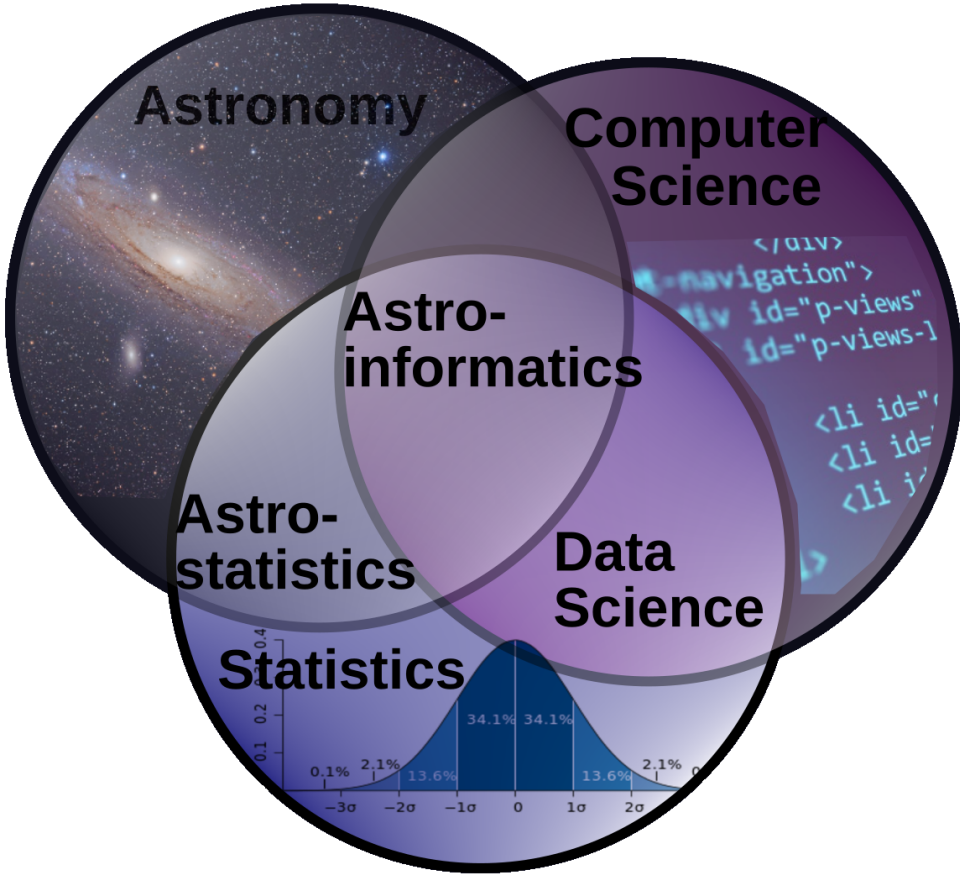
ASTR 3890 - Selected Topics: Data Science for Large Astronomical Surveys (Spring 2022)

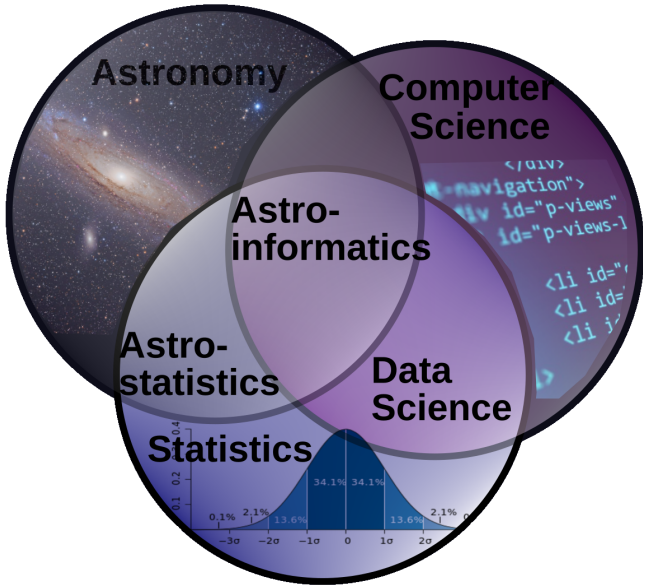
Introduction To Probability & Statistics: I

Dr. Nina Hernitschek
February 7, 2022

Astroinformatics, Astrostatistics



Astroinformatics, Astrostatistics



Astroinformatics = processing (extensive amounts of) astronomical data using computer science methods, including Astrostatistics

Astrostatistics = extracting knowledge from astronomical data

Knowledge = summary (physical or phenomenological) of data behavior

Data = result of measurements

Probability

$p(A) =$ the probability of A (or the probability density at A)

example: the probability that an observed object is a galaxy

Probability

$p(A)$ = the probability of A (or the probability density at A)

example: the probability that an observed object is a galaxy

The probability reflects our current state of knowledge of the object, and our belief that it is a galaxy.

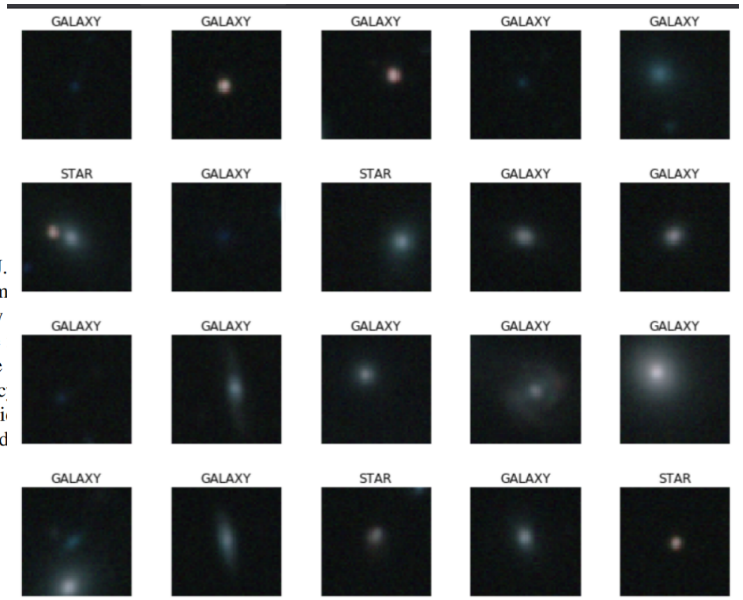
Deep Learning for Star-Galaxy Classification

Ganesh Ranganath Chandrasekar Iyer Krishna Chaithanya Vastare
University of California, San Diego
{grchandr, kvastare}@eng.ucsd.edu

Abstract

Conventional star-galaxy classifiers are based on the reduced summaries provided by the star-galaxy catalogs. However, these classifiers need careful feature selection and involvement of domain experts at various stages of classification. Thus, the current mechanism is not extremely scalable. It is important to develop a scalable probabilistic classifier

Recently, Edwardo Machado of CEFET/RJ.
lished a paper which encompassed and com
above mentioned algorithms for Star - Galaxy
Figure 1 shows the purity vs the magnitude
gorithms. It is evident that NN performs the
compares the algorithms based on the accurac
the ROC curve (AUC), Completeness galaxi
galaxies [8]. Again from his results it is evid



Notation

$A \cup B$ is the **union** of sets A and B. Read as A OR B.

Notation

$A \cup B$ is the **union** of sets A and B. Read as A OR B.

$A \cap B$ is the **intersection** of sets A and B. Read as A AND B.
Different notations $p(A \cap B) = p(A, B)$

We will use the comma notation throughout.

Kolmogorov Axioms

first axiom:

The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

Kolmogorov Axioms

first axiom:
The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

second axiom:
This is the assumption of unit measure: The probability that at least one of the elementary events in the entire sample space will occur is 1.
$p(\Omega) = 1$ where Ω is the set of all possible outcomes, i.e. the sum/ integral of all possible outcomes is 1.

Kolmogorov Axioms

first axiom:
The probability of an event is a non-negative real number:

$$p(A) \geq 0 \; \forall A$$

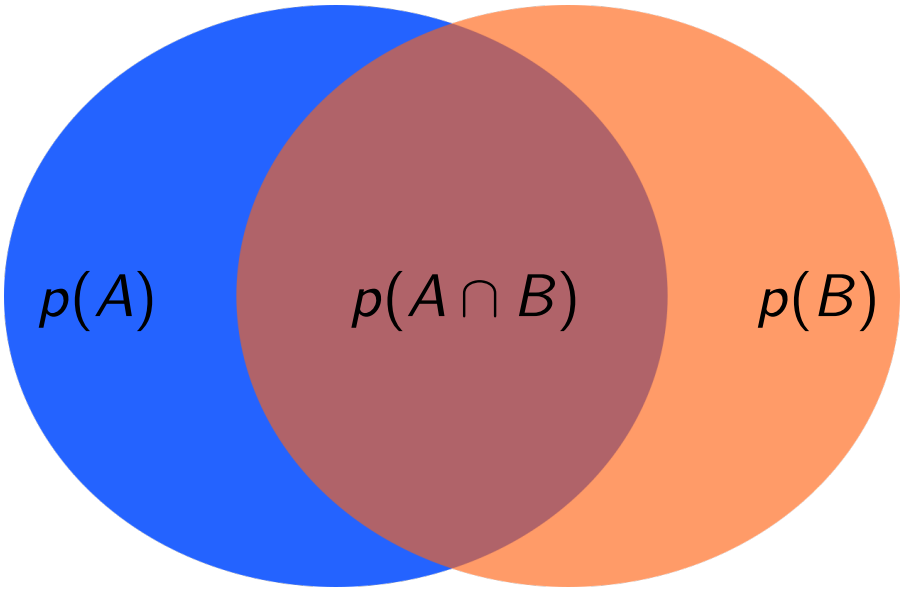
second axiom:
This is the assumption of unit measure: The probability that at least one of the elementary events in the entire sample space will occur is 1.
$p(\Omega) = 1$ where Ω is the set of all possible outcomes, i.e. the sum/ integral of all possible outcomes is 1.

third axiom:
Any countable sequence of disjoint sets (synonymous with mutually exclusive events) $A_1, A_2, \ldots$ satisfies

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$$

Consequences of Kolmogorov Axioms

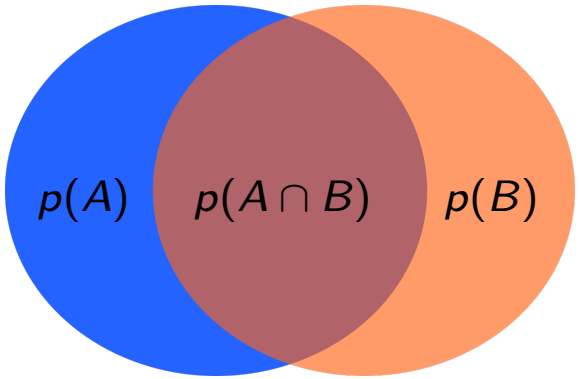
If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

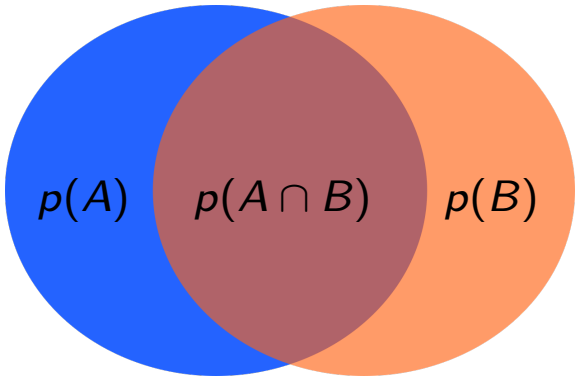
The probability that either A or B will happen (which could include both) is the union, given by

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

The term $-p(A \cap B)$ is necessary as A and B overlap, thus we would count it twice.

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

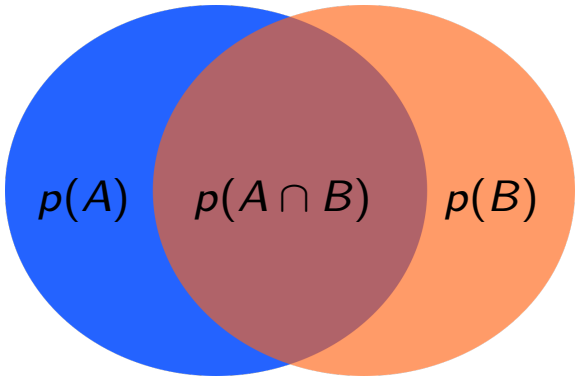
The probability that both A and B will happen, $p(A \cap B)$, is

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

where $p(A|B)$ is the probability of A given that B is true and is called the **conditional probability** (so the $|$ is short for "given").

Consequences of Kolmogorov Axioms

If we have two events, A and B, the possible combinations are illustrated by the following figure:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

The law of the total probability says that (for independent A_i, B_i),

$$p(A) = \sum_i p(A|B_i)p(B_i)$$

Consequences of Kolmogorov Axioms

It is important to realize that the following is always true:
$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$
However, if A and B are independent, then $p(A|B) = p(A)$ and $p(B|A) = p(B)$ and $p(A, B) = p(A)p(B)$.

Consequences of Kolmogorov Axioms

It is important to realize that the following is always true:
$p(A, B) = p(A|B)p(B) = p(B|A)p(A)$
However, if A and B are independent, then $p(A|B) = p(A)$ and $p(B|A) = p(B)$ and $p(A, B) = p(A)p(B)$.

example: classic marbles in bag scenario

If you have a bag with 5 marbles (3 yellow and 2 blue) and you want to know the probability of picking 2 yellow marbles in a row, that would be $p(Y_1, Y_2) = p(Y_1)p(Y_2|Y_1)$ with $p(Y_1) = \frac{3}{5}$ since you have an equally likely chance of drawing any of the 5 marbles.

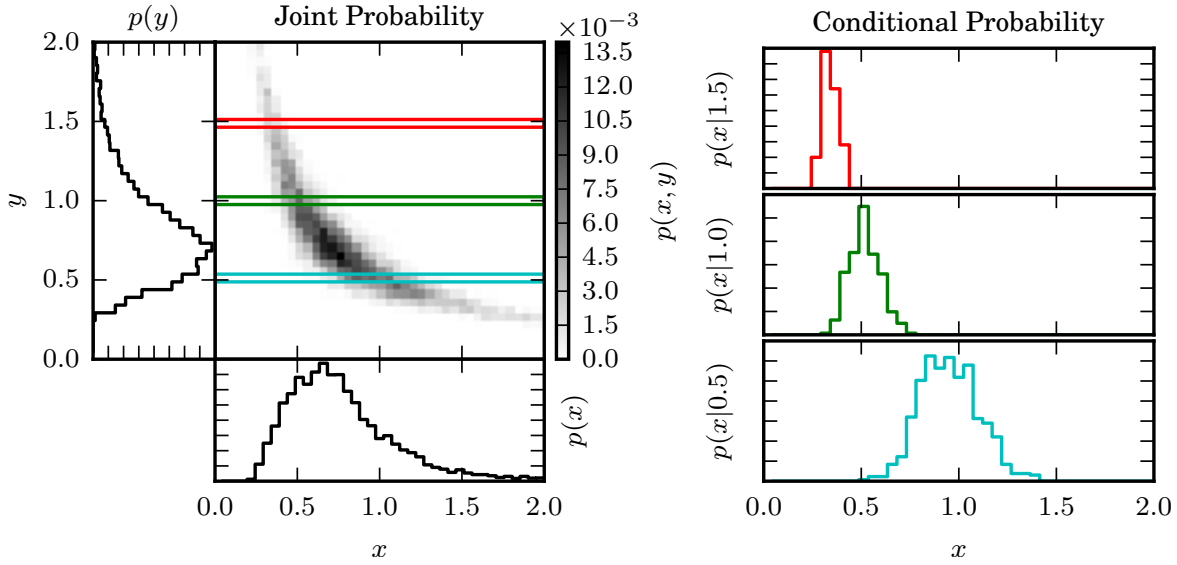
If you did not put the first marble back in the bag after drawing it (**sampling without replacement**), then the probability is $p(Y_2|Y_1) = \frac{2}{4}$, so that $p(Y_1, Y_2) = \frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$.

But if you put the first marble back (**sampling with replacement**), then $p(Y_2|Y_1) = \frac{3}{5} = p(Y_2)$, so that $p(Y_1, Y_2) = \frac{3}{5} \times \frac{3}{5} = \frac{9}{25}$.

In the first case Y_1 and Y_2 are not independend, but in the second they are.

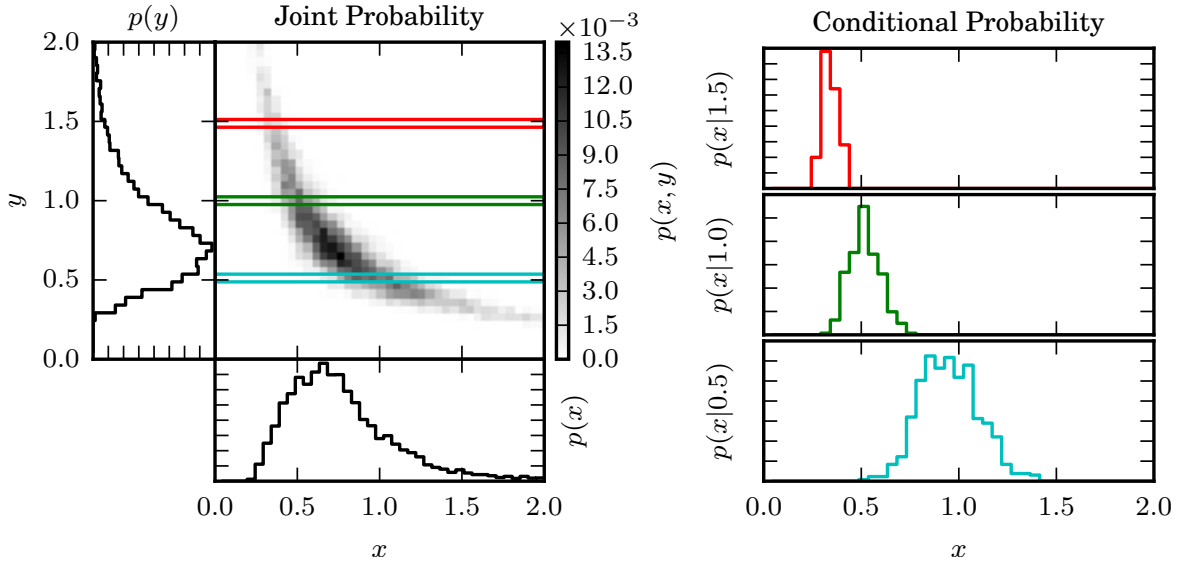
Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.



Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.



We have

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

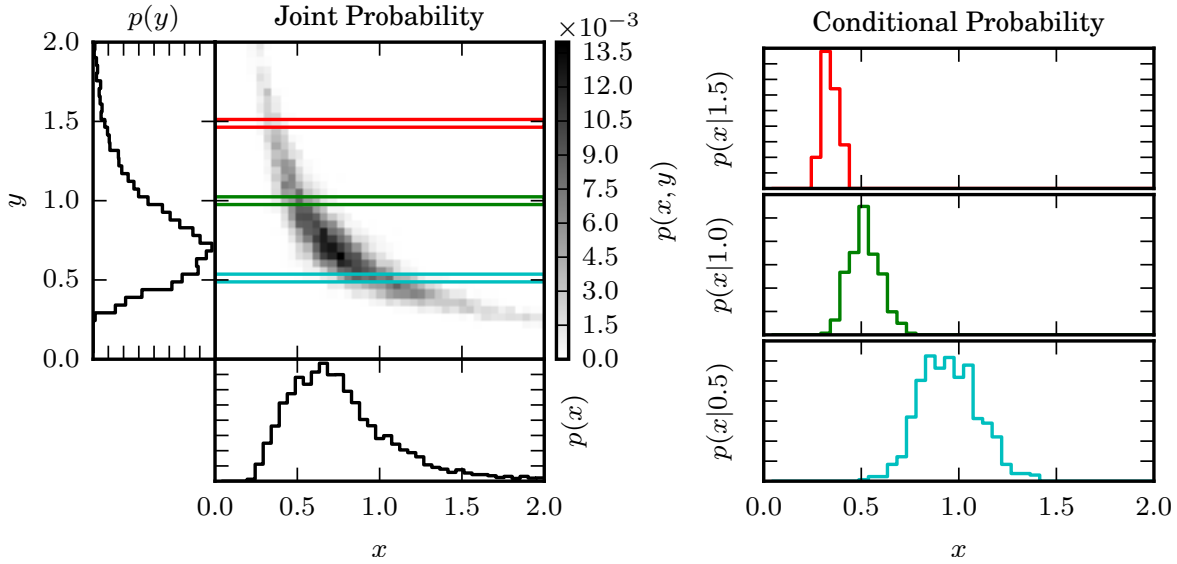
We can define the **marginal probability** as

$$p(x) = \int p(x, y)dy$$

where marginal means projecting onto one axis (integrating over the unwanted variable). The marginal distributions are shown on the left and bottom sides of the left panel.

Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.

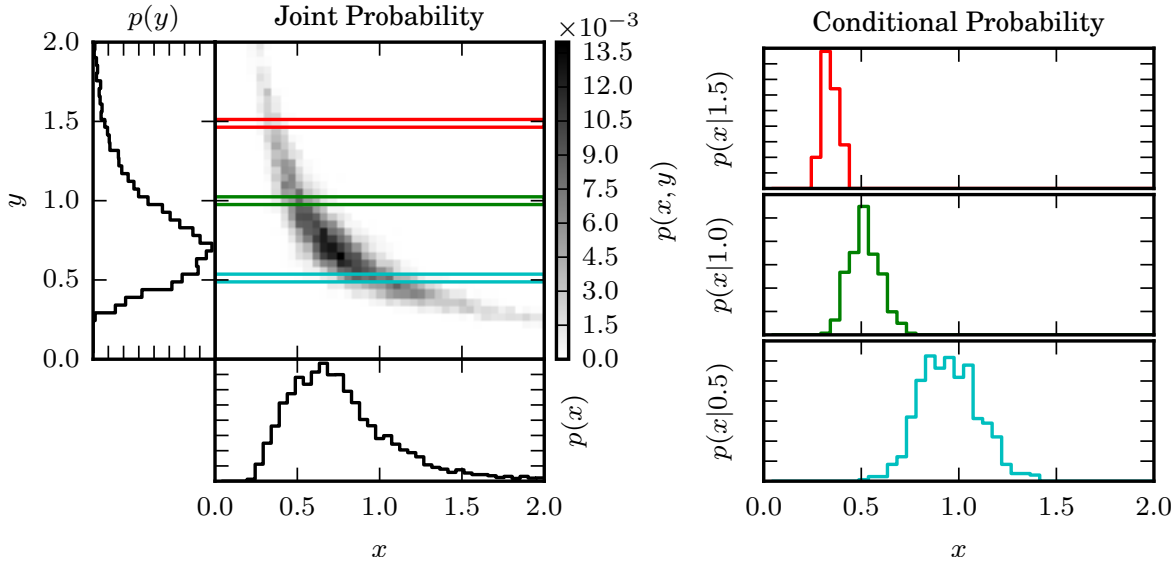


The three panels on the right show the **conditional probability** (of x) for three values: $p(x|y = y_0)$
These are just normalized "slices" through the 2-D distribution.
The marginal probability of x can be re-written as

$$p(x) = \int p(x|y)p(y)dy$$

Bayes' Theorem

In this 2-D distribution in $x - y$ parameter space, x and y are not independent as, once you pick a y, your values of x are constrained.



But since $p(x|y)p(y) = p(y|x)p(x)$, we can write

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\int p(x|y)p(y)dy}$$

which in words says that

the (conditional) probability of y given x is the (conditional) probability of x given y times the (marginal) probability of y divided by the (marginal) probability of x, where the latter is just the integral of the numerator.

Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")

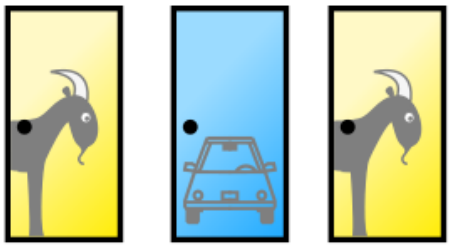
You are playing a game show and are shown 2 doors. One has a car behind it, the other two a goat. What are your chances of picking the door with the car?

Bayes' Theorem

Example: Monty Hall Problem (or "Deal Or No Deal")

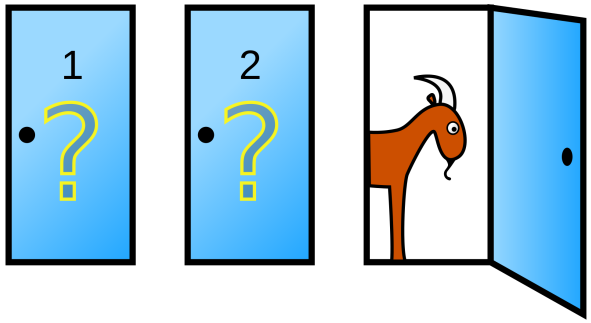
You are playing a game show and are shown 2 doors. One has a car behind it, the other two a goat. What are your chances of picking the door with the car?

You are playing a game show and are shown 3 doors. One has a car behind it, the other two a goat. What are your chances of picking the door with the car?

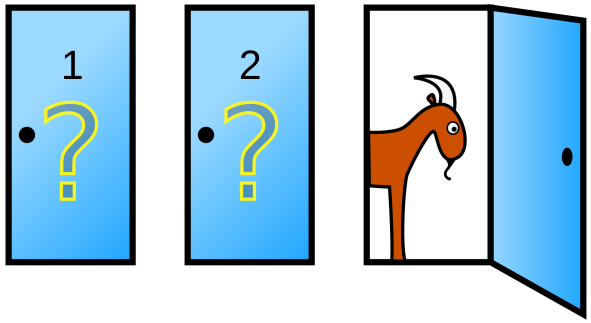


Bayes' Theorem

You are playing a game show and are shown 3 doors. The game show host asks you to pick a door, but not to open it yet. Then the host opens one of the other two doors (that you did not pick), making sure to select one with a goat. The host offers you the opportunity to **switch** doors. Do you?



Bayes' Theorem

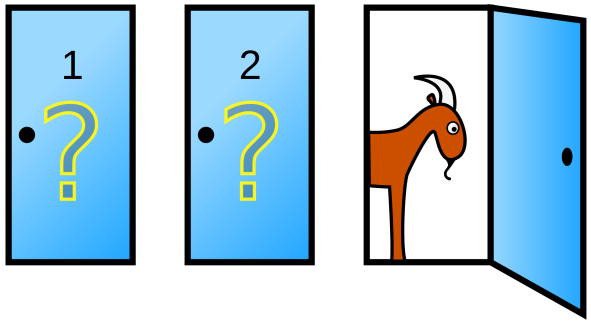


What we know about the **probabilities**:

Probability of car behind Door 1 $=1/3$

Probability of car behind Doors 2 or 3 $=2/3$

Bayes' Theorem



What we know about the **probabilities**:

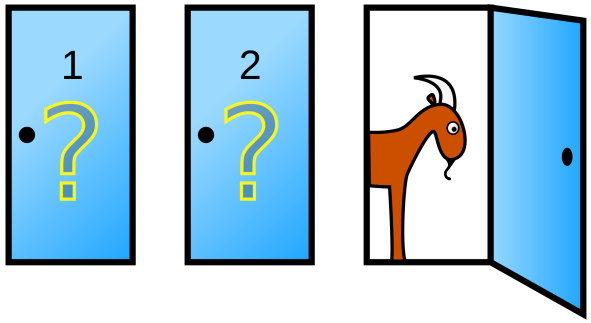
Probability of car behind Door 1 $=1/3$
Probability of car behind Doors 2 or 3 $=2/3$

Probability of you had picked the car 1 $=1/3$
$\Rightarrow$ host can open either of the other doors

Probability of you had picked a goat (one of the two goats) $=2/3$

$\Rightarrow$ host opens door that is also a goat, remaining door has the car

Bayes' Theorem



What we know about the **probabilities**:

Probability of car behind Door 1 $=1/3$
Probability of car behind Doors 2 or 3 $=2/3$

Probability of you had picked the car 1 $=1/3$
$\Rightarrow$ host can open either of the other doors

Probability of you had picked a goat (one of the two goats) $=2/3$
$\Rightarrow$ host opens door that is also a goat, remaining door has the car

The advice is to switch to the door you hadn't chosen and that's not open yet - precisely, switching doubles your chances.

This is an example of the use of **conditional probability**, where we have $p(A|B) \neq p(A)$.

Notation

In the textbook and here:

x is a scalar quantity that is measured N times to form a dataset

x_i is a single measurement with $i = 1, ..., N$

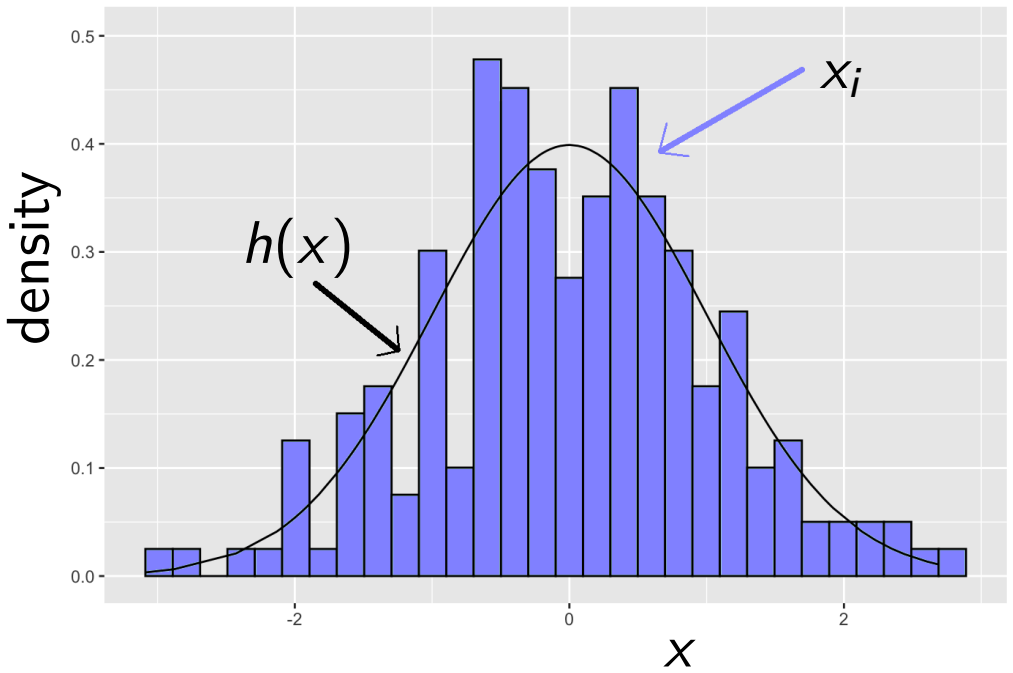
x_i refers to the set of all N measurements comprising the dataset

measurements (data) can be real numbers, discrete labels (strings or numbers), or even "missing values" (we sometimes pad our datasets with `NaN` in this case)

Goal of Statistical Inference

idea:

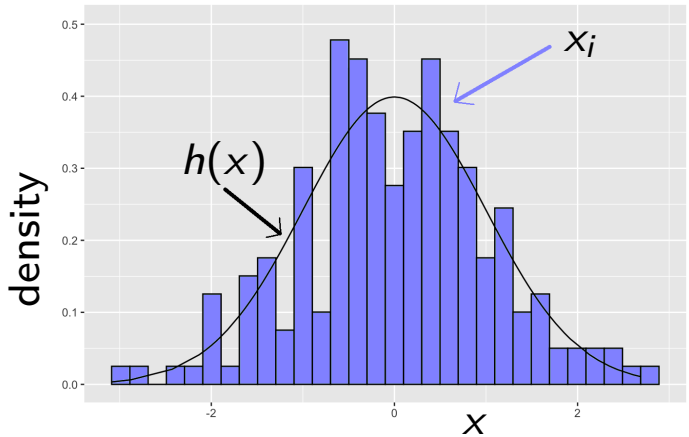
- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf



Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf

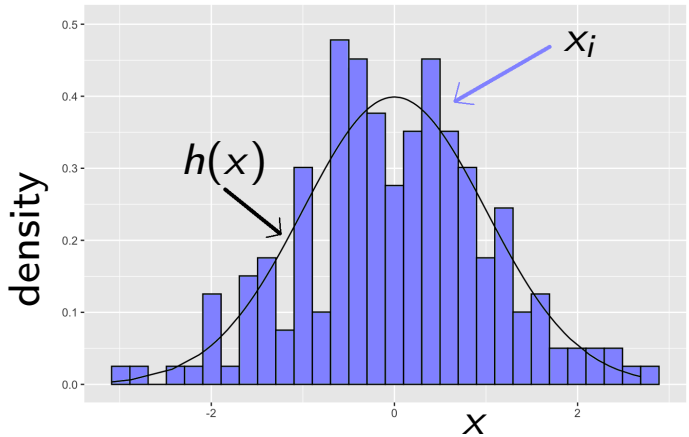


using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn

Goal of Statistical Inference

idea:

- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf

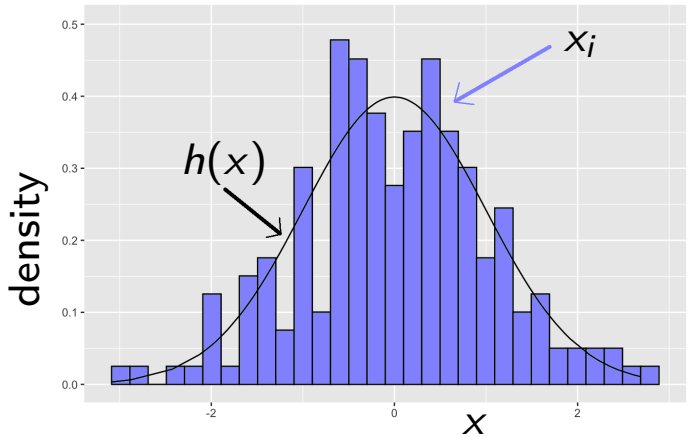


Question: What is the probability of a value lying between x and $x + dx$, where dx is infinitesimal?

Goal of Statistical Inference

idea:

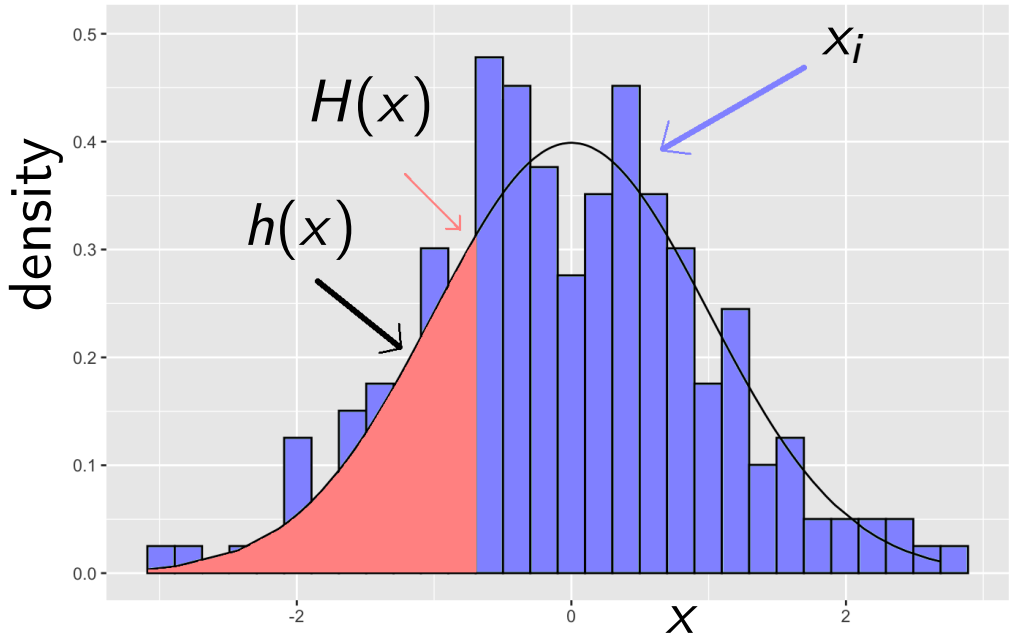
- measurements are drawn from an underlying probability distribution function (pdf) $h(x)$
- we can only observe the measurements x_i, not the underlying pdf



Question: What is the probability of a value lying between x and $x + dx$, where dx is infinitesimal?
Answer: $h(x)dx$

Probability Distribution Function

The "left to right" integral of $h(x)$ is the **cumulative distribution function** (cdf), $H(x) = \int_{-\infty}^{x} h(x')dx'$.



The inverse function of the cdf is the **quantile function**, answering the question: Which x value has e.g. 90% of the distribution below it?

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

While $h(x)$ is the underlying pdf (also called *population pdf*), what we measure from the data is the **empirical pdf** $f(x)$.

Empirical Distribution Function

Don't neglect measurement errors:

Using measurements x_i, we are trying to estimate the probability density (distribution) function or the *pdf* $h(x)$ from which the individual x_i are drawn.

While $h(x)$ is the underlying pdf (also called *population pdf*), what we measure from the data is the **empirical pdf** $f(x)$.

So, $f(x)$ is a model of $h(x)$. In principle, with infinite data $f(x) \rightarrow h(x)$, but in reality the blurring effect of **measurement errors** keep this from being strictly true. Likewise, the empirical cdf is denoted $F(x)$.

Errors and Uncertainties

Technically, **errors** are systematic biases that we can not mitigate through collecting lots and lots of data.

Statistical uncertainties are the result of random measurement uncertainty.

But "error" will be used for both, and denoted as either statistical errors (error bars) or systematic errors (biases).

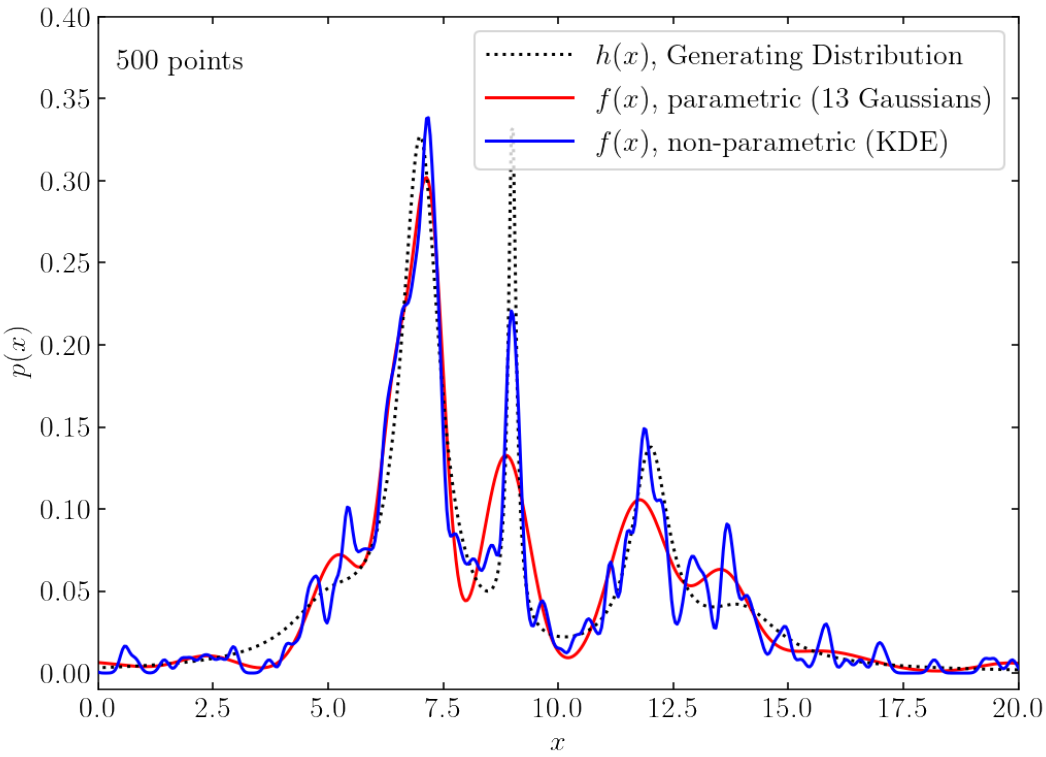
Statistical error distributions (error bars) that vary from data point to data point are called heteroscedastic errors (usually the case in astronomy). If they are the same for all points then they are homoscedastic errors.

Physical Models

idea: We can either:

- Describe the data, then the process is non-parametric, i.e. we are just trying to **describe** the data behavior in a compact practical way.
- Guess a physical model for $h(x)$, then the process is parametric.

  From a **model** we can generate new data that mimic measurements.

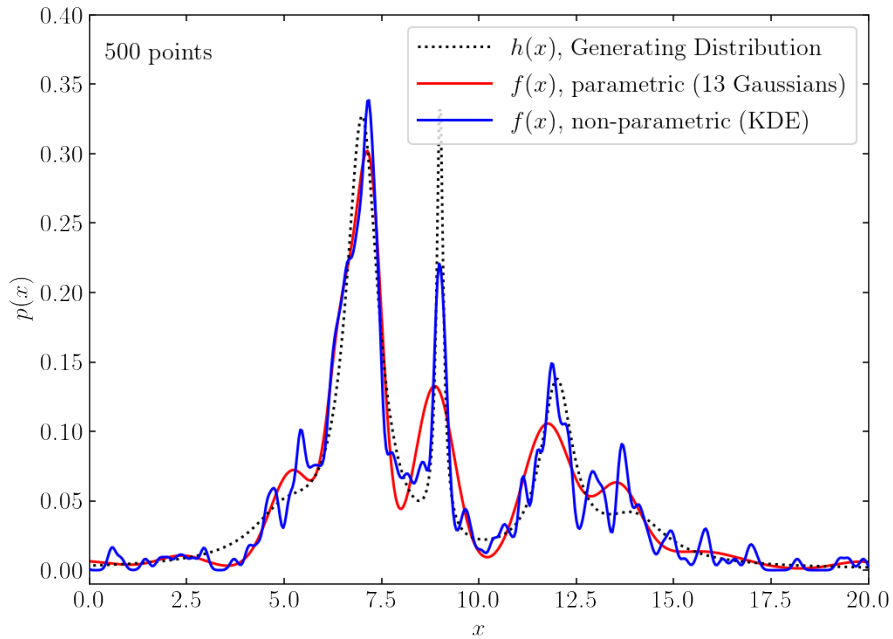


Physical Models

idea: We can either:

- Describe the data, then the process is non-parametric, i.e. we are just trying to **describe** the data behavior in a compact practical way.
- Guess a physical model for $h(x)$, then the process is parametric.

  From a **model** we can generate new data that mimic measurements.



$\Rightarrow$ We will see later on how to do that with Python!

Break & Questions

afterwards we continue with `lecture_3.ipynb` from the `github` repository