PROBABILISTIC INFERENCE AND LEARNING

LECTURE 02

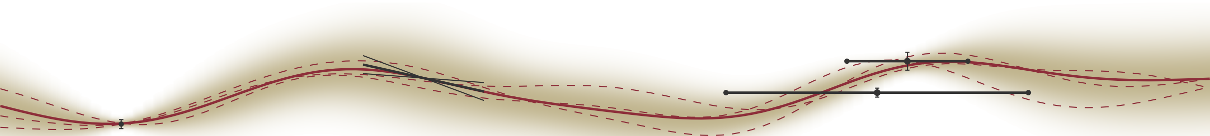
PROBABILITIES OVER CONTINUOUS VARIABLES

Philipp Hennig

22 October 2018

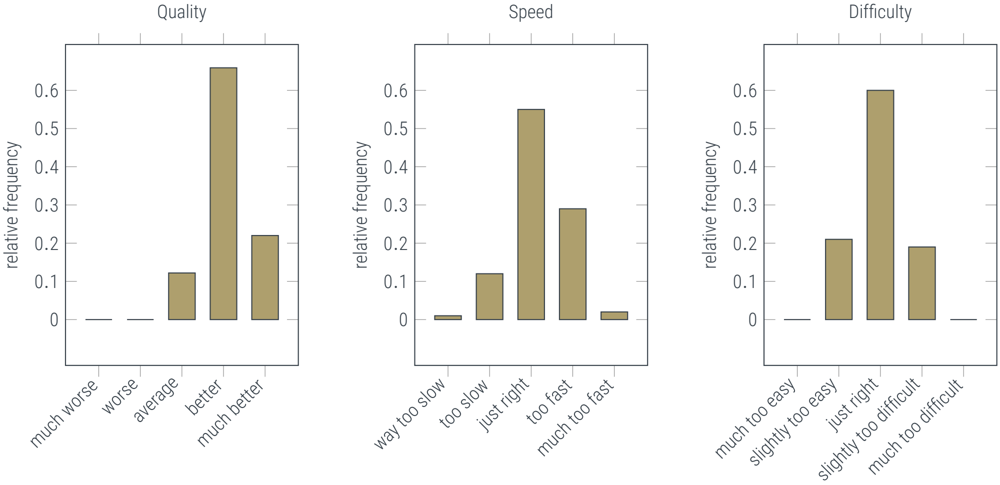
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF SCIENCE
DEPARTMENT OF COMPUTER SCIENCE
CHAIR FOR THE METHODS OF MACHINE LEARNING



Last Lecture: Debrief

Feedback dashboard



Last Lecture: Debrief

Detailed Feedback

Things you did not like:

- **no break!**
- **messy, illegible blackboard writing**
- going slowly through examples
- waiting for you to understand your notes
- “please move derivations to the slides”
- “please don’t compute actual numbers (computers should do that)”

Things you did not understand:

- derivations on the blackboard
- “What is A, B, C in $p(A, B \mid C)$, formally”?
- $A \perp\!\!\!\perp B \mid C$
- The earthquake example
- German quote at the end

Things you enjoyed:

- mix of blackboard, quotes, definitions, graphics
- (both) **examples**
- going slowly through the examples
- summary of last lecture
- **the quote at the end**
- formal definitions
- atomic independence structures
- downsides of DAGs
- great timing

Overview of Lectures so far:

0. Introduction to Reasoning under Uncertainty
    - Probabilities are the mathematical formalization of uncertainty
    - Two basic (sum & product) rules, and their Corollary, **Bayes' Theorem**, provide mathematical framework for inference.
1. Probabilistic Reasoning
    - Probabilities extend deductive reasoning to plausible reasoning
    - in multivariate distributions, (conditional) independence structure is crucial to control computational complexity

---

Today:

- Generalizing to distributions $p(x)$ over continuous variables $x \in \mathbb{R}$ requires some careful considerations.

Probabilities over Discrete Variables

Connection to Last Lecture

Probability theory as an extension of propositional logic

- finite set of propositional variables $A, B, \dots, Z \in \{0, 1\}$ jointly ranging over all boolean assignments
- sample space $\Omega = \{\text{all boolean assignments}\}$
- probability measure $p : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} p(\omega) = 1$

Discrete probability theory (includes the previous case)

- random variable ranging in a discrete set, e.g. $\{0, 1, 2, \dots, \}$
- sample space $\Omega = \{0, 1, 2, \dots\}$
- *probability measure* $p : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} p(\omega) = 1$

Probabilities over Discrete Variables

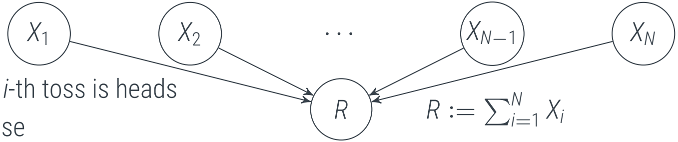
Example: Binomial distribution

A bent coin has probability *f* of coming up heads. The coin is tossed *N* times. What is the probability distribution of the number of heads *r*?

Probabilities over Discrete Variables

Example: Binomial distribution

A bent coin has probability *f* of coming up heads. The coin is tossed *N* times. What is the probability distribution of the number of heads *r*?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

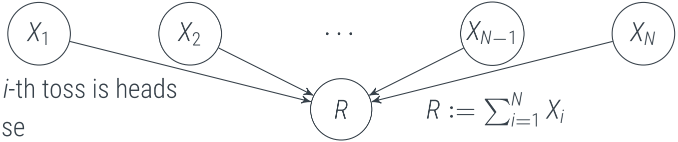
$$R := \sum_{i=1}^{N} X_i$$

$$p(R = r) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} \prod_{i=1}^{N} p(X_i = \text{face}_i(\omega)) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} f^r \cdot (1 - f)^{N-r} := p(r \mid f, N)$$

Probabilities over Discrete Variables

Example: Binomial distribution

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

$$R := \sum_{i=1}^{N} X_i$$

$$p(R=r) = \sum_{\omega \in \{\boldsymbol{X} \mid R=r\}} \prod_{i=1}^{N} p(X_i = \text{face}_i(\omega)) = \sum_{\omega \in \{\boldsymbol{X} \mid R=r\}} f^r \cdot (1-f)^{N-r} := p(r \mid f, N)$$

Note: In the remainder of the course, will often **abuse notation** as on the RHS above:

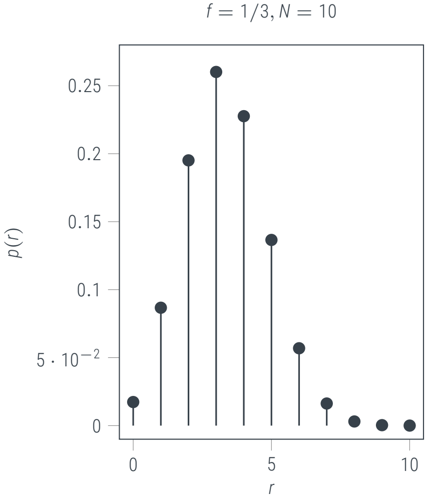
- $p(r)$ instead of $p(R=r)$ (recall again that $p(X) \neq p(Y)$!)
- $p(r \mid f, N)$ even though f, N are not variables in the graph (though they could be!)
  As a general rule, we will allow arbitrary **parameters** θ to be shown or dropped in the condition.
  This works because $p(x \mid \theta)$ is not i.g. a probability distribution of θ.

Probabilities over Discrete Variables

Example: Binomial distribution

A bent coin has probability *f* of coming up heads. The coin is tossed *N* times. What is the probability distribution of the number of heads *r*?

$$\begin{aligned} p(r \mid f, N) &= \#\text{ ways to choose } r \text{ from } N \cdot f^r \cdot (1-f)^{N-r} \\ &= \frac{N!}{(N-r)! \cdot r!} \cdot f^r \cdot (1-f)^{N-r} \\ &= \binom{N}{r} \cdot f^r \cdot (1-f)^{N-r} \end{aligned}$$



Probabilities over Discrete Variables

Example: Binomial distribution

$$p(r \mid f, N) = \binom{N}{r} \cdot f^r \cdot (1 - f)^{N-r}$$

Definition (expectation)

Let variable X take value $X = x \in \Omega$ with probability $p(X = x) =: p(x)$. The **expected value** of the function $a(x)$ is

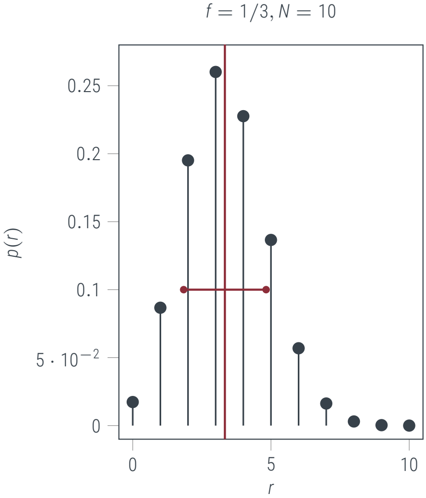
$$\mathbb{E}_p(a) := \sum_{x \in \Omega} a(x)p(x).$$

In particular, we have

mean of x $\mathbb{E}_p(x)$

variance of x $\mathbb{E}_p((x - \mathbb{E}_p(x))^2)$

Expected values are properties of p. They need not lie in the domain of X.



From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be a variable taking **real values**, $X \in \mathbb{R}$

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be a variable taking **real values**, $X \in \mathbb{R}$
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be a variable taking **real values**, $X \in \mathbb{R}$
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$
- note that $E_B = E_{A \vee W}$. Also, A and W are mutually exclusive, thus sum rule:

$$p(B) = p(A) + p(W) \qquad\qquad p(W) = p(B) - p(A)$$

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be a variable taking **real values**, $X \in \mathbb{R}$
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$
- note that $E_B = E_{A \vee W}$. Also, A and W are mutually exclusive, thus sum rule:

$$p(B) = p(A) + p(W) \qquad\qquad p(W) = p(B) - p(A)$$

- with $F(x) := p(X \leq x)$ and $f(x) = \frac{d}{dx}F(x)$ we get

$$p(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be a variable taking **real values**, $X \in \mathbb{R}$
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$
- note that $E_B = E_{A \vee W}$. Also, A and W are mutually exclusive, thus sum rule:

$$p(B) = p(A) + p(W) \qquad\qquad p(W) = p(B) - p(A)$$

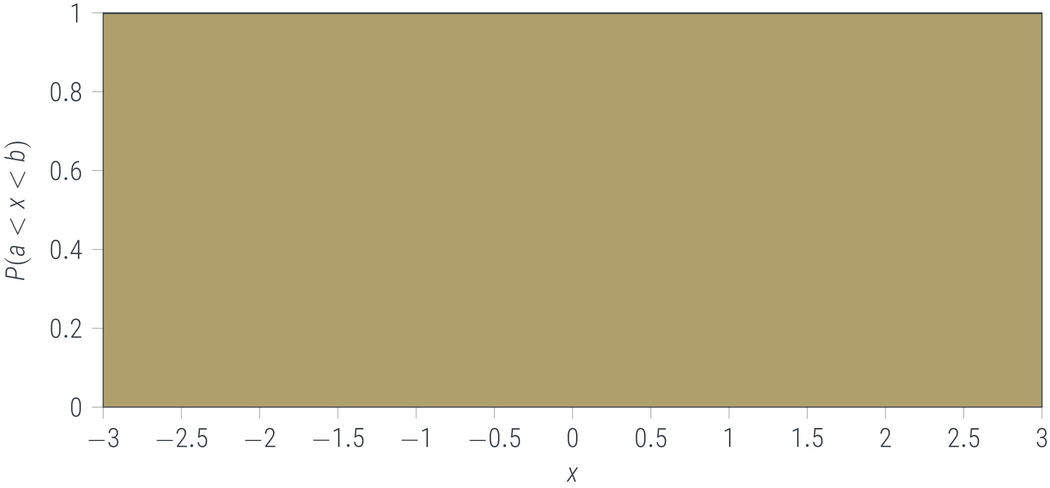
- with $F(x) := p(X \leq x)$ and $f(x) = \frac{d}{dx}F(x)$ we get

$$p(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

- F is called **cumulative distribution function** (CDF) and f **probability density function** (PDF)

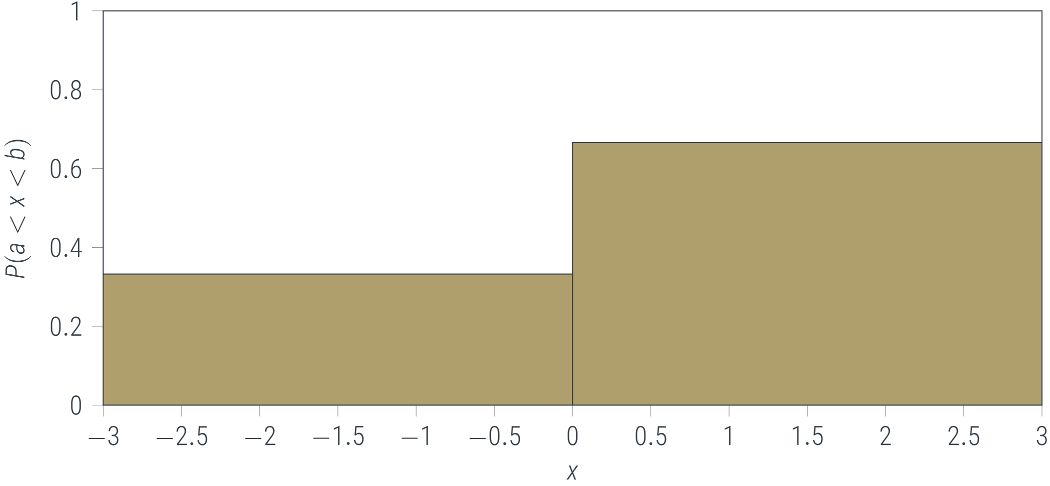
Probability Densities

the density of probability in a measurable region of the domain



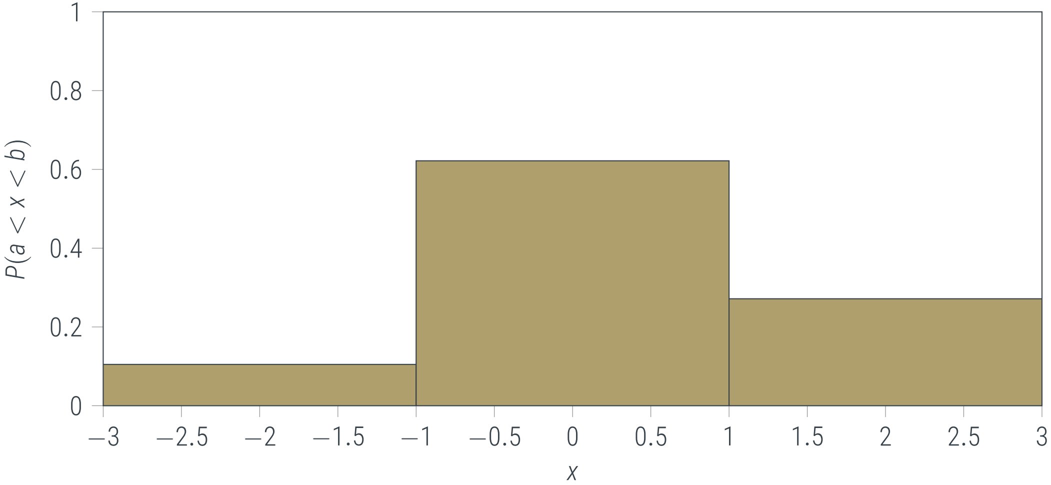
Probability Densities

the density of probability in a measurable region of the domain



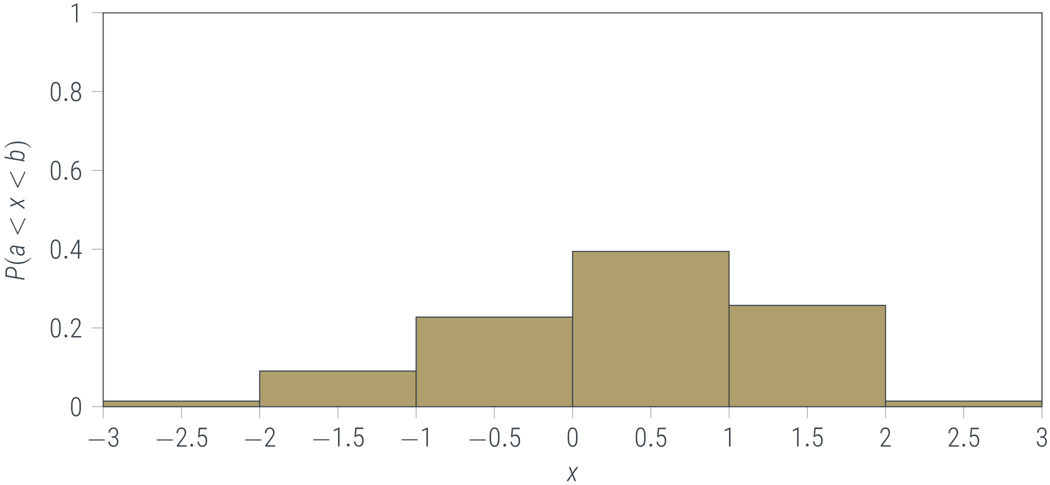
Probability Densities

the density of probability in a measurable region of the domain



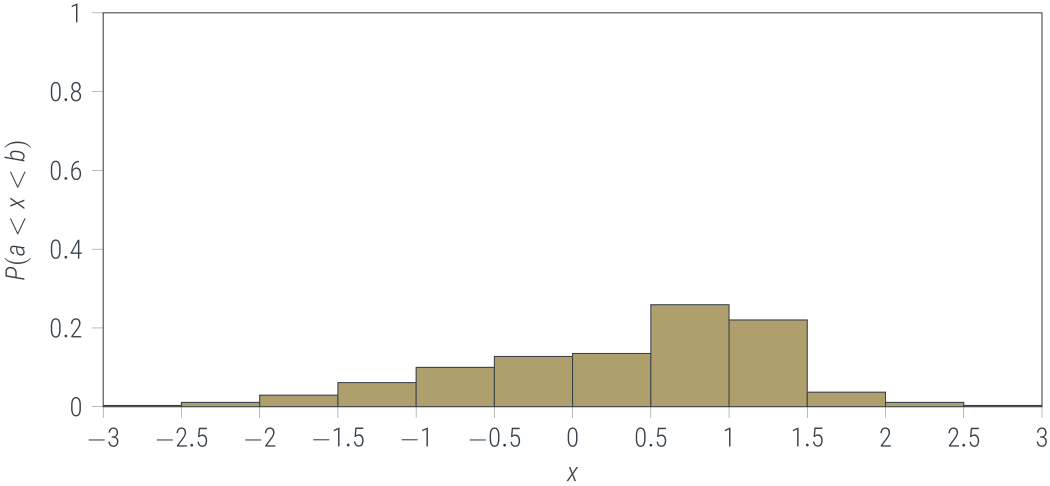
Probability Densities

the density of probability in a measurable region of the domain



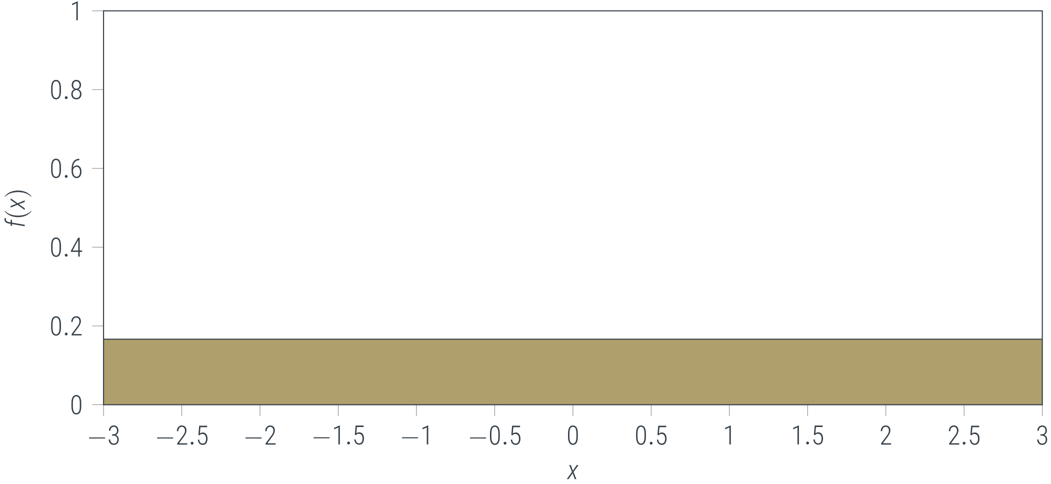
Probability Densities

the density of probability in a measurable region of the domain



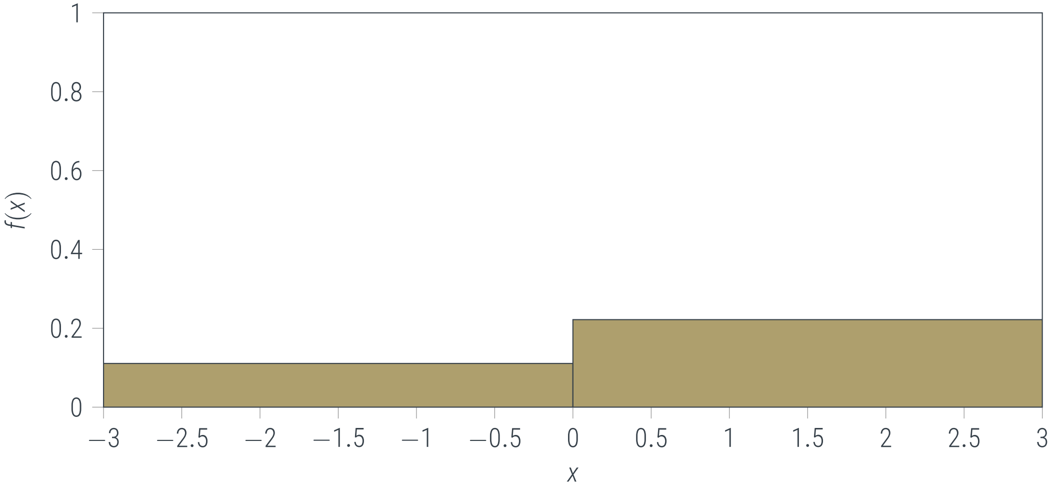
Probability Densities

the density of probability in a measurable region of the domain



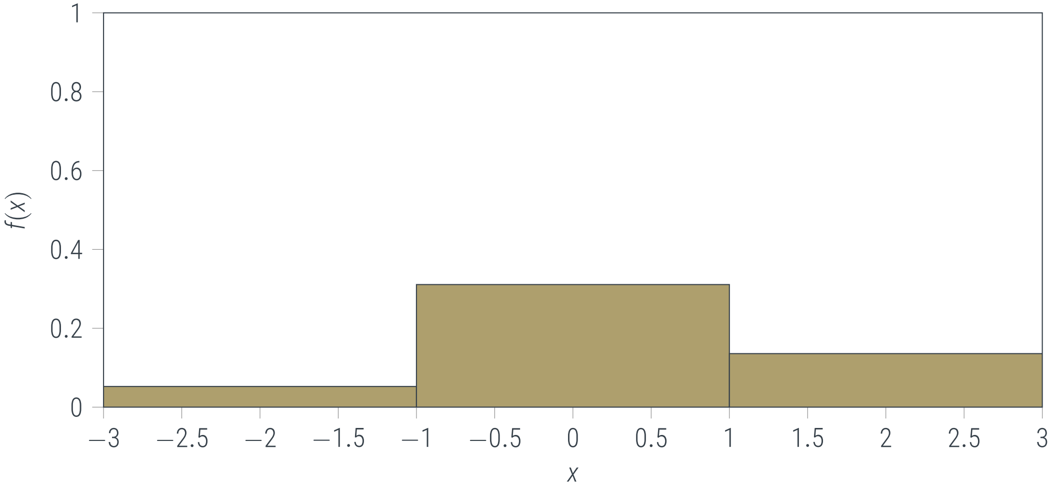
Probability Densities

the density of probability in a measurable region of the domain



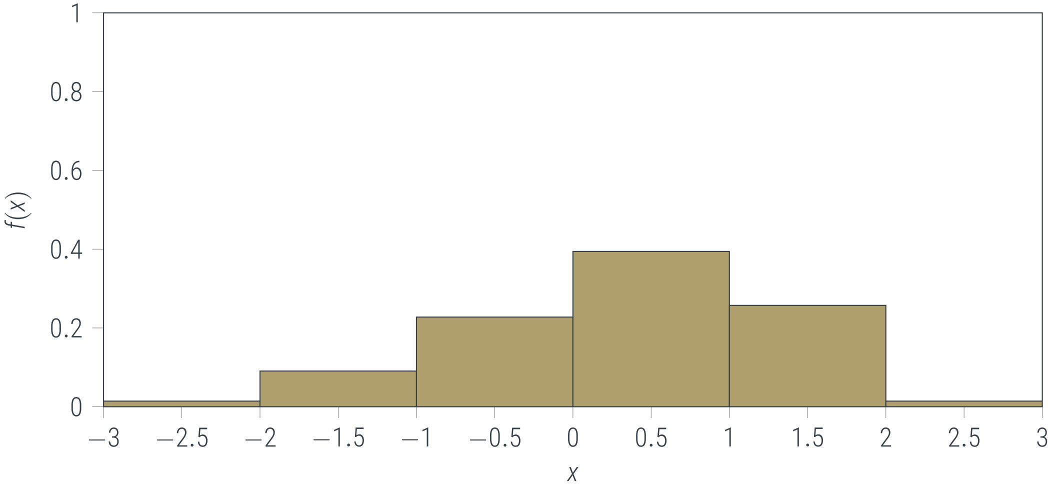
Probability Densities

the density of probability in a measurable region of the domain



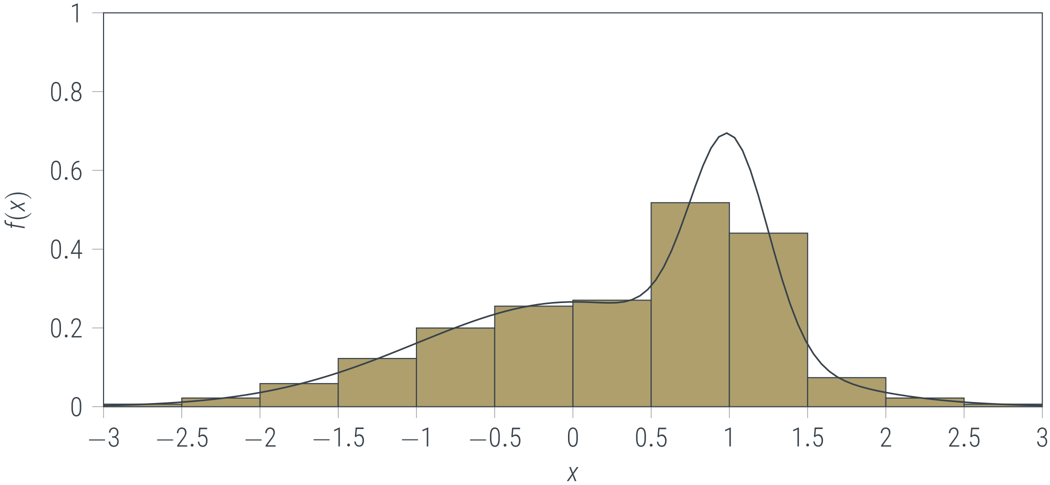
Probability Densities

the density of probability in a measurable region of the domain



Probability Densities

the density of probability in a measurable region of the domain



Discrete vs. continuous probabilities

an analogous extension

Probability theory as an extension of propositional logic

- finite set of propositional variables $A, B, \dots, Z \in \{0, 1\}$ jointly ranging over all boolean assignments
- sample space $\Omega = \{\text{all boolean assignments}\}$
- probability mass function $f : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} f(\omega) = 1$

Discrete probability theory (includes the previous case)

- random variable ranging in a discrete set, e.g. $\{0, 1, 2, \dots, \}$
- sample space $\Omega = \{0, 1, 2, \dots\}$
- *probability mass function* $f : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} f(\omega) = 1$

Continuous probability theory

- random variable ranging in a continuous set, e.g. real numbers $\mathbb{R}$
- sample space $\Omega = \mathbb{R}$
- *probability density function* $f : \Omega \to \mathbb{R}_+$, such that $\int_{\omega \in \Omega} f(\omega) d\omega = 1$

Discrete vs. continuous probabilities

some subtle differences

Probabilities measure the mass of subsets $E \subset \Omega$ of the sample space.

Discrete probabilities

- e.g. $\Omega = \mathbb{N}$
- probability mass function $f : \mathbb{N} \to [0, 1]$, such that $\sum_n f(n) = 1$
- probability $p(E) = \sum_{n \in E} f(n)$
- small letter p

Continuous probabilities

- e.g. $\Omega = \mathbb{R}$
- probability density function (PDF) $f : \mathbb{R} \to \mathbb{R}_+$, such that $\int f(x)dx = 1$
- probability $P(E) = \int_E f(x)dx$
- large letter P

Rules for continuous variables

In continuous domains, the pdf is analogous to the probabilities in the discrete case.

Theorem (rules for PDFs)

The standard rules of probability theory do hold for PDFs,

$$f(x,y) = f(x|y)\, f(y) \qquad \textit{product rule}$$

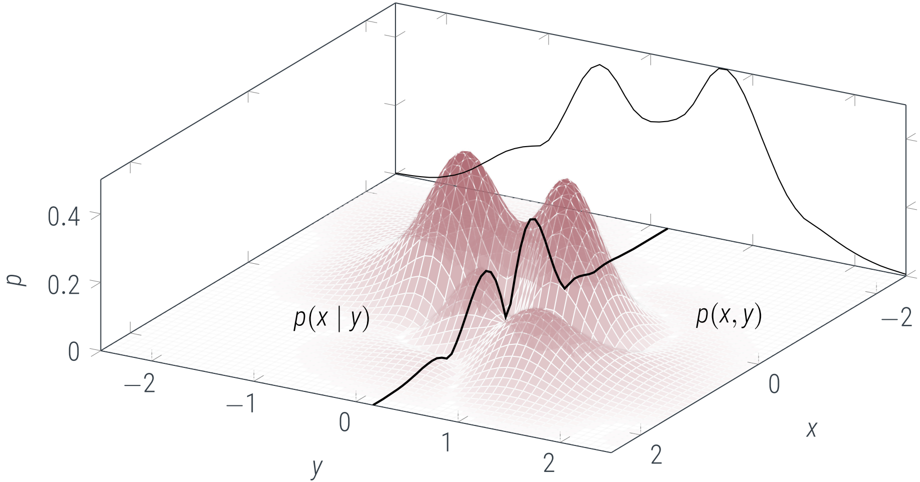
$$f(x) = \int f(x,y)\, dy \qquad \textit{sum rule}$$

For that reason we often write p for PDFs. Sums turned into integrals. Note that the product rule implies Bayes' theorem.

The same does **not** hold for CDFs (capital letter $P(E)$)! This is because, although $P(E)$ is a probability in the discrete sense $P(E) = p(X \in E)$, it's not a probability of the variable E. Here, notation bites us. Of course, the rules of probability still hold for discrete events $X_E = \mathbb{I}(X = x \in E)$ with $p(X_E) = p(E)$, but not for the real variable $E \in \mathbb{R}$.

Continuous Densities

a sketch



A Big Caveat

Probability density functions are only defined relative to a base measure

Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $f_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$f_Y(y) = f_X(v(y)) \cdot \left| \frac{dv(y)}{dy} \right| = f_X(v(y)) \cdot \left| \frac{du(x)}{dx} \right|^{-1}.$$

Proof: for $u'(X) > 0$: $\forall\ d_1 = u(c_1) < y < u(c_2) = d_2$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \leq v(y)) = \int_{c_1}^{v(y)} f(x)\,dx$$

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = f_X(v(y)) \cdot \frac{dv(y)}{dy} = f_X(v(y)) \cdot \left| \frac{dv(y)}{dy} \right|$$

A Big Caveat

Probability density functions are only defined relative to a base measure

Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $f_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$f_Y(y) = f_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right| = f_X(v(y)) \cdot \left|\frac{du(x)}{dx}\right|^{-1}.$$

Proof: for $u'(X) < 0$: $\forall\ d_2 = u(c_2) < y < u(c_1) = d_1$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \geq v(y)) = 1 - P(X \leq v(y)) = 1 - \int_{c_1}^{v(y)} f(x)\,dx$$

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = -f_X(v(y)) \cdot \frac{dv(y)}{dy} = f_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right|$$

Some Nomenclature

simplified. If you don't like this, wait for it

Probability (Measure)

The **probability** $p(E)$ of an event $E \subseteq \Omega$ is the **measure** assigned by the probability measure $p(E)$. In this sense, probability measure (map) and probability (output of map) are almost the same thing.

Probability Distribution

Consider a base space $(\Omega, \mathcal{F}, p)$ with a **probability measure** p. Then a **distribution** is the probability measure of any derived variable X that is a map from the base space to another measure space $(\mathcal{X}, \mathcal{A}, p_X)$. Such variables are called **Random Variables**. *Measure* and *distribution* are sometimes used interchangeably, but formally, a measure induces a distribution, which is not true the other way round.

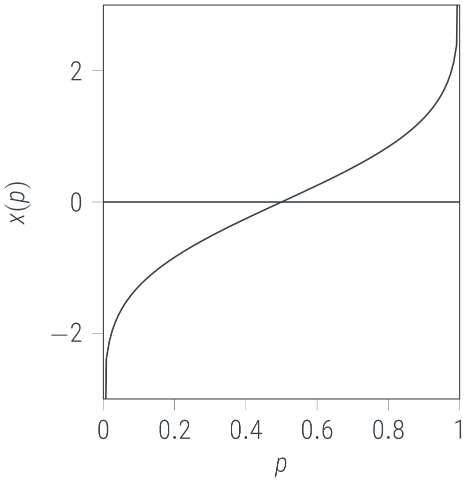
Probability Density (Function) — pdf

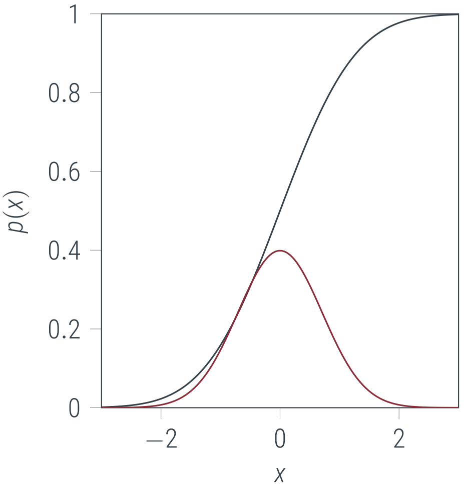
A probability density is the function $p(x) : \mathcal{X} \rightharpoonup \mathbb{R}$ such that if P is the measure of X,

$$P(X \in A \subset \mathcal{X}) = \int_A p(x)\, dx. \qquad \text{note: “}p(x) = \frac{dP(X)}{dx}\text{“}$$

Measure and Densities

in pictures





Some Nomenclature — Formal Definitions

If that's your sort of thing

Definition (probability measure — short version of full definition from Lecture 1)

Let $(\Omega, \mathcal{F}, p)$ be a measurable space (i.e. a measurable sample space Ω and σ-algebra $\mathcal{F}$). A positive measure p on $(\Omega, \mathcal{F})$ is called a **probability measure** if $p(\Omega) = 1$.

Definition (probability distribution)

Let X be a measurable function from $(\Omega, \mathcal{F}, p)$ to $(\mathcal{X}, \mathcal{A})$ (= a function between two measurable spaces such that the pre-image of every measurable set is measurable) (such functions are also called **random variables**). Then the **probability distribution** of X is the pushforward X_*p of p. That is, it is the measure satisfying $X_*p = pX^{-1}$.

Simplified takeaway: For our purposes, **probability measure** = **probability distribution**.

Some Nomenclature — Formal Definitions

If that's your sort of thing

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Definition (probability density function)

Let X be a random variable with distribution X_*P on $(\mathcal{X}, \mathcal{A})$. Let μ be a reference measure (e.g. for $\mathcal{X} = \mathbb{R}^N$: The Lebesgue measure). The **probability density function (pdf, aka. "density")** of X is a measurable function f on $(\mathcal{F}, \mathcal{A})$ such that for any measurable set $A \in \mathcal{A}$,

$$X_*P(A) = \int_{X^{-1}A} dX_*P = \int_A f\, d\mu.$$

This property is also written short-hand as

$$f = \frac{dX_*P}{d\mu}$$

and f is also called the **Radon-Nikodym derivative** of X_*P with respect to μ.

Apologies: I tend to mix up "distribution" and "density". Try to catch me if I do!

Densities have Units!

PDFs come with a few caveats

- **Probabilities** are unitless, but probability **densities** have units:

$$P(0 < x < 1) = 20\% \qquad \text{but} \qquad p(x = 0.5) = 0.3\frac{1}{\mathrm{m}}.$$

- Probability densities can be > 1
- Densities are only defined relative to a **base measure**
  (recall from above: Changing $x \rightarrow y = f(x)$ requires a change of measure)
- There is no **ignorant** density: At best, it's uniform wrt. a particular base measure

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere

If we sample independently, what is the likelihood for a positive or a negative observation?

$$p(X = 1 \mid \pi) = \pi; \qquad\qquad p(X = 0 \mid \pi) = 1 - \pi$$

An example

Based on a very famous argument

What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses
- Inference? Bayes' theorem!

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

- uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere

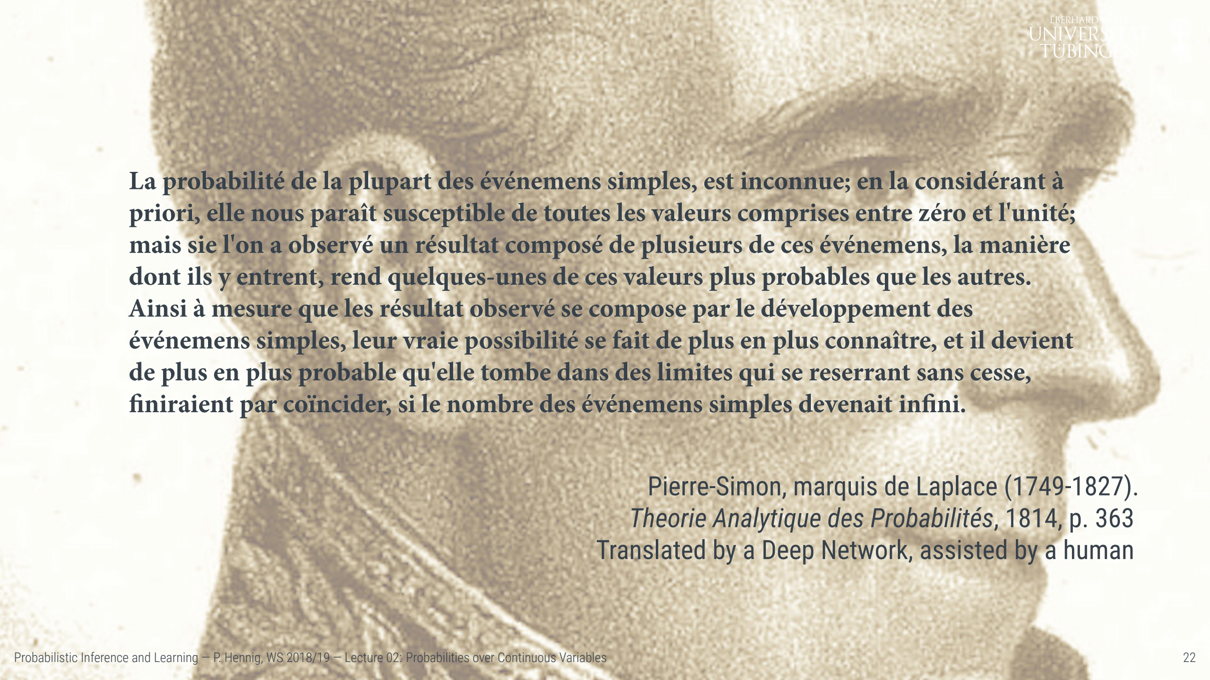
If we sample independently, what is the likelihood for a positive or a negative observation?

$$p(X = 1 \mid \pi) = \pi; \qquad p(X = 0 \mid \pi) = 1 - \pi$$

What is the posterior after n positive, m negative observations?

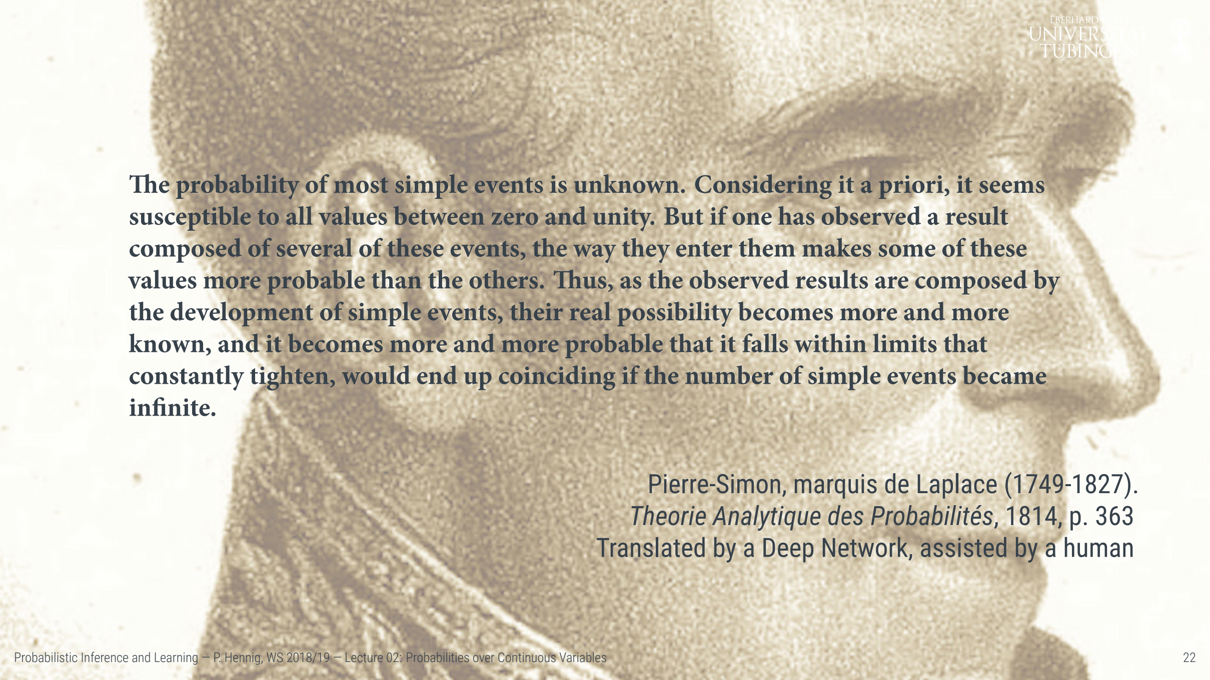
$$p(\pi \mid n, m) = \frac{\pi^n (1 - \pi)^m \cdot 1}{\int \pi^n (1 - \pi)^m \cdot 1\, d\pi} = \frac{\pi^n (1 - \pi)^m}{B(n + 1, m + 1)}$$

DEMO



La probabilité de la plupart des événemens simples, est inconnue; en la considérant à priori, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais sie l'on a observé un résultat composé de plusieurs de ces événemens, la manière dont ils y entrent, rend quelques-unes de ces valeurs plus probables que les autres. Ainsi à mesure que les résultat observé se compose par le développement des événemens simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui se reserrant sans cesse, finiraient par coïncider, si le nombre des événemens simples devenait infini.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 363
Translated by a Deep Network, assisted by a human



The probability of most simple events is unknown. Considering it a priori, it seems susceptible to all values between zero and unity. But if one has observed a result composed of several of these events, the way they enter them makes some of these values more probable than the others. Thus, as the observed results are composed by the development of simple events, their real possibility becomes more and more known, and it becomes more and more probable that it falls within limits that constantly tighten, would end up coinciding if the number of simple events became infinite.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 363
Translated by a Deep Network, assisted by a human

Let’s be more careful with notation!
(but only once more, then we’ll be sloppy)

Example — inferring probability of wearing glasses (2)

Now with more care

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Example — inferring probability of wearing glasses (2)

Now with more care

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Example — inferring probability of wearing glasses (2)

Now with more care

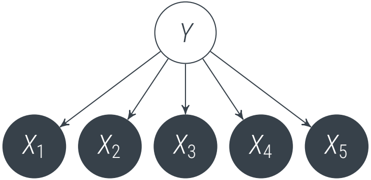
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Example — inferring probability of wearing glasses (2)

Now with more care

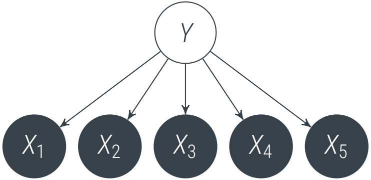
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Generative model and joint probability

- we abbreviate $Y = \pi$ as π, $X_i = x_i$ as x_i
- $p(\pi)$ is the prior of Y, written fully $p(Y = \pi)$
- $p(x_i|\pi)$ is the likelihood of observation x_i
- note that the likelihood is a function of π

Example – inferring probability of wearing glasses (3)

Bayesian inference of a Bernoulli probability

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Example — inferring probability of wearing glasses (3)

Bayesian inference of a Bernoulli probability

Probability of wearing glasses without observations

$$p(\pi | \text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi | x_1) = Z_1^{-1} p(x_1 | \pi) p(\pi)$$

Example — inferring probability of wearing glasses (3)

Bayesian inference of a Bernoulli probability

Probability of wearing glasses without observations

$$p(\pi | \text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi | x_1) = Z_1^{-1} p(x_1 | \pi) p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi | x_1, x_2) = Z_2^{-1} p(x_2 | x_1, \pi) p(x_1 | \pi) p(\pi) = Z_2^{-1} p(x_2 | \pi) p(x_1 | \pi) p(\pi)$$

Example — inferring probability of wearing glasses (3)

Bayesian inference of a Bernoulli probability

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = Z_1^{-1} p(x_1|\pi) p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi|x_1, x_2) = Z_2^{-1} p(x_2|x_1, \pi) p(x_1|\pi) p(\pi) = Z_2^{-1} p(x_2|\pi) p(x_1|\pi) p(\pi)$$

...

Probability of wearing glasses after five observations

$$p(\pi|x_1, x_2, x_3, x_4, x_5) = Z_5^{-1} \left(\prod_{i=1}^{5} p(x_i|\pi) \right) p(\pi)$$

Example — inferring probability of wearing glasses (4)

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1 - \pi & \text{for } x_1 = 0 \end{cases}$$

Example — inferring probability of wearing glasses (4)

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1 - \pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV N for the number of observations being 1 (with values n)
- RV M for the number of observations being 0 (with values m)

Example — inferring probability of wearing glasses (4)

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1-\pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV *N* for the number of observations being 1 (with values *n*)
- RV *M* for the number of observations being 0 (with values *m*)

Probability of wearing glasses after five observations

$$\begin{aligned} p(\pi|x_1,x_2,x_3,x_4,x_5) &= Z_5^{-1}\left(\prod_{i=1}^{5} p(x_i|\pi)\right) p(\pi) \\ &= Z_5^{-1}\pi^n(1-\pi)^m p(\pi) \\ &= p(\pi|n,m) \end{aligned}$$

Example — inferring probability of wearing glasses (5)

a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

Example — inferring probability of wearing glasses (5)

a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

Example — inferring probability of wearing glasses (5)

a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n, m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1}$$

with parameters $a > 0, b > 0$

the Beta ***distribution*** *with parameter a and b*

Example — inferring probability of wearing glasses (5)

a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

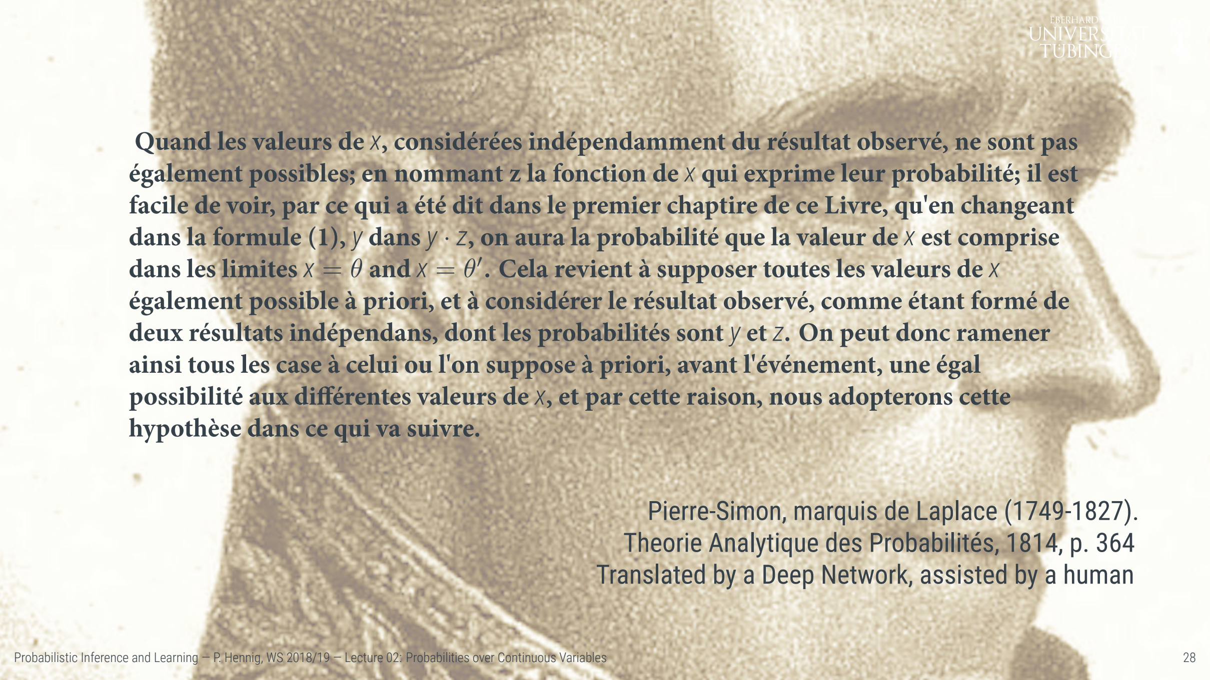
$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1} \qquad \text{with parameters } a > 0, b > 0$$

the Beta ***distribution*** *with parameter a and b*

Let's give the normalization factor Z of the beta distribution a name!

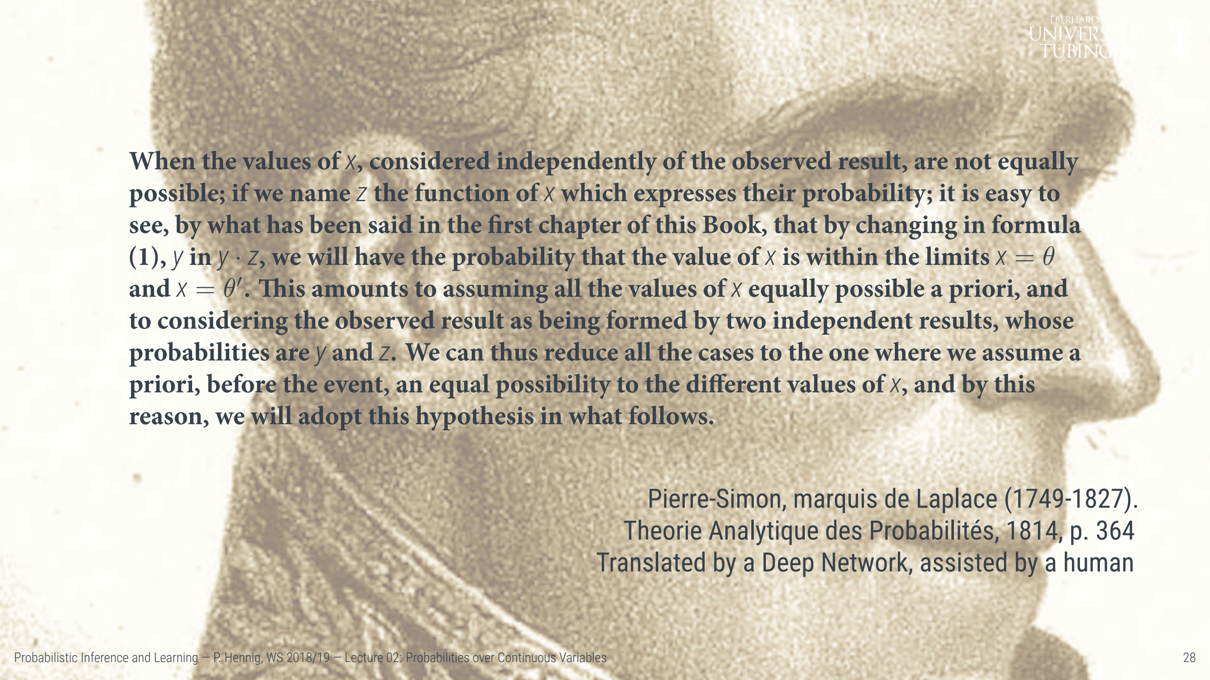
$$B(a,b) = \int_0^1 \pi^{a-1}(1-\pi)^{b-1}d\pi$$

the Beta ***function*** *with parameters a and b*



Quand les valeurs de x, considérées indépendamment du résultat observé, ne sont pas également possibles; en nommant z la fonction de x qui exprime leur probabilité; il est facile de voir, par ce qui a été dit dans le premier chaptire de ce Livre, qu'en changeant dans la formule (1), y dans $y \cdot z$, on aura la probabilité que la valeur de x est comprise dans les limites $x = \theta$ and $x = \theta'$. Cela revient à supposer toutes les valeurs de x également possible à priori, et à considérer le résultat observé, comme étant formé de deux résultats indépendans, dont les probabilités sont y et z. On peut donc ramener ainsi tous les case à celui ou l'on suppose à priori, avant l'événement, une égal possibilité aux différentes valeurs de x, et par cette raison, nous adopterons cette hypothèse dans ce qui va suivre.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 364
Translated by a Deep Network, assisted by a human



When the values of x, considered independently of the observed result, are not equally possible; if we name z the function of x which expresses their probability; it is easy to see, by what has been said in the first chapter of this Book, that by changing in formula (1), y in $y \cdot z$, we will have the probability that the value of x is within the limits $x = \theta$ and $x = \theta'$. This amounts to assuming all the values of x equally possible a priori, and to considering the observed result as being formed by two independent results, whose probabilities are y and z. We can thus reduce all the cases to the one where we assume a priori, before the event, an equal possibility to the different values of x, and by this reason, we will adopt this hypothesis in what follows.

Pierre-Simon, marquis de Laplace (1749-1827).
Theorie Analytique des Probabilités, 1814, p. 364
Translated by a Deep Network, assisted by a human

Aside: The Eulerian Integrals

For an evening at the fireplace: [Philip J. Davis. *Leonhard Euler's Integral: A Historical Profile of the Gamma Function*, 1959]

For $m, n \in \mathbb{N}$ and $x, y, z \in \mathbb{C}$:

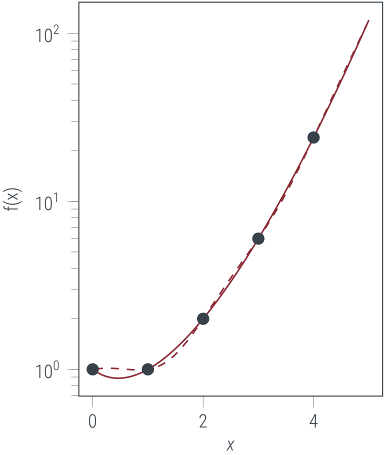
$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

$$= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \text{ if } x+\bar{x}, y+\bar{y} > 0$$

$$B(m,n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!}$$

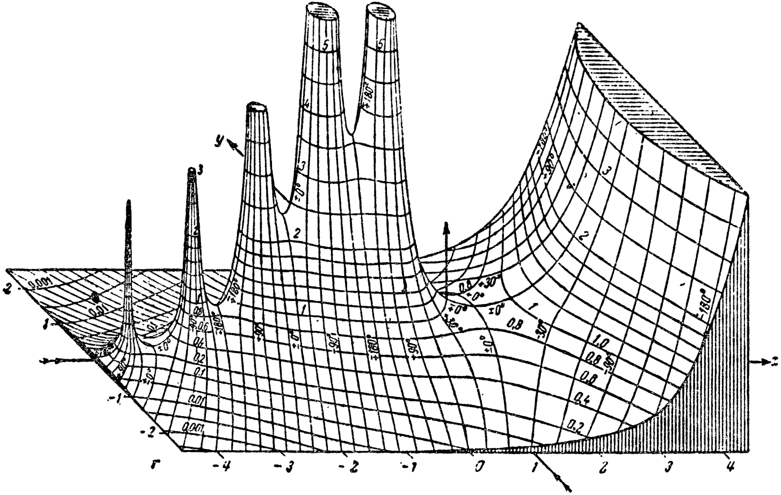
$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt$$

$$\Gamma(n) = (n-1)!$$



Aside: The Gamma Function

[from: E. Jahnke & F. Emde, *Tafeln höherer Funktionen, 4.ed., Leipzig 1948*]



Probability Distributions over Continuous Variables

- **probability density functions** (PDFs) distribute probability (“mass”) over continuous domains
    - they change nontrivially under changes of measure / units
    - they can have values > 1, but $\int p(x)\, dx = 1$
- Sum and Product Rule, Bayes’ Theorem transfer to PDFs

Example: Inferring a Bernoulli Probability

$$p(\pi) = \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} \qquad p(n,m \mid \pi) = \pi^{n}(1-\pi)^{m} \quad \Rightarrow \quad p(\pi|n,m) = \frac{\pi^{n+a-1}(1-\pi)^{m+b-1}}{B(a+n,b+m)}$$

- An example of a **conjugate prior** (more later)