we want to solve

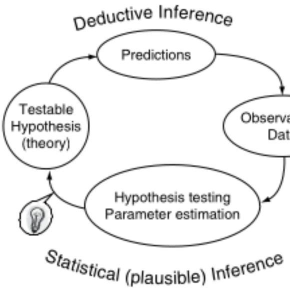


The problem we want to solve



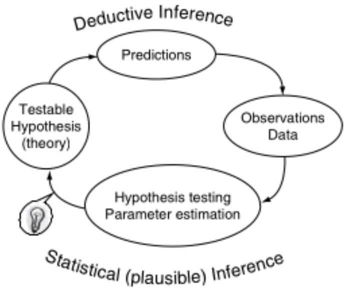
- Given one or more data sets :Which is the theoretical model that can better explain the data?, Is there only one?

The problem we want to solve



- Given one or more data sets :Which is the theoretical model that can better explain the data?, Is there only one?
- Which are the values of the free parameters of our model that can explain the data?

The problem we want to solve



- Given one or more data sets :Which is the theoretical model that can better explain the data?, Is there only one?
- Which are the values of the free parameters of our model that can explain the data?
- If several theoretical models are able to explain the data: which one should I chose?

Concepts we want to clarify

$$\boxed{H_0 = 73{,}0 \pm 0{,}4 \ \mathrm{km\,s^{-1}Mpc^{-1}}}$$

- What does this statement mean?

Concepts we want to clarify

$$\boxed{H_0 = 73{,}0 \pm 0{,}4 \ \mathrm{km\,s^{-1}Mpc^{-1}}}$$

- What does this statement mean?
- It never means that the true value of H_0 lies between $72{,}6\ \mathrm{kms^{-1}Mpc^{-1}}$ and $73{,}4\ \mathrm{kms^{-1}Mpc^{-1}}$.

Concepts we want to clarify

$$\boxed{H_0 = 73{,}0 \pm 0{,}4 \text{ km}\,\text{s}^{-1}\text{Mpc}^{-1}}$$

- What does this statement mean?
- It never means that the true value of H_0 lies between $72{,}6 \text{ kms}^{-1}\text{Mpc}^{-1}$ and $73{,}4 \text{ kms}^{-1}\text{Mpc}^{-1}$.
- It means that if we repeat the experiment many times, $68\,\%$ of the measurements will lie in that range (frequentist approach).

Concepts we want to clarify

$$\boxed{H_0 = 73{,}0 \pm 0{,}4\ \mathrm{km\,s^{-1}Mpc^{-1}}}$$

- What does this statement mean?
- It never means that the true value of H_0 lies between $72{,}6\ \mathrm{kms^{-1}Mpc^{-1}}$ and $73{,}4\ \mathrm{kms^{-1}Mpc^{-1}}$.
- It means that if we repeat the experiment many times, $68\,\%$ of the measurements will lie in that range (frequentist approach).
- It means that the probability distribution of H_0 is a Gaussian with mean value $73\ \mathrm{kms^{-1}Mpc^{-1}}$ and standard deviation $0{,}4\ \mathrm{kms^{-1}Mpc^{-1}}$ (Bayesian approach).

Statistical Methods

- When comparing data and models we are typically doing one of two things :
  - Hypothesis testing : we have a set of N measurements $x_i \pm \sigma_i$ which a theorist says should have values μ_i. How probable is it that these measurements would have been obtained, if the theory is correct?
  - Parameter estimation : we have a parameterized model which describes the data, such as $y = f(x, \theta)$, and we want to determine confidence intervals of those parameters.
- In this course we will focus on parameter estimation

Basics statistical concepts

- Population: complete set of elements about which we wish to conduct research or draw conclusions. These elements can be objects, events, situations, or groups of individuals. The population comprises all possible observations or cases relevant to the research question under study.
- Sample: subset of elements selected from a population in order to conduct a specific analysis.

Statistical inference

- Inferential statistics is a branch of statistics focused on studying properties and drawing conclusions about a population based on information obtained from a sample of that population.
- Statistical inference is the process of using a data sample to infer the distribution that generated those data. That is, given a sample $X_1, X_2, ..., X_n$, statistical inference allows us to determine the underlying distribution from which the data were drawn and/or its properties.

Statistical inference

- Inferential statistics is a branch of statistics focused on studying properties and drawing conclusions about a population based on information obtained from a sample of that population.
- Statistical inference is the process of using a data sample to infer the distribution that generated those data. That is, given a sample $X_1, X_2, ..., X_n$, statistical inference allows us to determine the underlying distribution from which the data were drawn and/or its properties.
- Some examples of problems that can be solved with statistical inference:
  - Given a sample $X_1, X_2, ..., X_n$, and a parametric statistical model, estimate the parameters of that model.

Statistical inference

- Inferential statistics is a branch of statistics focused on studying properties and drawing conclusions about a population based on information obtained from a sample of that population.
- Statistical inference is the process of using a data sample to infer the distribution that generated those data. That is, given a sample $X_1, X_2, ..., X_n$, statistical inference allows us to determine the underlying distribution from which the data were drawn and/or its properties.
- Some examples of problems that can be solved with statistical inference:
  - Given a sample $X_1, X_2, ..., X_n$, and a parametric statistical model, estimate the parameters of that model.
  - Given a sample $X_1, X_2, ..., X_n$ and a non-parametric statistical model, determine the underlying distribution and/or its properties.

Statistical inference

- Inferential statistics is a branch of statistics focused on studying properties and drawing conclusions about a population based on information obtained from a sample of that population.
- Statistical inference is the process of using a data sample to infer the distribution that generated those data. That is, given a sample $X_1, X_2, ..., X_n$, statistical inference allows us to determine the underlying distribution from which the data were drawn and/or its properties.
- Some examples of problems that can be solved with statistical inference:
  - Given a sample $X_1, X_2, ..., X_n$, and a parametric statistical model, estimate the parameters of that model.
  - Given a sample $X_1, X_2, ..., X_n$ and a non-parametric statistical model, determine the underlying distribution and/or its properties.
  - Given the following data set $(X_1, Y_1)...(X_n, Y_n)$, perform a regression analysis.

Basic statistical concepts

Given a population with elements $x_1, x_2, ..., x_N$. We define the following quantities:

- Population mean value

$$\boxed{\mu = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

- Population variance or dispersion around the mean value::

$$\boxed{\mathrm{V} = S^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

  where S is the standard deviation

- In general, a population may be too large for the researcher to collect precise data from every element due to size limitations.

A statistical estimator is a quantity that provides summarized information about a data set. Given a data set $x_1, x_2, ..., x_N$ of size N (a sample from the population), we can define the following estimators:

- The sample mean is an unbiased estimator of the population mean μ.

$$\boxed{\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i}$$

- The sample variace $(\hat{V})$ is an unbiased estimater of the population variance (V)

$$\boxed{\hat{V} = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

- An unbiased estimator is a statistic whose expected value is equal to the true value of the population parameter it estimates. This means that, on average, the estimator recovers the true value of the parameter when the sampling process is repeated many times. $< \hat{V} >= V$.

We can quote the error of each of these statistical estimators:

- $$\boxed{\text{Error in the sample mean}: \frac{\sigma}{\sqrt{N}}}$$
- $$\boxed{\text{Error in the sample variance}: \sigma^2\sqrt{\frac{2}{N-1}}}$$

Let us consider N independent measurements x_i of a quantity x, each with an associated error ϵ_i. What is the best estimator for x?

Let us consider N independent measurements x_i of a quantity x, each with an associated error ϵ_i. What is the best estimator for x?

- The simple average $\hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is not an optimal estimator, since we want to give more weight to the more precise measurements.
- The weighted average is a better estimator:

$$\hat{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

  where $w_i = 1/\epsilon_i^2$.
- A good estimator for the variance is:

$$\mathrm{Var}(\hat{x}) = \sigma^2 = \sum_{i=1}^{N} \frac{1}{\epsilon_i^2}$$

- A probability distribution is a function that assigns a probability to each particular value (or range of values) of a discrete or continuous variable.

Probability distributions

- A probability distribution is a function that assigns a probability to each particular value (or range of values) of a discrete or continuous variable.
- A probability distribution must be normalized:

$$\int_{-\infty}^{+\infty} P(x)dx = 1$$

or, if the variable is discrete,

$$\sum_{i=1}^{N} P(x_i) = 1$$

Probability distributions

- A probability distribution is a function that assigns a probability to each particular value (or range of values) of a discrete or continuous variable.
- A probability distribution must be normalized:

$$\int_{-\infty}^{+\infty} P(x)dx = 1$$

or, if the variable is discrete,

$$\sum_{i=1}^{N} P(x_i) = 1$$

- The probability that the variable x falls within the range $[x_1, x_2]$ is given by

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} P(x)dx$$

Probability distributions

A probability distribution can be characterized by its:

- mean value

$$\boxed{\mu_x = \int_{-\infty}^{+\infty} x P(x) dx}$$

- variance

$$\boxed{\sigma_x^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 P(x) dx}$$

Probability distributions

A probability distribution can be characterized by its:

- mean value

$$\boxed{\mu_x = \int_{-\infty}^{+\infty} x P(x) dx}$$

- variance

$$\boxed{\sigma_x^2 = \int_{-\infty}^{+\infty} (x - \mu_x)^2 P(x) dx}$$

- In the case of a discrete probability distribution, the mean value is :

$$\boxed{\mu_x = \sum_{i=1}^{N} x_i P(x_i)}$$

- and the variance is given by

$$\boxed{\sigma_x^2 = \sum_{i=1}^{N} (x_i - \mu_x)^2 P(x_i)}$$

Cumulative Probability Distribution

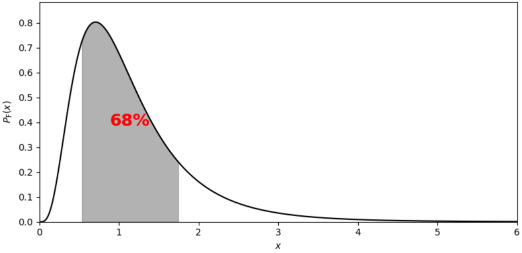
- The cumulative probability distribution $P(x)$ is defined as the probability that the variable is less than or equal to a given value.
- The cumulative distribution is bounded between 0 and 1 and is a monotonically increasing function.
- It can be applied to both continuous and discrete variables.
- It can be calculated as follows. If $p(x)$ is the probability distribution of the variable x:

$$P(x) = \int_{-\infty}^{x} p(x)dx$$

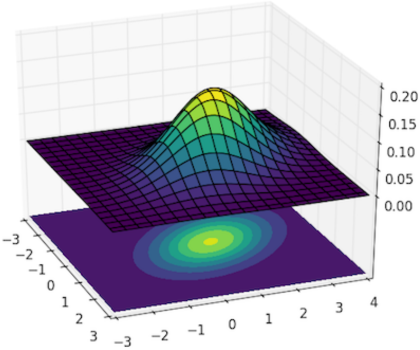
Confidence intervals

- A confidence interval is a range of values that is likely to contain the true value of a population parameter with a certain level of confidence.
- If a variable x is assumed to follow a certain probability distribution $P(x)$, the corresponding 68 % confidence interval $(a_{\mathrm{low}}, a_{\mathrm{high}})$ is defined:

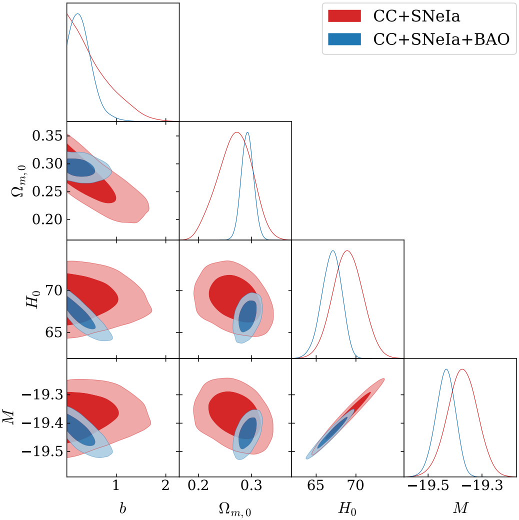
$$\boxed{\int_{a_{\mathrm{low}}}^{a_{\mathrm{high}}} P(x)dx = 0{,}68}$$



- What are the values of a and b such that $\int_{-a}^{a}\int_{-b}^{b} P(x,y)dxdy = 0{,}68$?



Confidence regions



- Confidence regions for the cosmological parameters assuming the

Outline

Random variables

- A random variable is a function that assigns a numerical value to each possible outcome of a random experiment:
  - It's a variable whose values depend on the outcome of a random phenomenon.
  - We cannot predict its exact value before performing the experiment.
  - It is characterized by its probability distribution.
- Random variables can be of two types: Continous and Discrete:
  - Discrete random variables take isolated and countable values (integers, finite values). Examples: number of AGN in a galaxy sample, number of photons reaching a detector.

Random variables

- A random variable is a function that assigns a numerical value to each possible outcome of a random experiment:
  - It's a variable whose values depend on the outcome of a random phenomenon.
  - We cannot predict its exact value before performing the experiment.
  - It is characterized by its probability distribution.
- Random variables can be of two types: Continous and Discrete:
  - Discrete random variables take isolated and countable values (integers, finite values). Examples: number of AGN in a galaxy sample, number of photons reaching a detector.
  - Continuous random variables can take any value in a continuous interval. Examples: Temperature of an object, decay time of a particle.

Random variables

- A random variable is a function that assigns a numerical value to each possible outcome of a random experiment:
  - It's a variable whose values depend on the outcome of a random phenomenon.
  - We cannot predict its exact value before performing the experiment.
  - It is characterized by its probability distribution.
- Random variables can be of two types: Continous and Discrete:
  - Discrete random variables take isolated and countable values (integers, finite values). Examples: number of AGN in a galaxy sample, number of photons reaching a detector.
  - Continuous random variables can take any value in a continuous interval. Examples: Temperature of an object, decay time of a particle.
  - In statistics, we work with samples of random variables to infer properties of the population or the random process that generates them.

Probability distributions

Some examples of probability distributions that we will use in this course are:

- Binomial distribution
- Poisson distribution
- Normal or Gaussian distribution

Binomial distribution

- It applies to situations where there is a random process with two possible outcomes, with probabilities p and $1 - p$. Example: tossing a coin.

Binomial distribution

- It applies to situations where there is a random process with two possible outcomes, with probabilities p and $1-p$. Example: tossing a coin.
- The binomial distribution is a discrete probability distribution that gives the probability of obtaining n occurrences of a particular outcome (with probability p) out of two possible outcomes, after performing N independent trials.

$$\boxed{P_{\text{binomial}}(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}}$$

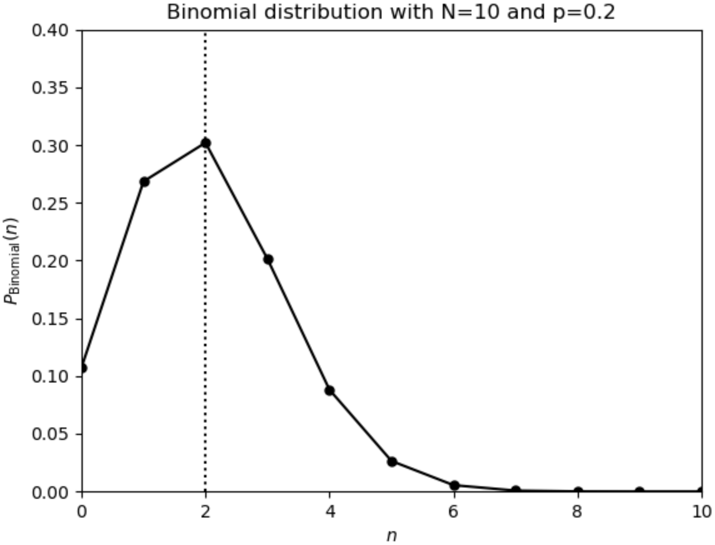
Binomial distribution

- It applies to situations where there is a random process with two possible outcomes, with probabilities p and $1-p$. Example: tossing a coin.
- The binomial distribution is a discrete probability distribution that gives the probability of obtaining n occurrences of a particular outcome (with probability p) out of two possible outcomes, after performing N independent trials.

$$\boxed{P_{\text{binomial}}(n) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}}$$

- The mean value of the binomial distribution is $\boxed{\bar{n} = pN}$
- The variance of the binomial distribution is $\boxed{\sigma^2(n) = Np(1-p)}$.

Binomial distribution



Example

- In the Hubble Space Telescope guide star catalogue, $60\,\%$ of the objects are binary stars. How large a sample should be chosen to ensure that the probability of the sample containing at least 2 non-binary stars is $99\,\%$?

Poisson distribution

- Applies to a discrete random process where we are counting something in a fixed interval
  Examples: radiative decay, photons arriving at a CCD.

Poisson distribution

- Applies to a discrete random process where we are counting something in a fixed interval
  Examples: radiative decay, photons arriving at a CCD.
- The Poisson distribution gives the probability of observing n events in a fixed interval, with mean rate μ:

$$\boxed{P_{\mathrm{Poisson}}(n) = \frac{\mu^n e^{-\mu}}{n!}}$$

Poisson distribution

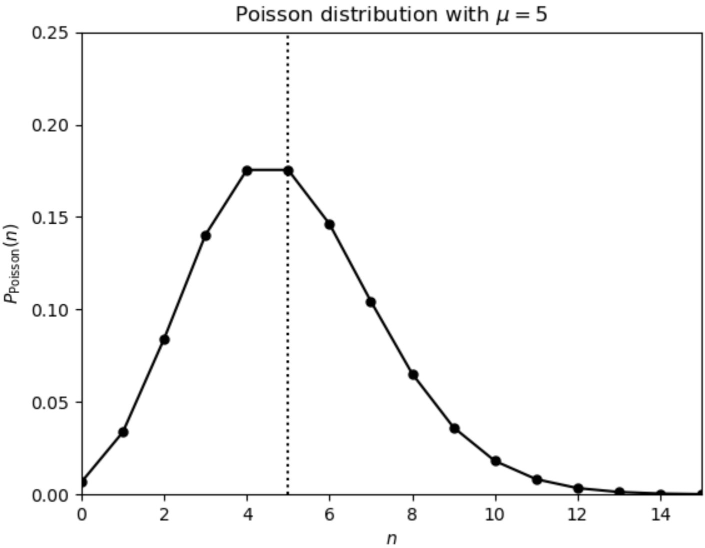
- Applies to a discrete random process where we are counting something in a fixed interval
  Examples: radiative decay, photons arriving at a CCD.
- The Poisson distribution gives the probability of observing n events in a fixed interval, with mean rate μ:

$$\boxed{P_{\text{Poisson}}(n) = \frac{\mu^n e^{-\mu}}{n!}}$$

- The mean and the variance of the Poisson distribution are equal:

$$\boxed{\bar{n} = \sigma^2(n) = \mu}$$

Poisson distribution



Example

- The density of quasars on the sky is known to be 20 per $\deg^2$. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?

Example

- The density of quasars on the sky is known to be 20 per deg^2. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?
- The mean number of quasars in an area A is $\mu = 20A$.

Example

- The density of quasars on the sky is known to be 20 per deg^2. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?
- The mean number of quasars in an area A is $\mu = 20A$.
- The probability of finding at least one quasar is $P(n \geq 1) = 1 - P(0)$.

Example

- The density of quasars on the sky is known to be 20 per deg^2. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?
- The mean number of quasars in an area A is $\mu = 20A$.
- The probability of finding at least one quasar is $P(n \geq 1) = 1 - P(0)$.
- P(0)= $\frac{\mu^0 e^{-\mu}}{0!} = e^{-\mu} = e^{-20\mathrm{A}}$.

Example

- The density of quasars on the sky is known to be 20 per deg^2. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?
- The mean number of quasars in an area A is $\mu = 20A$.
- The probability of finding at least one quasar is $P(n \geq 1) = 1 - P(0)$.
- P(0)= $\frac{\mu^0 e^{-\mu}}{0!} = e^{-\mu} = e^{-20\mathrm{A}}$.
- We want $P(n \geq 1) \geq 0{,}99$, which implies $e^{-20\mathrm{A}} \leq 0{,}01$.

Example

- The density of quasars on the sky is known to be 20 per $\deg^2$. What area of sky would we need to survey to ensure a 99 % chance of finding at least one quasar?
- The mean number of quasars in an area A is $\mu = 20A$.
- The probability of finding at least one quasar is $P(n \geq 1) = 1 - P(0)$.
- P(0)= $\frac{\mu^0 e^{-\mu}}{0!} = e^{-\mu} = e^{-20\mathrm{A}}$.
- We want $P(n \geq 1) \geq 0{,}99$, which implies $e^{-20\mathrm{A}} \leq 0{,}01$.
- So, we need to survey at least approximately $0{,}23$ square degrees of the sky to have a $99\,\%$ chance of finding at least one quasar.

Gaussian distribution

- The Gaussian or "normal" distribution for a variable x with mean value μ and standard deviation σ is:

$$\boxed{P_{\text{Gaussian}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

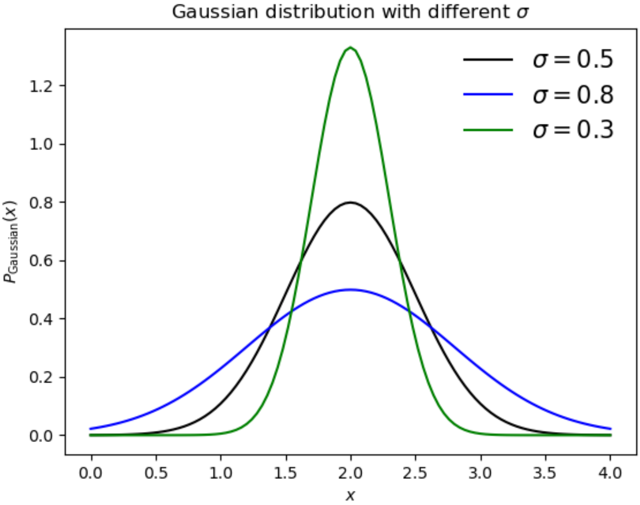
Gaussian distribution

- The Gaussian or "normal" distribution for a variable x with mean value μ and standard deviation σ is:

$$\boxed{P_{\mathrm{Gaussian}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

- In the limit of sufficiently large N, both the Binomial and Poisson distributions approach the Gaussian distribution.
- The central limit theorem says that if we average together variables drawn many times from any probability distribution, the resulting average will follow a Gaussian distribution, regardless of the original distribution of the variables.
- Examples of phenomena that follow a Gaussian distribution: measurement errors, heights of people, IQ scores.

Gaussian distribution



Central Limit Theorem

Given a set of samples $X_1, ..., X_n$ of equal size:

- The distribution of the sample means $\bar{X}$ approaches a normal distribution as the sample size increases.
- This normal distribution will have a mean value close to the population mean (μ).
- The variance of the sample mean (σ^2) will be close to the population variance divided by the number of samples ($\frac{S^2}{N}$).
- A sufficiently large sample size allows for accurate prediction of the characteristics of a population.
- Watch **this video** and **this other video** for more details on the Central Limit Theorem.

Outline

Confidence intervals

- Given a random variable that follows a Gaussian distribution, what is the range of values that contains 68 % probability?

Confidence intervals

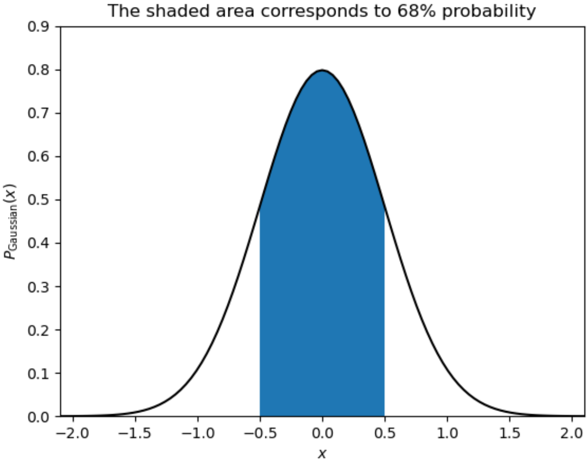
- Given a random variable that follows a Gaussian distribution, what is the range of values that contains 68 % probability?
- This question is equivalent to asking: What is the value of a such that:

$$\int_{\mu-a}^{\mu+a} P(x)dx = \int_{\mu-a}^{\mu+a} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 0{,}68$$

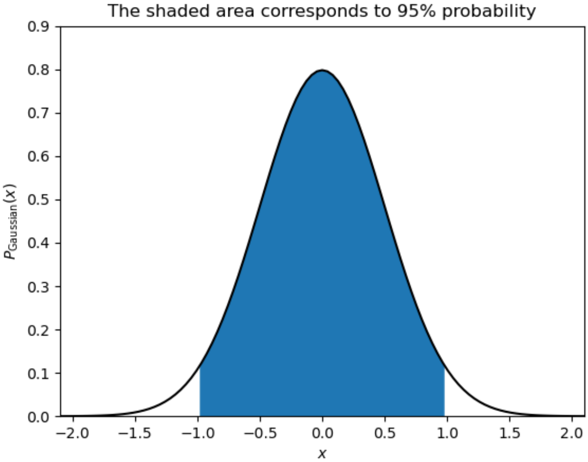
- Watch **this video** to learn more about the integral of the gaussian function.

Gaussian distribution

- The answer is $a = \sigma$ for 0.68, $a = 1{,}96\sigma$ for 0.95, and $a = 3\sigma$ for 0.99.

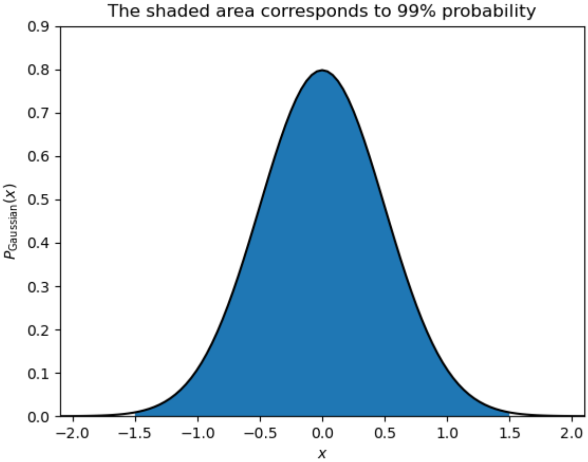


- The answer is $a = \sigma$ for 0.68, $a = 1{,}96\sigma$ for 0.95, and $a = 3\sigma$ for 0.99.



Gaussian distribution

- The answer is $a = \sigma$ for 0.68, $a = 1{,}96\sigma$ for 0.95, and $a = 3\sigma$ for 0.99.



Confidence intervals (frequentist approach)

- Given a random variable X that follows a Gaussian distribution, charachterized by mean value μ and standard deviation σ, the meaning of the confidence intervals is as follows:
  - The 68 % confidence interval $(\mu - \sigma, \mu + \sigma)$ has the following meaning:

    If you measure the variable X one time, the probability that the measured value falls within the interval $(\mu - \sigma, \mu + \sigma)$ is 68 %.

  - The 68 % confidence interval $(\mu - \frac{\sigma}{\sqrt{N}}, \mu + \frac{\sigma}{\sqrt{N}})$ has the following meaning:

    If you measure N times the variable X, the probability that the mean value of the measurements falls within the interval $(\mu - \frac{\sigma}{\sqrt{N}}, \mu + \frac{\sigma}{\sqrt{N}})$ is 68 %.

The frequentist approach does not assing probabilities to the true value of the variable X.

It only assigns probabilities to the measured values.

Confidence intervals (bayesian approach)

- Given a random variable X that follows a Gaussian distribution, charachterized by mean value μ and standard deviation σ, the meaning of the confidence intervals is as follows:
  - The 68 % confidence interval $(\mu - \sigma, \mu + \sigma)$ has the following meaning:

    The probability that the true value of X lies in the interval $(\mu - \sigma, \mu + \sigma)$ is 68 %.

  - The 68 % confidence interval $(\mu - \frac{\sigma}{\sqrt{N}}, \mu + \frac{\sigma}{\sqrt{N}})$ has the following meaning:

    If you measure N times the variable X, the probability that the the true mean value of X lies within the interval $(\mu - \frac{\sigma}{\sqrt{N}}, \mu + \frac{\sigma}{\sqrt{N}})$ is 68 %.

The bayesian approach assing probabilities to the true value of the variable X.

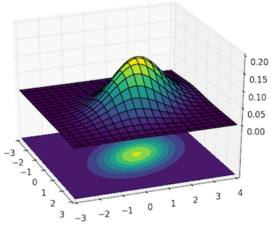
- Given a vector $\mathbf{X} = (X_1, X_2, ..., X_n)$ that follows a multivariate Gaussian distribution characterized by a mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$ and a covariance matrix $\boldsymbol{\Sigma}$, the probability density function is given by:

$$P(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})\right]$$

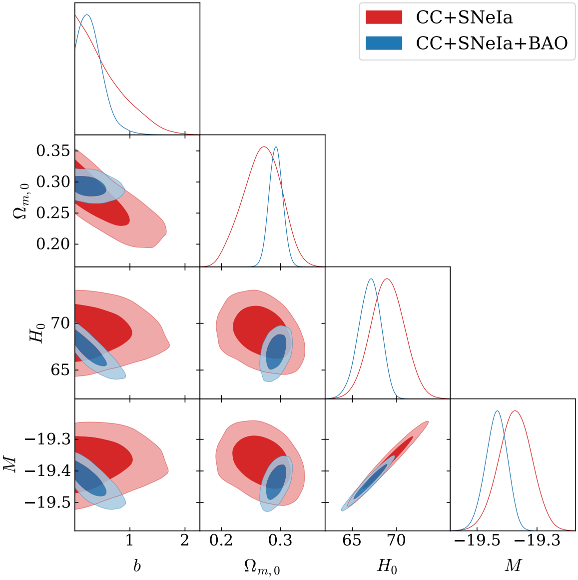
- the covariance matrix Σ contains the variances of each variable along the diagonal and the covariances between pairs of variables in the off-diagonal elements.

Confidence regions

- Given a vector $\mathbf{X} = (x, y)$ that follows a bivariate Gaussian distribution
$$P(\mathbf{X}) = \frac{1}{(2\pi)|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X}-\boldsymbol{\mu})\right]$$
where $|\Sigma|$ is the determinant of the covariance matrix.
- What are the values of a and b such that $\int_{-a}^{a}\int_{-b}^{b} P(x,y)dxdy = 0{,}68$?



Triangular plots



Outline

Frequentist approach vs Bayesian approach

- In the frequentist approach, a hypothesis (theoretical model) is assumed, and the probability of obtaining a given data set under that hypothesis is determined: $\longrightarrow P(\mathrm{Data}|\mathrm{Model})$
- The key question in the frequentist approach is whether the observed data set is reasonable under the assumed hypothesis (theoretical model).
- Philosophy: Reject a hypothesis (theoretical model) if $P(\mathrm{Data}|\mathrm{Model})$ is very small.
- The frequentist approach does not incorporate external information (prior).

Frequentist approach vs Bayesian approach

- In the Bayesian approach, given a data set, we determine the probability that these data can be described by a particular theoretical model: $\longrightarrow P(\mathrm{Model}|\mathrm{Data})$
- The key question in the Bayesian approach is: what is the probability that a given model describes the data under analysis? If the model is parametric, the question becomes: what is the probability of the parameter values of the model, given the observed data?
- External information (the prior) is incorporated. The probabilities obtained depend on the choice of priors.
- Philosophy: Models are not rejected; instead, probabilities are assigned to them. These probabilities represent our degree of knowledge about the behavior of nature.

Bayes theorem

- The concept of conditional probability is central to understanding Bayesian statistics.
- For example, the probability of getting a discount on a purchase is $P(D) \sim 0{,}2$. However, the probability of getting a discount if the purchase is made on a Friday is $P(D|\text{Friday}) = 1$.

Bayes theorem

- The concept of conditional probability is central to understanding Bayesian statistics.
- For example, the probability of getting a discount on a purchase is $P(D) \sim 0{,}2$. However, the probability of getting a discount if the purchase is made on a Friday is $P(D|\text{Friday}) = 1$.
- From the relationship $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, we can deduce the following result.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Bayes theorem

- The concept of conditional probability is central to understanding Bayesian statistics.
- For example, the probability of getting a discount on a purchase is $P(D) \sim 0{,}2$. However, the probability of getting a discount if the purchase is made on a Friday is $P(D|\text{Friday}) = 1$.
- From the relationship $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$, we can deduce the following result.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Bayes theorem

$$P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model})P(\text{Model})}{P(\text{Data})}$$

Worked Example (frequentist approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?

Worked Example (frequentist approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Estimate AGN fraction$= \dfrac{N_{\mathrm{AGN}}}{N_{\mathrm{total}}} = \dfrac{30}{100} = 0{,}3.$

Worked Example (frequentist approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Estimate AGN fraction$= \dfrac{N_{\text{AGN}}}{N_{\text{total}}} = \dfrac{30}{100} = 0{,}3$.
- There are 2 possible outcomes ("AGN" or "not an AGN") so the binomial distribution applies.

Worked Example (frequentist approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Estimate AGN fraction= $\dfrac{N_{\rm AGN}}{N_{\rm total}} = \dfrac{30}{100} = 0{,}3.$
- There are 2 possible outcomes ("AGN" or "not an AGN") so the binomial distribution applies.
- We can estimate the error using the variance of the binomial distribution for n success from N trials: $\sigma^2(n) = p(1-p)N$, noting that the variance of obtaining the fraction of successes $\dfrac{n}{N}$ in N trials is given by:

$$\frac{\sigma^2(n)}{N^2} = \frac{p(1-p)}{N}$$

Worked Example (frequentist approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Estimate AGN fraction= $\dfrac{N_{\text{AGN}}}{N_{\text{total}}} = \dfrac{30}{100} = 0{,}3$.
- There are 2 possible outcomes ("AGN" or "not an AGN") so the binomial distribution applies.
- We can estimate the error using the variance of the binomial distribution for n success from N trials: $\sigma^2(n) = p(1-p)N$, noting that the variance of obtaining the fraction of successes $\dfrac{n}{N}$ in N trials is given by:

$$\frac{\sigma^2(n)}{N^2} = \frac{p(1-p)}{N}$$

- The best estimate of the AGN fraction is $0{,}3 \pm 0{,}046$.

Worked Example (bayesian approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?

Worked Example (bayesian approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Use Bayes' theorem $P(\mathrm{p}|\mathrm{D}) \sim P(\mathrm{D}|\mathrm{p})P(\mathrm{p})$
- $P(\mathrm{p}|\mathrm{D})$ is the probability distribution of the random variable p (AGN fraction) given the data D, the quantity we aim to determine

Worked Example (bayesian approach)

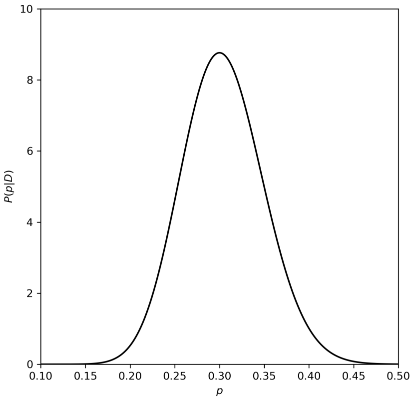
- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?
- Use Bayes' theorem $P(\mathrm{p}|\mathrm{D}) \sim P(\mathrm{D}|\mathrm{p})P(\mathrm{p})$
- $P(\mathrm{p}|\mathrm{D})$ is the probability distribution of the random variable p (AGN fraction) given the data D, the quantity we aim to determine
- $P(\mathrm{D}|\mathrm{p})$ is the probability of the data for a given value of p, which is given by the Binomial distribution as

$$\begin{aligned} P(\mathrm{D}|\mathrm{p}) &= P_{\mathrm{BINOM}}(n, N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\ &= \frac{100!}{(30)!(100-30)!} p^{30} (1-p)^{70} \end{aligned}$$

- $P(\mathrm{p})$ is the prior probability distribution of the AGN fraction, which we can assume to be uniform in the range $[0, 1]$.

Worked Example (bayesian approach)

- I observe 100 galaxies, 30 of which are AGN. What is the best estimate of the AGN fraction and its error?



Another example: Bayesian vs Frequentist approach

- A person flips a coin several times and finds that it always lands on heads. They then wonder: is this a fair coin or does it have two heads?

Another example: Bayesian vs Frequentist approach

- A person flips a coin several times and finds that it always lands on heads. They then wonder: is this a fair coin or does it have two heads?
- In the frequentist approach: we assume the coin is fair and perform 5 tosses to test the null hypothesis. If all 5 tosses result in heads, the question is: what is the probability of obtaining such an extreme result under the null hypothesis?
- In the Bayesian approach, we start with a prior belief about the probability that the coin is fair. After performing the 5 tosses, we update our probabilities taking into account both the observed results and the prior information.

Propositions and Probabilities

- In the frequentist approach $p(A)$ is the relative frequency with which A occurs in identical repeats of an experiment.
- In the Bayesian approach, $p(A|B)$ is a real number measure of the plausibility of a hypothesis A, given the truth of the information represented by proposition B.
- The basic rules for Bayesian probabilities are:
  - The sum rule:
    $$p(A|B) + p(\bar{A}|B) = 1$$
    where $\bar{A}$ means that the alternative proposition to A is true.
  - The product rule:
    $$\begin{aligned} p(A,B|C) &= p(A|C)p(B|A,C) \\ &= p(B|C)p(A|B,C) \end{aligned}$$

- Bayes' theorem follows directly from the product rule:

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

Bayes' Theorem

- Let $\{H_i\}$ denote a set of alternative hypotheses, D a set of data, and I the prior information.
- The plausibility of each alternative hypothesis, given the data, can be estimated by calculating the probability of that hypothesis conditioned on the data and prior information: $p(H_i|D, I)$.
- From the product rule, we can derive Bayes' theorem:

$$p(H_i|D, I) = \frac{p(H_i|I)p(D|H_i, I)}{p(D|I)}$$

  - $p(H_i|I)$ is the probability of H_i, before we have obtained the data (D), and is called the **prior** probability.
  - $p(H_i|D, I)$ is the probability of H_i after obtaining the data (D), and is called the **posterior** probability.
  - $p(D|H_i, I)$ is called the likelihood of H_i.
  - $p(D|I)$ is called the global likelihood for all hypotheses.

Bayes' Theorem

- Let $\{H_i\}$ denote a set of alternative hypotheses, D a set of data, and I the prior information.
- The plausibility of each alternative hypothesis, given the data, can be estimated by calculating the probability of that hypothesis conditioned on the data and prior information: $p(H_i|D,I)$.
- From the product rule, we can derive Bayes' theorem:

$$p(H_i|D,I) = \frac{p(H_i|I)p(D|H_i,I)}{p(D|I)}$$

  - $p(H_i|I)$ is the probability of H_i, before we have obtained the data (D), and is called the **prior** probability.
  - $p(H_i|D,I)$ is the probability of H_i after obtaining the data (D), and is called the **posterior** probability.
  - $p(D|H_i,I)$ is called the likelihood of H_i.
  - $p(D|I)$ is called the global likelihood for all hypotheses.

Bayes' Theorem

- The following must hold:

$$\sum_i p(H_i|D,I) = 1$$

Bayes' Theorem

- The following must hold:

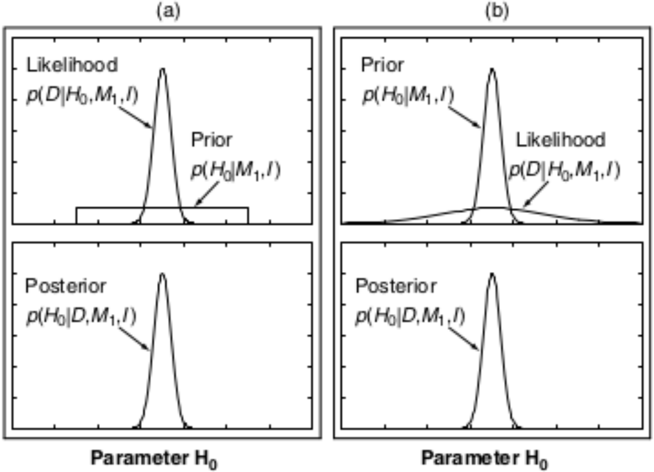
$$\begin{aligned}\sum_i p(H_i|D,I) &= 1 \\ &= \sum_i \frac{p(H_i|I)p(D|H_i,I)}{p(D|I)}\end{aligned}$$

- To satisfy this condition, $p(D|I)$ must be given by

$$p(D|I) = \sum_i p(H_i|I)p(D|H_i,I)$$

- In this way, $p(D|I)$, which does not depend on H_i, acts as a normalization constant.

- In the bayesian framework the result depends on the prior



Example

- Elementary particles in nature often originate from the decay of other particles. In a particle detector, a muon is observed. There are two hypotheses: 1) The muon originated from a pion; 2) The muon originated from a kaon. The decay probabilities inside the detector are known: $p(\mu|\pi) = 0{,}02$ and $p(\mu|K) = 0{,}10$. The ratio of pions to kaons in the original beam is $p(\pi) : p(K) = 3 : 1$. What is the probability that the observed muon originated from a pion or a kaon?

Example

- Elementary particles in nature often originate from the decay of other particles. In a particle detector, a muon is observed. There are two hypotheses: 1) The muon originated from a pion; 2) The muon originated from a kaon. The decay probabilities inside the detector are known: $p(\mu|\pi) = 0{,}02$ and $p(\mu|K) = 0{,}10$. The ratio of pions to kaons in the original beam is $p(\pi) : p(K) = 3 : 1$. What is the probability that the observed muon originated from a pion or a kaon?
  Applying Bayes' theorem,

$$p(\pi|\mu) = \frac{p(\pi)p(\mu|\pi)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)}$$

Example

- Elementary particles in nature often originate from the decay of other particles. In a particle detector, a muon is observed. There are two hypotheses: 1) The muon originated from a pion; 2) The muon originated from a kaon. The decay probabilities inside the detector are known: $p(\mu|\pi) = 0{,}02$ and $p(\mu|K) = 0{,}10$. The ratio of pions to kaons in the original beam is $p(\pi) : p(K) = 3 : 1$. What is the probability that the observed muon originated from a pion or a kaon?
  Applying Bayes' theorem,

$$\begin{aligned} p(\pi|\mu) &= \frac{p(\pi)p(\mu|\pi)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)} \\ &= \frac{3 \times 0{,}02}{3 \times 0{,}02 + 1 \times 0{,}1} \\ &= 0{,}375 \end{aligned}$$

Example

- Elementary particles in nature often originate from the decay of other particles. In a particle detector, a muon is observed. There are two hypotheses: 1) The muon originated from a pion; 2) The muon originated from a kaon. The decay probabilities inside the detector are known: $p(\mu|\pi) = 0{,}02$ and $p(\mu|K) = 0{,}10$. The ratio of pions to kaons in the original beam is $p(\pi) : p(K) = 3 : 1$. What is the probability that the observed muon originated from a pion or a kaon?
  Applying Bayes' theorem,

$$\begin{aligned} p(\pi|\mu) &= \frac{p(\pi)p(\mu|\pi)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)} \\ &= \frac{3 \times 0{,}02}{3 \times 0{,}02 + 1 \times 0{,}1} \\ &= 0{,}375 \end{aligned}$$

  Similarly, we can calculate $p(K|\mu)$:

$$p(K|\mu) = \frac{p(K)p(\mu|K)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)}$$

Example

- Elementary particles in nature often originate from the decay of other particles. In a particle detector, a muon is observed. There are two hypotheses: 1) The muon originated from a pion; 2) The muon originated from a kaon. The decay probabilities inside the detector are known: $p(\mu|\pi) = 0{,}02$ and $p(\mu|K) = 0{,}10$. The ratio of pions to kaons in the original beam is $p(\pi) : p(K) = 3 : 1$. What is the probability that the observed muon originated from a pion or a kaon?
  Applying Bayes' theorem,

$$\begin{aligned} p(\pi|\mu) &= \frac{p(\pi)p(\mu|\pi)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)} \\ &= \frac{3 \times 0{,}02}{3 \times 0{,}02 + 1 \times 0{,}1} \\ &= 0{,}375 \end{aligned}$$

  Similarly, we can calculate $p(K|\mu)$:

$$\begin{aligned} p(K|\mu) &= \frac{p(K)p(\mu|K)}{p(\pi)p(\mu|\pi) + p(K)p(\mu|K)} \\ &= \frac{1 \times 0{,}1}{3 \times 0{,}02 + 1 \times 0{,}1} \\ &= 0{,}625 \end{aligned}$$

Parameter Estimation

- In many cases, we address problems where a model (M) is assumed to be true, and the relevant hypothesis space is defined by the parameters of that model. In such situations, our interest shifts from probabilities to probability distributions (PDFs).
- For example, given prior information I about a model with a single parameter θ, $p(\theta|I)d\theta$ represents the prior probability that the parameter lies within the interval $\theta + \delta\theta$.
- We use the same symbol p for both probabilities and probability density functions.
- Bayes' theorem for PDFs can be written as

$$p(\theta|D, I) = \frac{p(\theta|I)p(D|\theta, I)}{p(D|I)}$$

with

$$p(D|I) = \int d\theta \, p(\theta|I)p(D|\theta, I)$$

where θ denotes the set of parameters of model M.

Example I

- The probability of a certain medical test being positive is 90 %, if a patient has disease D. 1 % of the population have the disease, and the test records a false positive 5 % of the time. If you receive a positive test, what is your probability of having D?

Example II

- A detector with finite resolution records the decay time t of a k meson. The time resolution corresponds to a Gaussian with variance σ^2. We are interested in the time θ at which the decay occurred. The mean lifetime τ of kaons is known. The probability density for the parameter θ before the measurement, the prior, is $p(\theta) = \frac{e^{-\frac{\theta}{\tau}}}{\tau}$, $\theta \geq 0$. The probability density for t with θ. fixed is the Gaussian.

Example II

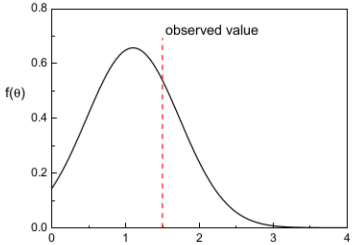
- A detector with finite resolution records the decay time t of a k meson. The time resolution corresponds to a Gaussian with variance σ^2. We are interested in the time θ at which the decay occurred. The mean lifetime τ of kaons is known. The probability density for the parameter θ before the measurement, the prior, is $p(\theta) = \frac{e^{-\frac{\theta}{\tau}}}{\tau}$, $\theta \geq 0$. The probability density for t with θ. fixed is the Gaussian.Applying Bayes' theorem, we obtain:

$$p(\theta|t) \quad = \quad \frac{p(\theta)p(t|\theta)}{p(t)} = \frac{p(\theta)p(t|\theta)}{\int_0^\infty p(\theta)p(t|\theta)d\theta}$$

Example II

- A detector with finite resolution records the decay time t of a k meson. The time resolution corresponds to a Gaussian with variance σ^2. We are interested in the time θ at which the decay occurred. The mean lifetime τ of kaons is known. The probability density for the parameter θ before the measurement, the prior, is $p(\theta) = \frac{e^{-\frac{\theta}{\tau}}}{\tau}$, $\theta \geq 0$. The probability density for t with θ. fixed is the Gaussian.Applying Bayes' theorem, we obtain:

$$
\begin{aligned}
p(\theta|t) &= \frac{p(\theta)p(t|\theta)}{p(t)} = \frac{p(\theta)p(t|\theta)}{\int_0^\infty p(\theta)p(t|\theta)d\theta} \\
&= \frac{\frac{e^{-\frac{\theta}{\tau}}}{\tau}\ e^{-\frac{(t-\theta)^2}{\sigma^2}}}{\int_0^\infty \frac{e^{-\frac{\theta}{\tau}}}{\tau}\ e^{-\frac{(t-\theta)^2}{\sigma^2}} d\theta}
\end{aligned}
$$



How to install the libraries that we will use in this course

- **To install Miniconda go to**
  https://www.anaconda.com/docs/getting-started/miniconda/install
  and follow the instructions for your operating system.
- **To Install pip3**
  sudo apt update
  sudo apt install python3-pip
- **Create a virtual environment using conda**
  conda create –name myenv python=yourpythonversion conda list
- **Install libraries**
  conda install -c conda-forge jupyterlab
  conda install -c conda-forge numpy scipy matplotlib
  conda install emcee h5py
- **Install libraries using pip3**
  pip3 install matplotlib numpy

How to install the libraries that we will use in this course

- **Combinations that work**
  - python 3.9.23 numpy==3.3.0 emcee==3.1.1 matplotlib==3.9.4 h5py==3.3.0
  - python 3.11.13 numpy==2.0.2 matplotlib==3.9.4 h5py==3.1.6 emcee==3.1.6

Suggested Lectures

- Bayesian Logical Data Analysis for the Physical Sciences Phil Gregory: Chapter 1