Introduction

Plan for the first four lectures

1. Background from probability theory, focusing on conditional probability. Introduction to Bayesian statistical models and Bayesian inference. Point estimates, prediction, conjugate priors.
2. Deep dive into advanced regression, fancy regression, Bayesian regression. Ridge, lasso, elastic net. Generalization performance and cross-validation.
3. Classifiers. Logistic regression, support vector machine and neural networks. Backpropagation and stochastic gradient descent.
4. Applications of the first three lectures to buy-side quant research, especially to multi-factor models and portfolio optimization.

Outcomes and Events

- Let us begin by recalling the basic setup under which one can do probability theory.

- Historically, much of statistics (and probability theory) arose from the need to understand the outcomes of scientific experiments, and the data which was collected in the process.
- Almost always, performing an experiment involves some degree of randomness in the outcome, if only due to the finite precision of the measuring device.
- Some experiments are almost completely random, such as rolling a die.

- If the experiment is rolling a die, the *outcomes* are the possible values for the number that comes up:

$$1, 2, \ldots, 6.$$

- The event described as "rolling an even number" is not a single outcome, but a set of outcomes:

$$\{2, 4, 6\}.$$

- Very often, what we want to assign probabilities to are *sets of outcomes*, not single outcomes.
- A set of outcomes is therefore called an *event*, and probability theory is about assigning probabilities to events.

- If E is an event, then the event that E does *not* happen is represented by the complement.
- The complement is any outcome *except* for those in E,

$$\begin{aligned} E^c &= \{x \in \Omega : x \notin E\} \\ &= \Omega \setminus E \end{aligned}$$

Partitions and the representation of information

- Suppose that a random experiment is performed, and its outcome ω is unknown, but we are given some information that is enough to narrow down the possible value of ω.
- We can then create a list of events that are sure to contain the actual outcome, and other sets that are sure not to contain it.

Definition 1.1

A (countable) partition $\mathcal{P}$ of a sample space $\Omega \neq \emptyset$ is a countable collection of mutually disjoint nonempty subsets whose union is Ω. That is,

$$\mathcal{P} = \{A_i\}_{i \geq 1} = \{A_1, A_2, \ldots\},$$

with

$$\bigcup_{i \geq 1} A_i = \Omega, \text{ and } i \neq j \Rightarrow A_i \cap A_j = \emptyset.$$

The subsets A_i are called atoms of the partition $\mathcal{P}$.

- In finance, partitions are often used to represent partial information.
- Intuitively, if partial information about the outcome ω is available in the form of a partition, then we are able to say which event from the partition has occurred.
- For example, if

$$E \subseteq \Omega$$

is an event, then the partition

$$\mathcal{P}_E = \{E, E^c\}$$

represents the information of whether the event E has occurred or not.

- As a second example, roll two dice and suppose that we can't see the individual dice, but we do get to see the sum of values on the facing-up sides.
- The partition representing this level of information consists of the 11 sets of the form

$$A_k = \{(i,j) : 1 \leq i,j \leq 6 \text{ and } i+j=k\}.$$

- If $\mathcal{P}$ is a countable partition then the collection of all unions of sets in the partition (including the empty set) is a σ-algebra of "measurable sets" since we expect to be able to "measure" the probability of any such set.

Probability spaces

- A probability space is a triple $(\Omega, \mathcal{F}, P)$.
- The first object Ω is the set of possible outcomes, sometimes called the sample space.
- The second object $\mathcal{F}$ is a collection of events, that is, a set whose elements are certain subsets of Ω.
- The third object, P, maps an event to its probability and is called a probability measure.
- We specify the axioms it must satisfy below.

- The set $\mathcal{F}$ must contain all events we need in order to analyze a particular problem, or to represent a certain set of information we have.
- It also has to satisfy certain logical consistency requirements.
- If we know whether E happened, then obviously we must also know whether E did not happen.

$$\text{if } \; E \in \mathcal{F} \text{ then also } E^c \in \mathcal{F}.$$

- Similarly, if we know whether A happened, and separately we know whether B happened, we must also know whether the composite event "A or B", represented by the union of the underlying outcomes $A \cup B$, happened.
- The full set of conditions that we need is simply that $\mathcal{F}$ must be a σ-algebra.

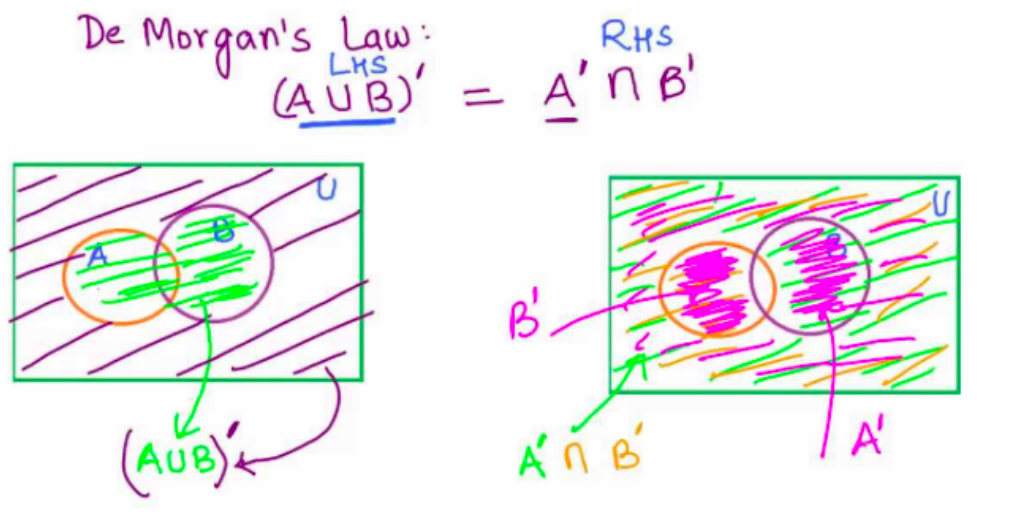
Definition 1.2

Given a set Ω, a σ-algebra (also called a σ-field) in Ω is a collection $\mathcal{F}$ of subsets of Ω that contains the empty set, and is closed under complements and countable unions.

- Many other reasonable conditions are implied by the definition of σ-algebra.
- One may show (exercise) that a σ-algebra is also closed under finite intersections and even countable intersections, by applying *De Morgan's Laws*:

$$(E \cup F)^c = E^c \cap F^c \tag{1.1}$$
$$(E \cap F)^c = E^c \cup F^c \tag{1.2}$$

Moreover, Ω is in $\mathcal{F}$ since it is the complement of the empty set.



The probability measure P associates a real number $P(A)$ to every event A in the σ-algebra $\mathcal{F}$, such that

1. For every event A, $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. If $A_1, A_2, \ldots$ is a sequence of pairwise disjoint events, then

$$P(\bigcup_i A_i) = \sum_i P(A_i)$$

- The definition of σ-algebra neatly entails that the arguments to the probability measure P are guaranteed to be events in $\mathcal{F}$; in particular,
$$\bigcup_i A_i \in \mathcal{F}.$$
- Without axiom 2, P is simply called a *measure* but does not necessarily represent probability.

Example 1.3

The "fair dice" example:

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

For any $A \subset \Omega$, define

$$P(A) = |A|/6.$$

Example 1.4

Let $\Omega = \{1, 2, \ldots\}$ *and suppose we are given numbers*

$$p_1, p_2, \ldots \geq 0$$

with

$$p_1 + p_2 + \cdots = 1.$$

For any $A \subset \Omega$*, define*

$$P(A) = \sum_{i \in A} p_i$$

Example 1.5

Let

$$\Omega = \{(x, y) : x^2 + y^2 < 1\},$$

and

$$P(A) = \text{area}(A)/\pi.$$

This is a case where A cannot be an arbitrary subset of the circle – for some sets area cannot be defined!

Example 1.6

Suppose a single outcome is a length-N sequence

$$\omega = (x_i)_{1 \leq i \leq N}$$

where each $x_i \in \{1, 0\}$ *where* 1 *symbolically represents an up move in a stock price, and* 0 *symbolically represents a down move. The sample space* Ω *consists of all sequences like this. For some number* $0 < p < 1$, *define*

$$P(\omega) = p^{\sum_i x_i}(1-p)^{N-\sum_i x_i}.$$

This model is used in option pricing.

- We now list a few fairly-straightforward consequences of the axioms.
- Most of these are properties that, we can convince ourselves intuitively, any reasonable model of probability should satisfy.

(C0) $P(\emptyset) = 0$.

In Axiom 3, take all sets to be $\emptyset$.

(C1) If $A_1 \cap A_2 = \emptyset$, then $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

In Axiom 3, take all sets other than first two to be $\emptyset$.

(C2) $P(A^c) = 1 - P(A)$.

Apply (C1) to $A_1 = A, A_2 = A^c$.

(C3) $0 \leq P(A) \leq 1$.

Use that $P(A^c) \geq 0$ in (C2).

(C4) If $A \subset B$, then

$$P(B) = P(A) + P(B \setminus A) \geq P(A).$$

Use (C1) for $A_1 = A$ and $A_2 = B \setminus A$.

(C5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

The proof is an exercise.

- Consider a collection of σ-algebras indexed by time

$$\{\mathcal{F}_t : t \geq 0\}$$

with the additional property that, if $t \geq s$ and E is any event in $\mathcal{F}_s$ then E is also in $\mathcal{F}_t$.

- This is a good way of representing the knowledge of a non-forgetful trader.
- If an event was measurable at an earlier time, it remains measurable at all later times.

Open sets

- First, define an *open ball* to be the inside of a sphere without the boundary:

$$B_\epsilon(x) = B(x;\epsilon) := \{y \in \mathbb{R}^n : d(x,y) < \epsilon\}$$

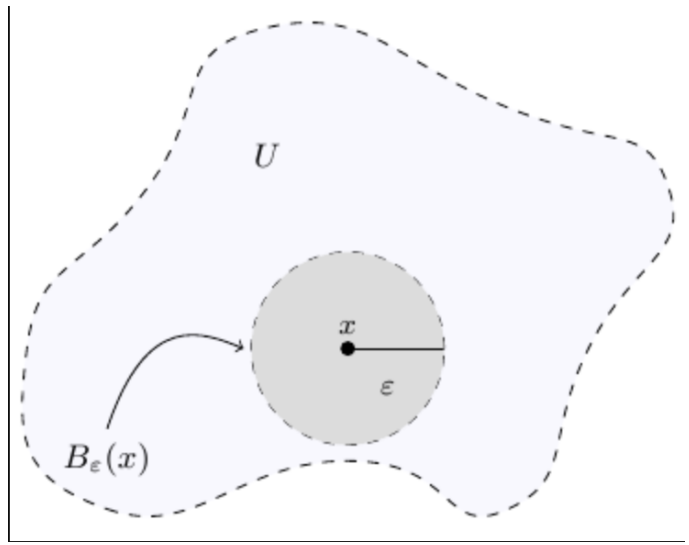
where $d(x,y)$ is the standard Euclidean distance.

A set

$$U \subseteq \mathbb{R}^n$$

is said to be *open* if for any $x \in U$, there exists $\epsilon > 0$ such that

$$B(x; \epsilon) \subset U.$$



Definition 1.7

The smallest (i.e. containing the fewest possible measurable sets) σ-algebra in $\Omega = \mathbb{R}^n$ containing all open sets is called the Borel σ-algebra. Any set which is in the Borel σ-algebra is called a Borel set.

- Consider the kind of sets this must include.
- Any countable union or intersection of open sets (including open intervals) is Borel.
- A closed set is just the complement of an open set, so all closed sets are also Borel – for example, any closed interval $[a, b]$ is the complement of the open set

$$(-\infty, a) \cup (b, +\infty)$$

  and is hence Borel.
- Then consider all countable unions of the sets we have already mentioned; then consider all countable intersections of those.
- Repeatedly take all countable unions and all countable intersections of sets previously obtained.

- This definition is sufficiently inclusive that it takes considerable cleverness to construct a subset of $\mathbb{R}$ that is actually *not* a Borel set!
- In any case, the Borel σ-algebra is extremely important in probability theory for describing continuous random variables.

Definition 1.8

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be continuous if the inverse image of every open set is open, i.e. $f^{-1}(U)$ is open whenever U is open.

- For functions $f : \mathbb{R} \to \mathbb{R}$, our definition agrees with the more basic definition using ϵ and δ.
- In calculus, f is said to be continuous at c if for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all x,

$$|x - c| < \delta \;\Rightarrow\; |f(x) - f(c)| < \varepsilon.$$

- In other words, for any $f(x)$ in an ε-ball around $f(c)$, the preimage $f^{-1}(f(x))$ lies in a δ-ball around c.

Random variables

- Intuitively, a random variable is a number whose value depends upon the outcome of a random experiment.
- Mathematically, a random variable X is a real-valued function on Ω, the space of outcomes:

$$X : \Omega \to \mathbb{R}.$$

- Consider rolling a die as an "experiment" in which the set of possible outcomes, Ω, consists of the six distinct faces of the die.
- The numerical value inscribed on the face of the die is then a random variable, because it associates a real number with each outcome (face).

- As another example, suppose that, disliking snow, so I purchase an insurance contract that pays \$1 if it snows in Chicago tomorrow.
- The amount of money I will receive from that insurance is a random variable on the sample space

$$\Omega = \{ \text{ snow, no-snow } \}.$$

- Imagine the confusion that would occur in trying to model a random variable X if the probability of the event "$X > a$ and $X < b$" were undefined.
- This would not be a useful model, so we have to add constraints which ensure that any set of outcomes for which we need to calculate probability is actually an event.
- In other words we need to ensure that the set of outcomes

$$\{\omega \in \Omega : X(\omega) \in (a, b)\}$$

will always be an event in any case.

- More generally, we want to be able to describe the probability that X lies in any countable union or intersection of intervals, complements of those and so forth.
- Ultimately, given any Borel set B, we will need $X^{-1}(B)$ to be an event in our probability space.
- This is a strong requirement, since there are so many Borel sets, but it will be satisfied for all of the random variables we need to model.
- If this requirement is satisfied, X is said to be Borel measurable.

- The *cumulative distribution function* of X is the function

$$F_X : \mathbb{R} \to [0, 1]$$

given by

$$F_X(x) = P(X \leq x)$$

- The notation $P(X \leq x)$ is shorthand for

$$P(X^{-1}(B)) \text{ where } B = (-\infty, x]$$

and note that B is, of course, a Borel set.

- A random variable X is said to be *discrete* if it takes at most countably many distinct values; also, X is said to be *continuous* if the function F_X is continuous.
- For a discrete random variable, we define its *probability mass function* as

$$p_X(x) = P(\{\omega : X(\omega) = x\}).$$

- The more common case in finance is that the space of outcomes is $\Omega = \mathbb{R}$, or perhaps the positive reals for modeling price, and the superset of all possible events $\mathcal{F}$ is the Borel σ-algebra.
- The measure of a general Borel set can be built up by defining the measure of its basic components, the open intervals.
- Perhaps the simplest measure of an open interval (a, b) is its length $b - a$; this is called the Lebesgue measure.
- However, the Lebesgue measure is not a probability measure, and cannot be made into one by dividing by a constant, because the measure of $\Omega = \mathbb{R}$ is infinite.

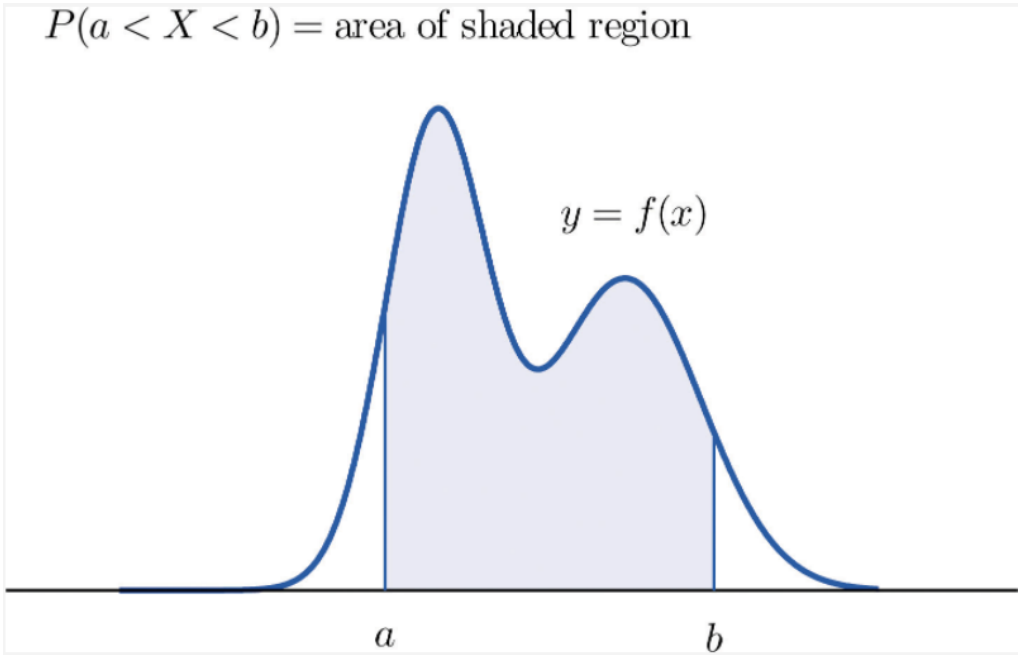
Instead, if we have a non-negative integrable function, $f(x)$, satisfying

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

then we can define a probability measure on $\Omega = \mathbb{R}$, by initially defining it for open intervals as

$$P[(a, b)] = \int_{a}^{b} f(x)dx$$

and this definition can be extended to any Borel set.



- More generally, suppose that $X : \Omega \to \mathbb{R}$ is any random variable on any probability space.
- Further suppose there exists $f_X : \mathbb{R} \to \mathbb{R}$ such that for any $a, b \in \mathbb{R}$ with $a < b$,

$$P[X \in (a, b)] = \int_a^b f_X(x)dx$$

then we say f_X is the *probability density function* of the random variable.

- We may omit the subscript on f_X when the meaning is clear.
- The values of $f_X(x)$ do not themselves give probabilities, but we can think heuristically of

$$f_X(x)dx$$

as the probability that X takes values in a very small interval $[x, x + dx]$.

- A random variable on $\mathbb{R}$ is said to be *Gaussian* if its density takes the form

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

  with the obvious extension to $\mathbb{R}^n$.
- A random variable has a density function if and only if its cumulative distribution function $F(x)$ is absolutely continuous.
- In this case, F is almost everywhere differentiable, and its derivative is the density f.

- Consider rolling n dice simultaneously, and collecting their numerical values into a vector; this would then be a *random vector*.
- Much of what we have done for $\mathbb{R}$-valued random variables can be extended in a natural way to a random vector, which is a function

$$X : \Omega \to \mathbb{R}^n.$$

- In particular, the density for a random vector (also called a *multivariate density*) is a function $f_X : \mathbb{R}^n \to \mathbb{R}$ defined by

$$P(X \in B) = \int_B f(\boldsymbol{x})\, d^n\boldsymbol{x} \tag{1.3}$$

where B is any Borel set in $\mathbb{R}^n$, where $\boldsymbol{x} \in \mathbb{R}^n$, and where $d^n\boldsymbol{x}$ is short for $dx_1 dx_2 \ldots dx_n$.

- Given a multivariate density

$$f_X(x_1, \ldots, x_n),$$

the *marginal density* of a variable x_i is defined as

$$f_X(x_i) = \int_{\mathbb{R}^{n-1}} f(x_1, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n \quad (1.4)$$

where, note that dx_i is purposely left out, and the integration is over $\mathbb{R}^{n-1}$ because we are not integrating over x_i.

- Let $X : \Omega \to \mathbb{R}$ be any random variable.
- Define the σ-algebra generated by X to be the smallest σ-algebra containing all preimages $X^{-1}(B)$ of all Borel sets B.
- Roughly speaking, if you have observed the value of X then you can answer the question of whether that value was in any given Borel set B, and so all such events are become measurable with respect to $\sigma(X)$.
- This definition extends in a natural way to the sigma-algebra generated by a finite collection of random variables.

Stochastic processes

- A discrete-time stochastic process is a collection of random variables

$$X_n, n \in \mathbb{N} = \{0, 1, 2, \ldots\}$$

  all defined on the same probability space.
- The key property that we will use is that the σ-algebras

$$\mathcal{F}_n = \sigma(X_0, X_1, X_2, \ldots, X_n)$$

  form a filtration, or a representation of the information that we continue to observe X without forgetting its past values.

Conditional Probability

- The *conditional probability* of event $E \in \mathcal{F}$ given event $F \in \mathcal{F}$ is

$$P(E \mid F) := \frac{P(E \cap F)}{P(F)} \tag{1.5}$$

provided $P(F) > 0$.

- Conditional probability may be thought of as relative probability: $P(E \mid F)$ is the probability of E relative to the reduced sample space consisting of only those outcomes in the event F.
- Stephen Stigler wrote that, in a sense, all probabilities are conditional since even "unconditional" probabilities are relative to the sample space Ω, and it is only by custom that we write $P(E)$ instead of the equivalent $P(E \mid \Omega)$.

- Suppose we ask a question like, "given that rain is freezing and forming ice, what is the probability that your flight will be canceled?"
- The answer is presumably higher than the overall cancellation rate across all flights, because ice can interfere with the proper operation of the aircraft.
- The set of outcomes relevant for this question is

$$\Omega = \{IC, IR, NC, NR\}$$

  where I = ice, N = no ice, C = canceled, and R = running.
- If E is the event of flight cancellation, then $E = \{IC, NC\}$ and if F is the event of ice, then $F = \{IC, IR\}$.
- If we know that conditions are icy, then we at least know that event NC has not occurred, so we eliminate that event when we consider $E \cap F = \{IC\}$.

- Since I live just outside New York city and teach in Chicago, this is a real example for me.

- Suppose that X and Y are random variables in a discrete probability space.
- Considering (1.5), take A as the event

$$\{\omega \in \Omega : X(\omega) = x\}$$

and similarly, B as the event

$$\{\omega \in \Omega : Y(\omega) = y\}.$$

- We can then write the conditional probability (1.5) of these two events in a simplified notation reminiscent of density functions:

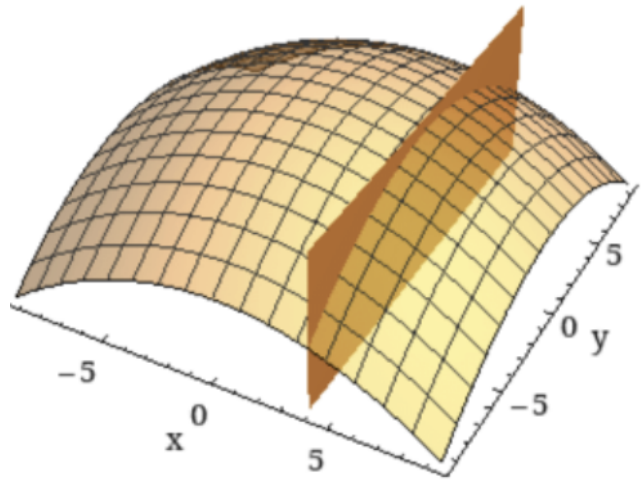
$$p(y \,|\, x) = \frac{P(X = x \text{ and } Y = y)}{P(X = x)} = \frac{p(x, y)}{p_X(x)}. \qquad (1.6)$$

- If X and Y are continuous random variables, the definition in (1.6) serves as an inspiration for the definition of the analogous conditional density:

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

which we agree to leave undefined for x such that $f_X(x) = 0$.

To visualize the conditional density, imagine $f(x, y)$ as a surface above the (x, y)-plane, and then imagine a vertical slice is taken at a constant x-value, as in the following picture.



- The resulting function $y \mapsto f(x, y)$ for fixed x is non-negative, but may not be a density because it is not normalized properly.
- The normalization factor that is needed is obtained by integrating out y from $f(x, y)$, which equals the marginal density value $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$.

Bayesian inference

- For any probability space $(\Omega, \mathcal{F}, P)$ and any events $E, F \in \mathcal{F}$ with $P(E) > 0$ and $P(F) > 0$,

$$P(E \mid F)P(F) = P(E \cap F) = P(F \cap E) = P(F \mid E)P(E)$$

and therefore,

$$P(E \mid F) = P(F \mid E)\frac{P(E)}{P(F)}\,. \tag{1.7}$$

- This result, known as *Bayes' theorem*, appears trivial but historically was a major conceptual step because it allows the "inversion" of probabilities.

Example 1.1

A patient goes to see a doctor. The doctor performs a test with 99 percent reliability–that is, 99 percent of people who are sick test positive and 99 percent of the healthy people test negative. The doctor knows that only 1 percent of the people in the country are sick. If the patient tests positive, what are the chances the patient is sick?

- The intuitive answer is 99 percent, but the correct answer is 50 percent, and Bayes' theorem gives us the relationship between what we know and what we want to know in this problem.
- Let s be the event that the patient is sick.
- What we are given is

$$P(\text{pos} \mid s),$$

ie. "the probability of testing positive given that the patient is sick;" what we want to know is

$$P(s \mid \text{pos}),$$

or "the probability of being sick given that the patient tested positive."

- According to Bayes' theorem,

$$\begin{aligned} P(\text{pos}) &= P(\text{pos} \mid s)P(s) + P(\text{pos} \mid \neg s)P(\neg s) \\ &= 0.99 * 0.01 + 0.01 * 0.99 \\ &= 0.0198 \end{aligned}$$

$$\begin{aligned} P(s \mid \text{pos}) &= \frac{P(\text{pos} \mid s)P(s)}{P(\text{pos})} \\ &= 0.99 * 0.01/0.0198 \\ &= 0.5 \end{aligned}$$

- Let's also reach the same conclusion using common sense.
- Imagine that the above story takes place in a small country, with 10,000 people.
- From the prior $P(s) = 0.01$, we know that 1 percent, or 100 people, are sick, and 9,900 are healthy.
- If we administer the test to everyone, the most probable result is that 99 of the 100 sick people test positive.
- Since the test has a 1 percent error rate, however, it is also probable that 99 of the healthy people test positive.
- If everyone who tests positive goes to the national hospital, the hospital will contain an equal number of healthy and sick patients.

- What if there are several different diseases (or variants of the same disease) that can each cause the same effect, but they have noticeably different prevalence in the population and also different levels of virulence?

- Suppose we have a list of events in a probability space,

$$E_1, E_2, \ldots, E_n,$$

  each of which (if it occurs) could possibly cause a certain effect F.
- Assume it is a comprehensive and mutually exclusive list.
- We know the probabilities of the effect

$$P(F \mid E_i)$$

  associated to each of the possible causes, and the probabilities $P(E_i)$ in the whole population.
- Suppose that an experiment is run, and we observe that the outcome ω is an element of F.

- The inner detective within us would like to know which of the possible causes is most likely, given the effect we observed.
- In other words we would like to know

$$P(E_i \mid F)$$

for all $i = 1, \ldots, n$.

- By Bayes' theorem,

$$P(E_i \mid F) = \frac{P(F \mid E_i)P(E_i)}{P(F)}$$

- The quantities in the numerator are among the things we assumed known, but what about the denominator?

- Note that if $\mathcal{P}$ is a partition and F is any event with positive probability, then sets of the form

$$\{E \cap F : E \in \mathcal{P}\}$$

form a partition of F and moreover,

$$\sum_{E \in \mathcal{P}} P(E \cap F) = P(F)$$

- Therefore, if $\{E_i\}$ is a partition,

$$P(E_i \mid F) = \frac{P(F \mid E_i)P(E_i)}{\sum_{i=1}^{n} P(E_i)P(F \mid E_i)}$$

which is now entirely in terms of the known quantities.

- Essentially the same argument that established the elementary version of Bayes' Theorem works more generally.
- We give it here for the continuous case, although the analogous statements are also true for the discrete case or for the mixed discrete-continuous case.
- Suppose we have available the marginal distribution of a random variable Y, namely its density $f_Y(y)$, and the conditional distributions of X given Y, with conditional density $f(x \mid y)$.
- If we wish to find the conditional distribution of Y given X, $f(y \mid x)$, we proceed as follows.

- First, if $f_X(x) > 0$ we have

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

- Then, plug in $f(x, y) = f(x \mid y) f_Y(y)$ to the numerator, and we have

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{f_X(x)} \tag{1.8}$$

- Since the left hand side satisfies

$$\int_{\mathbb{R}} f(y \mid x)\, dy = 1$$

by the properties of probability densities, the right side must also integrate to 1, with respect to y.

- Hence the denominator $f_X(x)$ is given by the integral, over y, of the numerator.

Statistical Models

- Most models contain one or more parameters

$$\theta_1, \theta_2, \ldots$$

that require numerical values before the model can make concrete numerical predictions.

- If there is more than one parameter, we will collect them into a vector

$$\theta \in \mathbb{R}^k.$$

- When we first begin to try to understand empirical data through the lens of a given mathematical model, the typical case is that the model parameters θ are not known, so we then ask, what process determined these parameters in the first place?
- If the process that generated θ was a *stochastic* process, then it would make sense to model θ as a random vector having its own density denoted $p(\theta)$.

- One can alter a die so that the probabilities of landing on the six sides are unequal; this is called a *loaded* die.
- Suppose that a gambler has two identical-looking dice in his pocket.
- One is fair, and the other is loaded.
- The gambler reaches into his pocket and picks one of the two dice at random, then rolls it.
- In this case, the experiment was rolling the selected die, while the parameters – the face probabilities – were themselves the result of a random process.

Definition 1.9

A (parametric) Bayesian statistical model consists of two random vectors

$$X, \Theta,$$

typically not independent of one another, where X (called the "data") takes values in $\mathbb{R}^d$, while Θ (called the "parameter vector") takes values in $\mathbb{R}^k$.

- Although in this form of the definition, it sounds as if X and Θ are on equal footing, actually they have different interpretations and different uses when it comes to modeling data.
- Realizations θ of random vector Θ are possible parameter vectors in some known mathematical model.
- Realizations of X are possible observational data points, of the same dimension as the data we are collecting and trying to analyze.
- Accordingly we refer to $\mathbb{R}^d$ as the *data space* and $\mathbb{R}^k$ as the *parameter space*.

- The conditional density

$$p(x \mid \theta)$$

is called the *likelihood*, while the marginal density

$$p(\theta)$$

is called the *prior*.

- The likelihood is actually the core of any model for understanding observational data, because it has the absolutely fundamental job of specifying how the data x and the parameter θ are linked to one another.
- It would not be too far off to say that the likelihood *is the model* and the rest of the items we discuss are details needed to work with it properly.
- The difference between a linear regression model and a neural network is all contained in the likelihood $p(x \mid \theta)$.

- Applying (1.8) to a Bayesian statistical model, one has

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{p(x)} \tag{1.9}$$

- The left-hand side $p(\theta \mid x)$ is known as the *posterior*.
- The posterior is an object of central importance in statistics because it encapsulates *what we know about the parameters, after having seen some data*.
- We are following the convention that P denotes the probability measure, as applied to events, and lowercase p denotes a density function.

- Both sides of (1.9) are valid probability densities (ie. they integrate to 1) on Θ; hence, if we do not know $p(x)$, but we can do the integral

$$\int_{\Theta} p(x \mid \theta) p(\theta) \, d\theta$$

  then we can recover $p(x)$ – it is just a normalization factor.
- For this reason, (1.9) is most easily remembered as follows:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{normalization}}$$

- Also note that, like many probability calculations, (1.9) is more convenient in log-space:

$$\log p(\theta \mid x) = \log p(x \mid \theta) + \log p(\theta) - \log p(x)$$

- The usual Gaussian density becomes a parabola in log-space, so we can see how looking at Bayes' theorem in log space could be quite convenient.

- Ordinary multivariate linear regression, suitably interpreted, provides an important example of a Bayesian statistical model.
- Typically one explains multivariate linear regression as a model of the form

$$\begin{aligned} y &= \theta_1 a_1 + \theta_2 a_2 + \cdots + \theta_k a_k + \epsilon \\ &= \boldsymbol{\theta} \cdot \boldsymbol{a} + \epsilon \end{aligned}$$

  where

$$\epsilon \sim N(0, \sigma^2)$$

  is the residual, assumed to follow a normal distribution and to be independent for different samples.
- The parameters $\boldsymbol{\theta}$ are called the *coefficients*, and there is one vector of coefficients which is assumed to apply across all samples.
- The numbers $\boldsymbol{a}$ are called *attributes* or *features*, and are usually different for each sample that is collected.

- Let's understand the ordinary linear model in the language of definition 1.9.
- Fix a list of n objects, indexed by

$$i = 1, \ldots, n,$$

such that the i-th object has a known, non-random, k-dimensional vector

$$\boldsymbol{a}_i \in \mathbb{R}^k$$

containing its attributes.
- The "data" as in definition 1.9 consists of a vector

$$\boldsymbol{y} \in \mathbb{R}^n$$

of n responses $y_1, \ldots, y_n$, and the parameter space is $\mathbb{R}^k$.
- For later use, stack the n distinct attribute vectors $\boldsymbol{a}_i$ into an $n \times k$ matrix A, so that $\boldsymbol{a}_i$ is the i-th row of A.

- To finish specifying the model, we must specify the likelihood

$$p(\boldsymbol{y} \mid \theta)$$

  and the prior $p(\theta)$.
- In ordinary least squares, we assume the residuals are independent and identically distributed normal variates.
- This means that the likelihood $p(\boldsymbol{y} \mid \theta)$ is a multivariate normal density with mean $A\theta$ and covariance matrix $\sigma^2 I$ where I is the $n \times n$ identity matrix and $\sigma > 0$ as before.
- Explicitly,

$$p(\boldsymbol{y} \mid \theta) = \frac{\exp\left(-\frac{1}{2}\|\boldsymbol{y} - A\theta\|^2/\sigma^2\right)}{(2\pi\sigma^2)^{n/2}}$$

- Various choices of the prior are possible, and the most common (or most analytically-tractable) choices have been given special names in the statistics literature.
- If we take the prior $p(\theta)$ to be another normal distribution, then the overall model is called *ridge regression*.
- The random variable Θ has a Laplace$(0, b)$ distribution if its probability density function is

$$\frac{1}{2b} \exp\left(-\frac{|\theta|}{b}\right)$$

- If we choose a Laplace prior, the Bayesian linear regression is called *the lasso*.

Bayesian billiards

- Thomas Bayes' original essay in 1763 contained an example, still relevant today, which is vaguely reminiscent of the game of billiards (aka pool).

Example 1.10 (Bayes (1763))

A ball A is rolled along the unit interval $[0, 1]$*, with a uniform probability of stopping at any point. It stops at a point* θ *between 0 and 1, and is not moved subsequently. A second ball B is then rolled n times under the same assumptions. Let x denote the number of times the ball B stopped before passing A. Without knowing* θ *but given x, what inference can we make on* θ*?*

- Once θ is known, stopping to the left versus the right of θ is conceptually like flipping a biased coin with $p(\text{heads}) = \theta$, although good luck finding an actual coin which is biased in this way.
- An experiment with two outcomes, usually called "success" and "failure" and conveniently denoted by 1 or 0, is called a *single Bernoulli trial*.

- Using combinatorics, we can prove that the probability of getting exactly x successes in n independent Bernoulli trials is given by

$$p(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

  for $x = 0, 1, 2, \ldots, n$.
- The distribution having this probability mass function is called the *Binomial distribution* due to the presence of the binomial coefficient.

- Therefore, in Example 1.10, $x \,|\, \theta \sim \mathrm{B}(n, \theta)$.
- Bayes' theorem tells us

$$p(\theta \,|\, x) \propto p(x \,|\, \theta) p(\theta). \tag{1.10}$$

- By assumption θ is uniform on $[0, 1]$, so $p(\theta) = 1$.
- Hence the left-hand side of (1.10), which is a density in θ, is proportional to

$$\binom{n}{x} \theta^x (1 - \theta)^{n-x} \tag{1.11}$$

  which we emphasize is now being considered as a function of $\theta \in [0, 1]$.
- What density is that?

- The *beta distribution* Beta(α, β) is a continuous distribution on $[0, 1]$, the space where θ lives, which has probability density given by

$$f_{\alpha,\beta}(\theta) = \text{constant} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{1.12}$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{1.13}$$

  where the constant is determined by integration.
- Therefore what we have in (1.11) must be a beta distribution, with $\alpha = 1 + x$ and $\beta = 1 + n - x$.

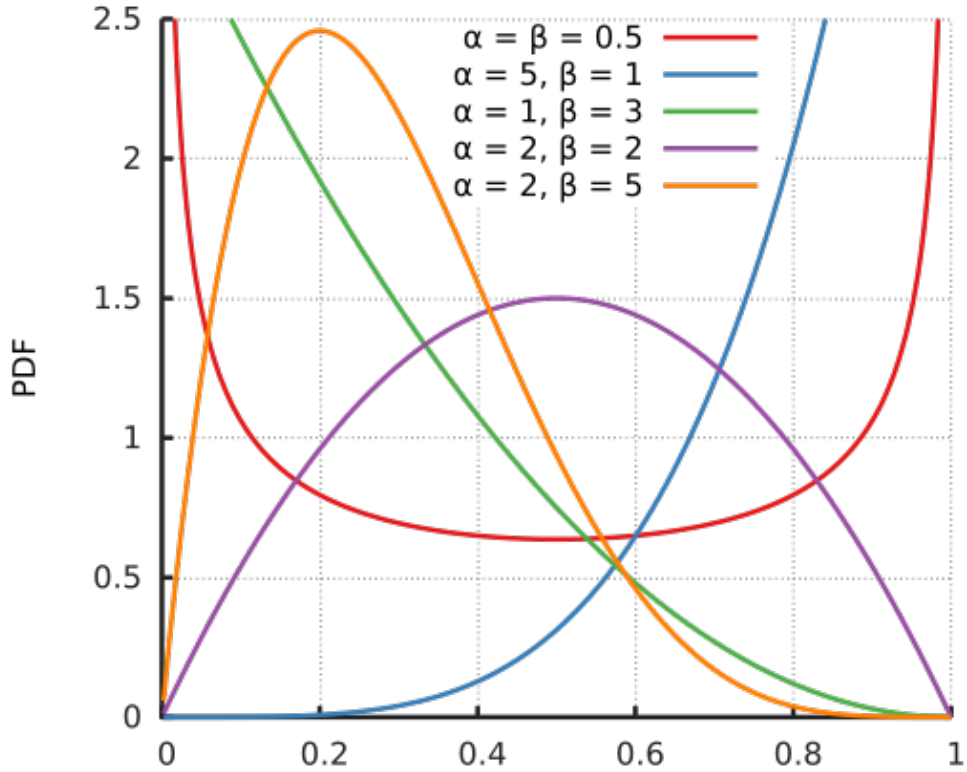


Figure: Probability density function (pdf) of Beta(α, β) for various values of α, β.

- The prior is so named because it summarizes the information that was known about the parameter *before* observation of the data presently being considered.
- That information is not necessarily a subjective opinion; it could itself simply be what was learned from earlier observations, or, as in Example 1.10, from the assumed physics of rolling the first ball.
- Bayesian analysis entails doing inference about the parameter, conditional on the data x, and the Bayesian paradigm gives a proper probabilistic meaning to this conditioning by allocating a probability distribution to θ.

Point Estimates

Many of you will have seen the standard way to form point estimates in statistics: maximum-likelihood estimation.

- The maximum-likelihood estimate of the parameter θ is defined as

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \log p(x \mid \theta)$$

- The log is inserted for convenience, since many likelihood functions contain the exponential function, and moreover, joint probabilities of independent events factorize as products, and the log converts all such products to sums.

- In Bayesian analysis, all inference is based around the posterior.
- The Bayesian analog of the MLE is the maximum a-posteriori estimator:

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname*{argmax}_{\theta} \log p(\theta \mid x) \\ &= \operatorname*{argmax}_{\theta} \{\log p(x \mid \theta) + \log p(\theta)\}\end{aligned}$$

- Hence the function being maximized is the log-likelihood augmented by an additional penalty term $\log p(\theta)$.
- Often, this term acts as a regularizer.

Prediction

- An important property of any statistical model is its ability to make predictions which apply outside of the data that was used to train the model.
- Some say *machine learning* **is** *prediction*.

- Let's consider prediction from the original Bayes model, which has binomial likelihood and, we derived, a beta distribution as the posterior.
- The probability that the next occurrence is a success is

$$p(\text{success} \,|\, x) = \int_0^1 p(\text{success} \,|\, \theta, x) p(\theta \,|\, x)\, d\theta \qquad (1.14)$$

- Recall that in this model, θ equals the success probability, so trivially,

$$p(\text{success} \,|\, \theta, x) = \theta.$$

- Moreover, by (1.11), we know that $p(\theta \,|\, x)$ is a beta distribution, with

$$\alpha = 1 + x \text{ and } \beta = 1 + n - x.$$

- Therefore, the probability (1.14) that the next flip is a success is

$$\int_0^1 \theta\, f_{\alpha,\beta}(\theta)\, d\theta, \quad \alpha = 1 + x,\ \beta = 1 + n - x$$

  where the beta density $f_{\alpha,\beta}$ is as given by (1.12).
- Note that the forecast is just the posterior mean.
- The mean of Beta(α, β) is

$$\alpha/(\alpha + \beta)$$

  so plugging in

$$\alpha = 1 + x \text{ and } \beta = 1 + n - x$$

  immediately gives

$$p(\text{success} \,|\, x) = \frac{x+1}{n+2}$$

Conjugate priors

- Suppose that, instead of the uniform prior implied by Bayes' example, we take Beta(α, β) as our prior.
- In other words, we take as prior the following:

$$p(\theta \mid \alpha, \beta) \;\propto\; \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The posterior is then

$$\begin{aligned} p(\theta \mid x) &\propto p(x \mid \theta) p(\theta) \\ &\propto \theta^{x}(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\sim \text{Beta}(\alpha + x, \beta + n - x). \end{aligned}$$

- Thus the posterior is of the same form as the prior, but with different parameters.
- The beta distribution is said to be a *conjugate prior* for this likelihood.
- We also say the prior has been "updated" by the data, x in this case.

- If more data come in, the posterior is suitable for use as a prior in the next round of updating.
- The order in which i.i.d. observations are collected does not matter, and updating the prior one observation at a time, or all observations together, does not matter.
- This kind of model lends itself well to inference with big data sets, since all of the relevant information about the full data set – no matter how large it may be – is contained in a small number of parameters.

- The Beta-Binomial model generalizes to the case in which each observation is one of k possible outcomes.
- For example, suppose we are interested in how frequently various English words appear in written text.
- Then there are k possible words, and θ_j for $j = 1, \ldots, k$ represents the frequency of the j-th word.
- We would like to infer whatever we can concerning the vector

$$\theta = (\theta_1, \ldots, \theta_k)$$

  from some examples x.
- Happily, this example lends itself to a conjugate prior, as we now explain.

- If n_i counts the number of observations of the i-th outcome, then

$$p(n_1, \ldots, n_k \,|\, \theta) \;\propto\; \prod_{j=1}^{k} \theta_j^{n_j}, \quad \theta_j \geq 0, \quad \sum_{j=1}^{k} \theta_j = 1. \quad (1.15)$$

- The conjugate prior associated to the likelihood (1.15) is the *Dirichlet distribution*:

$$p(\theta \,|\, \alpha) \;\propto\; \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}.$$

- The resulting posterior is also from the Dirichlet family, but with parameters $\alpha_j + n_j$.

Confidence Intervals

- For ease of explanation, we use notation appropriate to one-dimensional data

$$x \in \mathbb{R},$$

but all of these concepts can be generalized to multivariate data.

Definition 1.11

An interval $[\ell, u]$ *has Bayesian coverage* a *if* $p(\ell < \theta < u \,|\, x) = a$.

- The interpretation of this interval is that it describes your information about the location of θ after you have observed $X = x$.
- This is different from the frequentist interpretation of coverage probability, which describes the probability that the interval will cover the true value before the data are observed.

- A "Bayesian confidence interval" is an interval with some pre-specified Bayesian coverage, often 0.95 or 0.99.
- Clearly there can, in general, be many intervals with the same Bayesian coverage.
- One simple way to obtain a Bayesian confidence interval is to use posterior quantiles.
- To make a $100 \times (1-\alpha)\%$ quantile-based confidence interval, find $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

$$p(\theta < \theta_{\alpha/2} \,|\, x) = \frac{\alpha}{2} \quad \text{and} \quad p(\theta > \theta_{1-\alpha/2} \,|\, x) = \frac{\alpha}{2}\,.$$

- So the numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ, so one can see easily

$$p(\theta \in [\theta_{\alpha/2}, \theta_{1-\alpha/2}] \,|\, x) = 1 - \alpha\,.$$

Example 1.12

Suppose out of $n = 10$ conditionally independent draws of a binary random variable we observe $x = 2$ "heads" or "ones." Under a uniform prior for θ, the posterior is

$$\theta \,|\, x = 2 \sim \mathit{Beta}(1 + 2, 1 + 8).$$

A 95% posterior confidence interval can be obtained from the 0.025 and 0.975 quantiles of this beta distribution, computed in R as `qbeta(c(0.025, 0.975), 3, 9)`*, yielding the interval*

$$I_1 = [0.06022, 0.51776].$$

So the posterior probability that $\theta \in I_1$ is 95%.

- Quantile-based confidence intervals can be rather counter-intuitive for distributions that are either skewed or multi-modal.

Consider the following example:

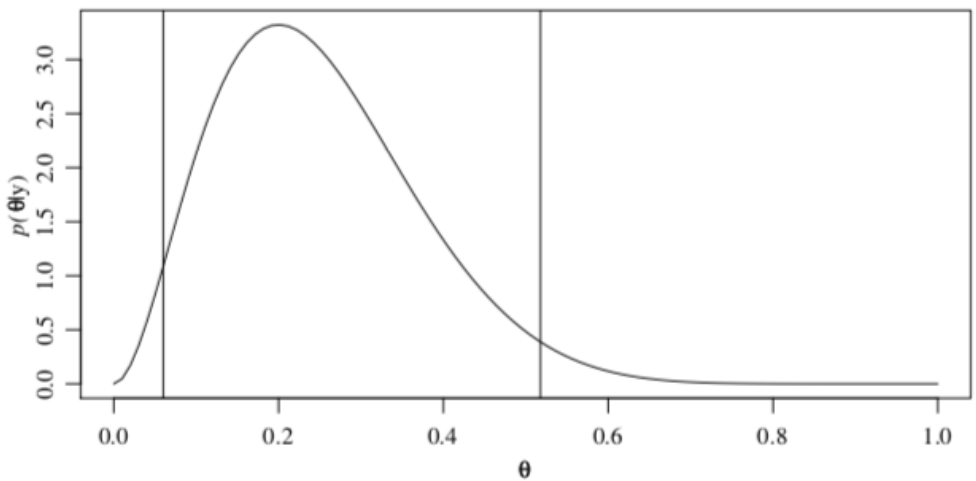


Figure: A beta posterior distribution, with vertical bars indicating a 95% quantile-based confidence interval. Notice that there are θ-values outside the quantile-based interval that have higher probability density than some points inside the interval.

- This suggests a more restrictive type of interval: the HPD interval.

Definition 1.2

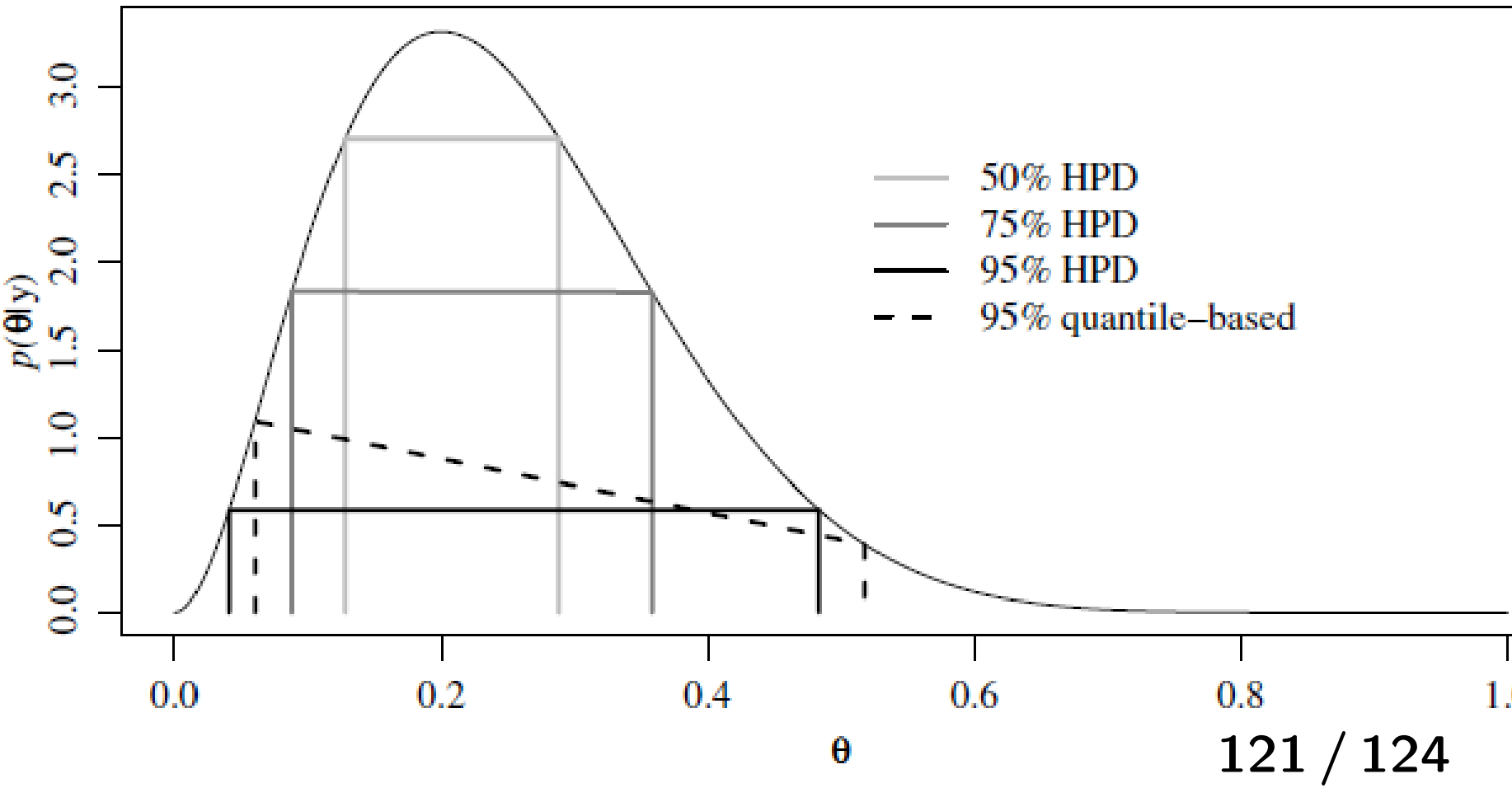
A $100 \times (1-\alpha)\%$ *HPD region* consists of a subset S of the parameter space $S \subseteq \Theta$ such that:

1. $p(\theta \in S \,|\, x) = 1 - \alpha$
2. If $\theta_a \in S$ and $\theta_b \notin S$ then $p(\theta_a \,|\, x) > p(\theta_b \,|\, x)$.

- This definition generalizes in a straightforward way to multi-dimensional parameter vectors.
- All points in an HPD region have a higher posterior density than points outside the region.
- Note that an HPD region might not be an interval if the posterior density is multimodal.

- The idea is as follows: gradually move a horizontal line down across the density, including in the HPD region all θ-values having a density above the horizontal line.

Stop moving the line down when the posterior probability of the region reaches $1 - \alpha$.



For Example 1.12, the 95% HPD region can be calculated in R with the command

```
require('HDInterval');
hdi(qbeta, 0.95, shape1 = 3, shape2 = 9)
```

- which yields $[0.04056, 0.48372]$.

Unlike the quantile-based confidence interval, the HPD region does not include the possibility that the coin is fair.

 Bayes, Thomas (1763). “An essay towards solving a problem in the doctrine of chances.”. In.