# GPH-GU2372 GPH-GU3372 Applied Bayesian Analysis in Public Health

## Lecture 1: Introduction to Bayesian Inference

Hai Shu, PhD

09/12/2022

# Administrative details

- Instructor: Hai Shu, PhD, hs120@nyu.edu
- Office hours: 6:05pm–6:35pm @ 708 Broadway, Room 759
- Course Assistant: Sooyoung Kim, sk9076@nyu.edu
  If any question, ask the CA first, then me.
- Textbook: *"A First Course In Bayesian Statistical Methods"*
  by Peter D. Hoff, Springer.
- Class Attendance (10% of final grade), 6 HW Assignments(30%),
  Midterm Exam (25%), and Final Exam (35% for 2372; 20% for 3372)
  Final project (only for 3372; 15%)

# Scope of the course and topics

- This course is designed to introduce graduate students to Bayesian data analysis.
- Emphasis is on the practical use of Bayesian inference in public-health/medical problems.
- Data analysis will be carried out using R.
- The following topics will be covered:
  - *Overview and basics of Bayesian inference*
  - *Bayesian analysis of basic models*
  - *Bayesian computing: Markov Chain Monte Carlo (MCMC), Metropolis-Hastings algorithms*
  - *Hierarchical models*
  - *Bayesian (generalized) linear (mixed) models*
  - *Probit regression*
  - *Bayesian model selection*

# Introduction

- Some review of probability
- Basics of Bayesian inference
- Advantages of Bayesian methods

# Review of probability

- What is probability? Formally, we think of probability as a function that assigns a real number to an event.
- We consider an experiment and we call the set of all its possible outcomes a sample space $\mathcal{S}$. A subset $E$ of the sample space $\mathcal{S}$ is called an event.
- A function defined on the set of all possible subsets of $\mathcal{S}$ with values in $[0, 1]$ is called a probability function $P$ if it satisfies the following conditions:
  1. $0 \leq P(E) \leq 1$.
  2. $P(\mathcal{S}) = 1$ and $P(\emptyset) = 0$.
  3. If $E$ and $F$ are two events such that $E \cap F = \emptyset$ (i.e., $E$ and $F$ are disjoint events), then $P(E \cup F) = P(E) + P(F)$.

# Review of probability

- If $E$ is an event, we denote its complement by $E^c$. From properties 2 and 3 of a probability function, we have

$$P(E^c) = 1 - P(E).$$

- Conditional probability. Let $E$ and $F$ be two events (i.e. two subsets of the sample space $\mathcal{S}$), the conditional probability $P(E|F)$ of $E$ **given** that $F$ has occurred is:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

- Joint probability. From the definition of conditional probability, we can derive a formula for the probability $P(E \cap F)$ that both events $E$ and $F$ occur:

$$P(E \cap F) = P(E|F)P(F).$$

# Review of probability

- Independent events. Two events $E$ and $F$ are said to be independent (denoted by $E \perp\!\!\!\perp F$ ) if knowing that $F$ has occurred does not change the probability of $E$ occurring. Formally,

$$E \perp\!\!\!\perp F \quad \text{if} \quad P(E|F) = P(E).$$

- Multiplication rule: If $E \perp\!\!\!\perp F$, then $P(E \cap F) = P(E)P(F)$.
- A collection of subsets $\{E_1, ..., E_k\}$ of the sample space $\mathcal{S}$ form a partition of $\mathcal{S}$ if:
  1. the events are disjoints, i.e., $E_i \cap E_j = \emptyset$ for any $i \neq j$,
  2. the union of all the events is $\mathcal{S}$, i.e., $\bigcup_{i=1}^{k} E_i = \mathcal{S}$.
- Rule of total probability: From properties 2 and 3 of a probability function, if $\{E_1, ..., E_k\}$ is a partition of $\mathcal{S}$, then

$$\sum_{i=1}^{k} P(E_i) = 1.$$

# Review of probability

- Rule of marginal probability: Let $A$ be an event and let $\{E_1, ..., E_k\}$ be a partition of $\mathcal{S}$. Then,

$$P(A) = \sum_{i=1}^{k} P(A \cap E_i) \overset{\text{joint prob.}}{=} \sum_{i=1}^{k} P(A|E_i)P(E_i).$$

- The rule of marginal probability is often interpreted as: reinterpreting the probability of the event $A$ by *"marginalizing"* over the events in the partition.

# Review of probability

- Application. Consider the following table with probabilities of survival and stage of disease:

| | $E$=Early stage | $E^c$=Late stage |
|---|---|---|
| $F$=Survive | 0.72 | 0.02 |
| $F^c$=Die | 0.18 | 0.08 |

- What are the joint probability $P(F \cap E^c)$, the marginal probability $P(E^c)$, and the conditional probability $P(F|E^c)$?

# Bayes' Theorem

- By the definition of conditional and joint probability, Thomas Bayes (170?-1761) formulated Bayes' Theorem:
  For two events $A$ and $D$, if $P(D) \neq 0$, then the conditional probability of $A$ **given** $D$ is:
  $$P(A|D) = \frac{P(D|A)P(A)}{P(D)}.$$

- Bayes' theorem is at the basis of Bayesian statistics and it provides a model for learning.
  Suppose we have an initial or "a priori" belief about the probability that the event $A$ occurs. Suppose we observe some data, i.e., we observe the event $D$ occuring. Then by Bayes' theorem, we can "a posteriori" update the probability that the event A occurs in light of the data. This is $P(A|D)$!

# Bayes' Theorem

- If $A$ and $D$ are events (i.e., subsets of the sample space $\mathcal{S}$) and $\{E_1, ..., E_k\}$ is a partition of $\mathcal{S}$, then Bayes' theorem can be re-expressed as:

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

$$\overset{\text{rule of marginal probability}}{=} \frac{P(D|A)P(A)}{\sum_{i=1}^{k} P(D|E_i)P(E_i)}$$

- The denominator is the marginal probability of event D.

# Examples of application of Bayes' Theorem

- Bayes' theorem is often used in diagnostic tests for cancer.
- Example: A young person was diagnosed as having an extremely rare cancer in young people. He was very upset. A friend told him that it was probably a mistake. His friend reasoned as follows. No medical test is perfect: there are always incidences of false positives and false negatives.

  Let $C$ stand for the event that he has cancer and $+$ denote the event that an individual responds positively to the test.

  Assume that:
  - $P(C) = \frac{1}{1{,}000{,}000} = 10^{-6}$
  - $P(+|C) = 0.99$
  - $P(+|C^c) = 0.01$

  Find the probability that the young man has cancer given that the test has a positive response.

# Examples of application of Bayes' Theorem

- Before the cost of DNA testing became accessible, Bayes' theorem was also used extensively in legal cases of disputed paternity where the probability that the alleged father was in fact the father was calculated given blood test evidence.

- In this class, Bayes' Theorem will provide the foundation upon which we will develop the Bayesian paradigm for parametric inference.

- We will work with random variables and parametric models to describe, respectively, their probability mass function (pmf) or their probability density function (pdf).

# Random variables

- A (real valued) random variable $Y$ is a function from the sample space $\mathcal{S}$ to a subset, $\mathcal{Y}$, of $\mathbb{R}$ (the set of real numbers).
- $\mathcal{Y}$ can now be considered the sample space for the random variable $Y$.
- If the sample space $\mathcal{Y}$ of $Y$ is countable, i.e., $\mathcal{Y}$ can be written as $\mathcal{Y} = \{y_1, y_2, y_3, ...\}$, then $Y$ is called a discrete random variable.

  If $\mathcal{Y}$ is not countable and is *"equal"* to $\mathbb{R}$, then $Y$ is called a continuous random variable.

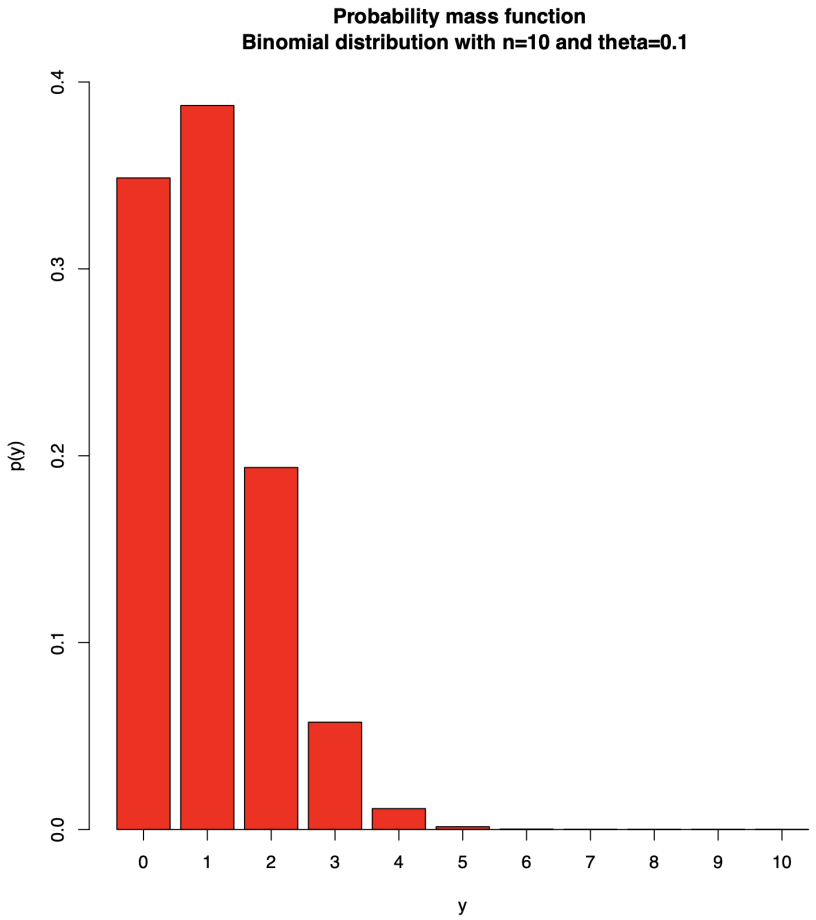
# Discrete random variables

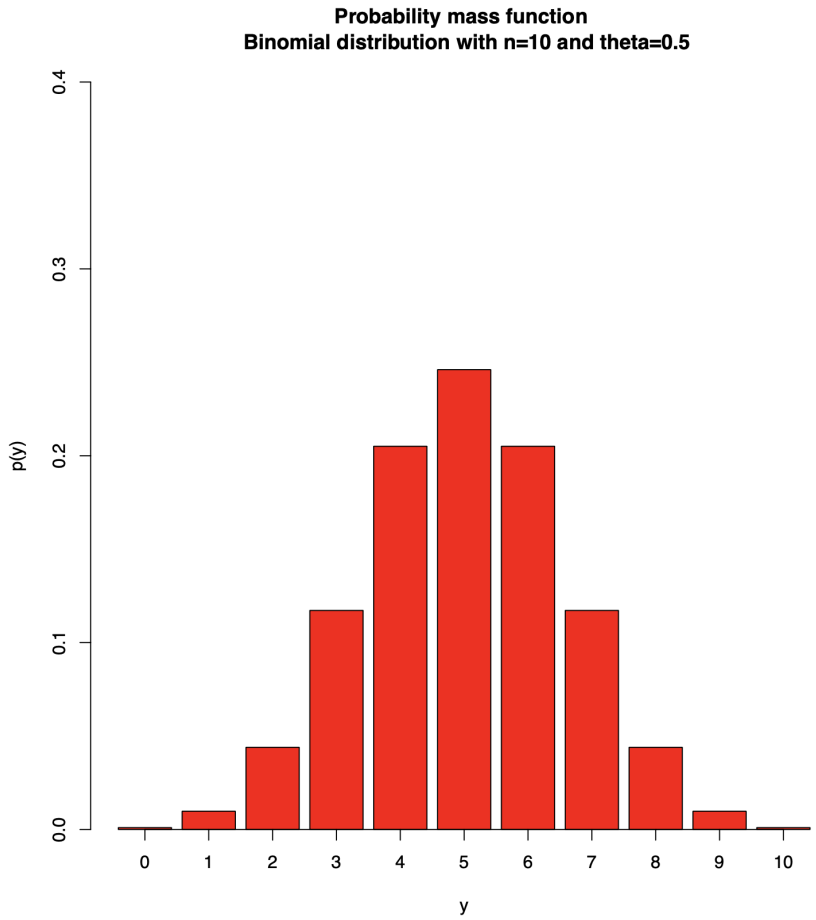
- If $Y$ is a discrete random variable, then any event, that is, any subset of $\mathcal{Y}$, can be obtained as disjoint union of *basic sets* of the form $E = \{y\}$, for $y \in \mathcal{Y}$.
- In light of this, using the rules of probability, if we want to know the probability of any event $A \subset \mathcal{Y}$, it is enough that we know the probability for the basic sets $E = \{y\}$.
- If $Y$ is a discrete random variable, the function, $p(y)$, that expresses the probability for the basic sets $E = \{y\}$ is called the probability mass function (pmf) and it satisfies the following conditions:
  1. For any $y \in \mathcal{Y}$, $0 \leq P(\{y\}) = p(y) \leq 1$
  2. $\sum_{y \in \mathcal{Y}} p(y) = 1$
- Then, e.g., if $A = \{y_1, y_2\}$, the probability $P(A)$ of $A$ is given by:

$$\begin{aligned} P(A) &= P(\{y_1, y_2\}) = P(\{y_1\} \cup \{y_2\}) \\ &= P(\{y_1\}) + P(\{y_2\}) = p(y_1) + p(y_2). \end{aligned}$$

# Discrete random variables: parametric models

- For certain discrete random variables $Y$, the pmf can be expressed as a function of certain parameters. In this case, we say that we are adopting a parametric model for the pmf of $Y$.
- Examples: $Y$ is the number of successes in $n = 10$ trials, then $p(y)$ can be modeled using a Binomial distribution$(n = 10, \theta)$.



# Discrete random variables: parametric models

- Other examples:
  - $Y$ is the number of shots I miss before making $r = 5$ baskets $\Longrightarrow p(y)$ is a Negative binomial distribution$(r = 5, \theta)$
  - $Y$ is the number of cars that between 8 and 9 AM pass in front of a given stop sign without making a full stop $\Longrightarrow p(y)$ is a Poisson distribution$(\theta)$

# Continuous random variables

- If $Y$ is a continuous random variable, then any event, that is, any subset of $\mathcal{Y} = \mathbb{R}$, can be obtained as union or intersection of basic sets of the form $E = (a, b]$ or $E = (-\infty, b]$ or $E = (a, \infty)$, for $a, b \in \mathbb{R}$.
- If $Y$ is a continuous random variable, the function, $F(y)$, that expresses the probability for the basic sets of the type $E = (-\infty, y]$ is called the cumulative distribution function:
$F(y) = P((-\infty, y]) = P(Y \leq y)$.
- The cumulative distribution function (cdf) $F(y)$ satisfies the following conditions:
  1. $F(-\infty) = 0, F(\infty) = 1$
  2. $F$ is non-decreasing, that is: $F(a) \leq F(b)$ if $a < b$.
- From the cdf $F$, using the rules of probability, probabilities of other events can be derived:
  1. For any $a \in \mathbb{R}$, $P((a, \infty)) = P(Y > a) = 1 - P(Y \leq a) = 1 - F(a)$
  2. For any $a, b \in \mathbb{R}$, $P(a < Y \leq b) = F(b) - F(a)$

# Continuous random variables

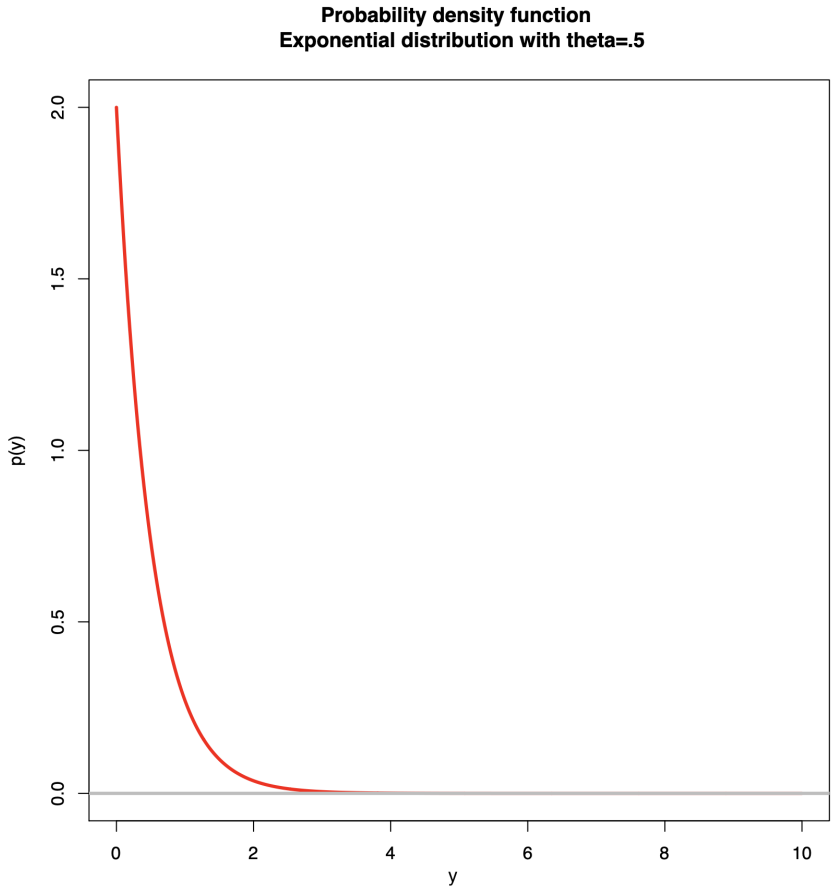
- From a fundamental measure theory theorem, we know that for any cdf $F$, there exists a nonnegative function $p(y)$ such that:

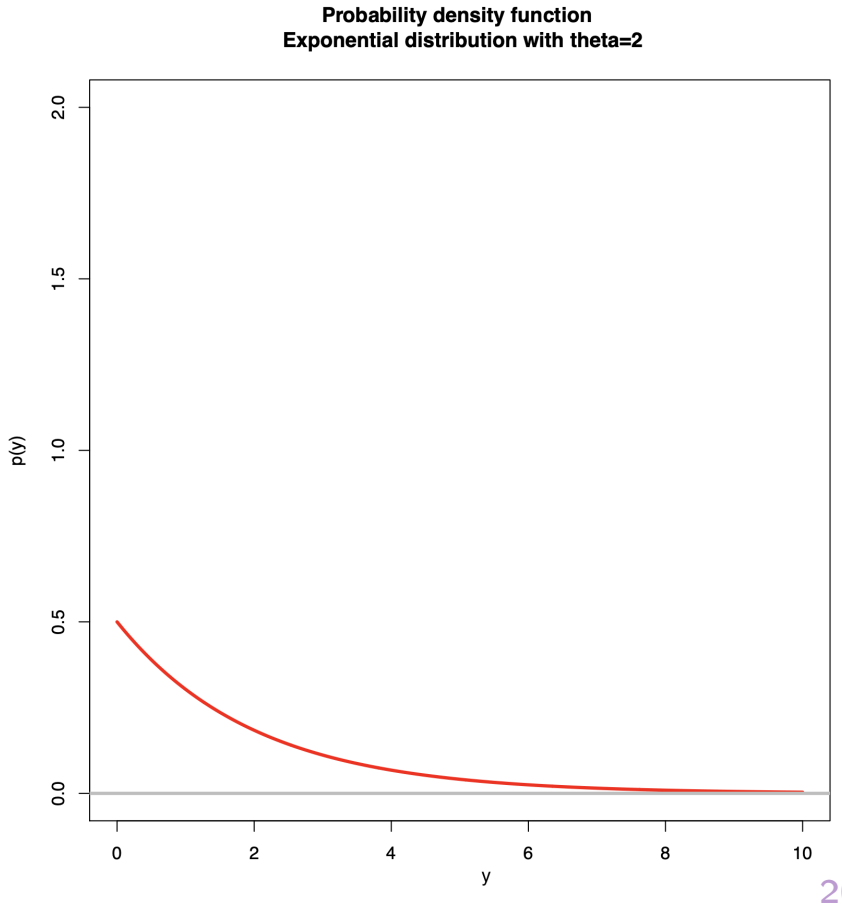
$$F(a) = \int_{-\infty}^{a} p(y)dy$$

- The function $p(y)$ is called the probability density function (pdf) of the continuous random variable $Y$ and satisfies analogous conditions to those for probability mass functions:
  1. For any $y \in \mathbb{R}, 0 \leq p(y)$
  2. $\int_{y \in \mathbb{R}} p(y)dy = 1$
- Probabilities for any event $A \subset \mathbb{R}$ can be derived from the pdf: $P(A) = \int_A p(y)dy$.

# Continuous random variables: parametric models

- For certain continuous random variables $Y$, it is possible to express the pdf as a function of certain parameters, and adopt a parametric model for the pdf of $Y$.
- Examples: $Y$ is the lifetime (in years) of the battery of a certain brand and model of laptops. Then, $p(y)$ can be modeled using an exponential distribution($\theta$).



# Continuous random variables: parametric models

- Other examples:
  - $Y$ is the height of US women in a certain age range $\Longrightarrow p(y)$ is a Normal distribution$(\theta_1, \theta_2)$
  - ...

# Statistical inference

- The goal of statistical inference is to learn about the general population by taking a subset from that population.
- The steps in classical statistical parametric inference are:
  1. Collect a random sample from the population.
  2. Determine a probability model for the data $y_1, \ldots, y_n$ obtained (e.g., we are collecting data on the height of US women: each women's height can be thought as a random variable with a normal distribution). This is called the sampling model.
  3. The sampling model will depend on a set of parameters, $\theta$: $p(y_i|\theta), i = 1, \ldots, n$.
  4. Using the sampling model, construct the likelihood function $L(\theta; y)$.
  5. Determine the maximum likelihood estimate (MLE) of $\theta$ by maximizing the likelihood function $L(\theta; y)$.
- In the classical or frequentist approach to statistical inference, the parameters of a sampling model are assumed to be constant but **unknown**.

# Bayesian paradigm for inference

- The Bayesian paradigm allows to incorporate subjective prior beliefs regarding the parameter(s) $\theta$ of the sampling model in our learning procedure.
- In Bayesian statistics, before we observe the data, we have some belief regarding possible values, or range of values for the set of parameters $\theta$. This knowledge is expressed in the form of a prior distribution.
- Since we are assigning a distribution to $\theta$, we are now assuming that $\theta$ is a random variable (univariate if $\theta$ is just one parameter, multivariate if $\theta$ consists of more than one parameter).
- Since $\theta$ is a random variable, there is a sample space associated with $\theta$. We call this the parameter space and we indicate it with $\Theta$.
  The parameter space $\Theta$ is the set of all possible values for $\theta$.

# Bayesian paradigm for inference

- Once data $y_1, \ldots, y_n$ has been collected, we specify a sampling model for the data: $p(y_i|\theta), i = 1, \ldots, n$. This represents the probability of observing $y_i$ if we knew $\theta$.

- In light of the data observed, using Bayes' theorem we update our belief about $\theta$ and we derive the posterior distribution $p(\theta|y_1, \ldots, y_n)$.

# Advantages of Bayesian methods

- Interpretation: Having a distribution for the unknown parameter(s) $\theta$ makes it easier to understand what a point estimate and a confidence interval mean.
  Example: suppose the parameter of interest here is the population mean $\theta$. We have collected data and we build a 95% confidence interval for $\theta$. This is given by:

$$\bar{y} \pm 1.96 \frac{s}{\sqrt{n}}$$

  After the sample is collected and the interval is created, it either contains $\theta$ or it does not. So, **95% is NOT the probability that $\theta$ falls in the interval.** For a frequentist, 95% is not a coverage probability, it is just a tag that informs how the interval performs over the long haul. By contrast, Bayesian confidence intervals, called credible sets, convey the posterior probability that the parameter falls in the interval.

# Advantages of Bayesian methods

- Likelihood Principle: Bayesian inference obeys the likelihood principle. The likelihood principle states that if two sampling models yield proportional likelihoods for $\theta$, then inference about $\theta$ should be identical under the two models.
  Frequentist inference does NOT always obey the likelihood principle!

  Example: suppose we toss a coin 12 times and we observe 9 heads and 3 tails. We want to test the hypothesis that the probability of heads, $\theta$, is $H_0 : \theta = 0.5$ versus $H_1 : \theta > 0.5$.
  Suppose that we set up two sampling models for these data:
  1. $Y$, the number of heads, is a Binomial$(n = 12, \theta)$
  2. $Y$, the number of heads observed before completing the experiment (which is: getting 3 tails), is a Negative Binomial$(r = 3, 1 - \theta)$

- The two likelihoods are proportional but they lead to two different p-values!!!!

# Advantages of Bayesian methods

- Bayesian theorem provides a rational method for learning.

  Suppose you collect data $y_1, \ldots, y_n$ and compute your posterior distribution $p(\theta|y_1, \ldots, y_n)$. Suppose that later on you collect an additional observation $y_{n+1}$. Then, the posterior distribution $p(\theta|y_1, \ldots, y_n)$ could be used as a prior for the current Bayesian data analysis.

# Advantages of Bayesian methods

- Bayesian inference does not require large sample theory.

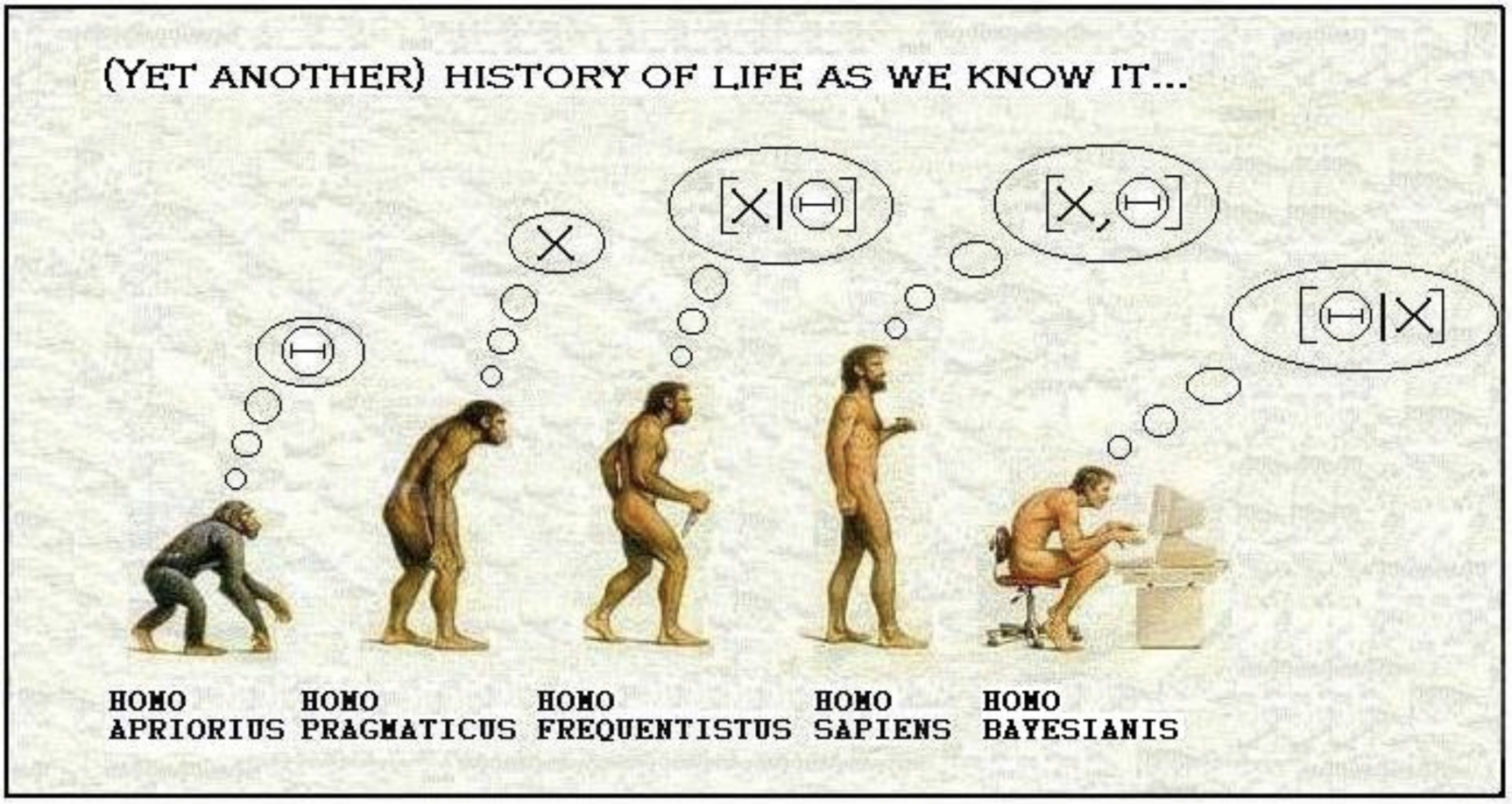
  Bayesian methods do not require asymptotics for valid inference. Small sample Bayesian inference proceeds in the same way as if one had a large sample.

- Bayesian inference often has frequentist inference as a special case.

  One can often obtain frequentist answers by choosing a uniform prior for the parameters. In this case, frequentist answers can be obtained from the posterior distribution (the MLE is then the posterior mode).

# Reading

- If you are interested, you can read "What is Bayesian statistics and why everything else is wrong" by Michael Lavine.
  `https://people.math.umass.edu/~lavine/whatisbayes.pdf`



# Next

- Joint densities
- Bayes' theorem for density functions
- Independent and exchangeable random variables
- De Finetti's theorem
- Prior distributions
- Binomial model

# Joint densities: discrete random variables

- Suppose $Y_1$ and $Y_2$ are two discrete random variables with sample spaces the countable sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$.
- We can consider the two random variables jointly. Then, the sample space is $\mathcal{Y}_1 \times \mathcal{Y}_2$ and events are sets of the form $A \times B = \{(a, b) : a \in A, b \in B\}$ with $A \subset \mathcal{Y}_1$ and $B \subset \mathcal{Y}_2$.
- As in the case of a single random variable, the probability of any event $A \times B \subset \mathcal{Y}_1 \times \mathcal{Y}_2$ can be derived using the rules of probability from the joint density $p_{Y_1,Y_2}(y_1, y_2)$:

$$p_{Y_1,Y_2}(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})$$

- The marginal density $p_{Y_1}(y_1)$ of $Y_1$ can be computed from the joint density $p_{Y_1,Y_2}(y_1, y_2)$:

$$\begin{aligned} p_{Y_1}(y_1) &= P(\{Y_1 = y_1\}) \overset{\text{rule of marginal probability}}{=} \sum_{y_2 \in \mathcal{Y}_2} P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &= \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1,Y_2}(y_1, y_2) \end{aligned}$$

# Joint densities: discrete random variables

- We can also define the conditional density $p_{Y_2|Y_1}(y_2|y_1)$ of $Y_2$ **given** that $Y_1 = y_1$:

$$p_{Y_2|Y_1}(y_2|y_1) = P(\{Y_2 = y_2\}|\{Y_1 = y_1\}) = \frac{P(\{Y_2=y_2\}\cap\{Y_1=y_1\})}{P(\{Y_1=y_1\})}$$
$$= \frac{p_{Y_1,Y_2}(y_1,y_2)}{p_{Y_1}(y_1)}$$

- The conditional density $p_{Y_2|Y_1}(y_2|y_1)$ of $Y_2$ **given** that $Y_1 = y_1$ is a probability mass function. Thus, $\sum_{y_2\in\mathcal{Y}_2} p_{Y_2|Y_1}(y_2|y_1) = 1$.
- Similarly, one can define the marginal density $p_{Y_2}$ of $Y_2$ and the conditional density $p_{Y_1|Y_2}$ of $Y_1$ given $Y_2$.
- Thus, from $p_{Y_1,Y_2}(y_1, y_2)$, one can derive $p_{Y_1}(y_1)$, $p_{Y_2}(y_2)$ and $p_{Y_1|Y_2}(y_1|y_2)$, $p_{Y_2|Y_1}(y_2|y_1)$.
  From $p_{Y_1}(y_1)$ and $p_{Y_2|Y_1}(y_2|y_1)$ (or from $p_{Y_2}(y_2)$ and $p_{Y_1|Y_2}(y_1|y_2)$), one can derive $p_{Y_1,Y_2}(y_1, y_2)$.
  However, from $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$, it is **NOT** possible to derive $p_{Y_1,Y_2}(y_1, y_2)$ unless $Y_1$ and $Y_2$ are independent!

# Joint densities: continuous random variables

- Suppose now that $Y_1$ and $Y_2$ are two continuous random variables with sample spaces $\mathcal{Y}_1$ "=" $\mathbb{R}$ and $\mathcal{Y}_2$ "=" $\mathbb{R}$.
- Now, events are subsets of the form $A \times B \subset Y_1 \times Y_2$ with $A \subset Y_1$ and $B \subset Y_2$ and their probabilities can be derived using the rules of probability from those of the basic sets $(-\infty, a] \times (-\infty, b]$ with $a, b \in \mathbb{R}$. This probabilities are given by the joint cumulative distribution function (cdf) $F_{Y_1,Y_2}(a, b)$

$$\begin{aligned} F_{Y_1,Y_2}(a,b) &= P(Y_1 \leq a, Y_2 \leq b) = P(\{Y_1 \leq a\} \cap \{Y_2 \leq b\}) \\ &= P(\{Y_1 \in (-\infty, a]\} \cap \{Y_2 \in (-\infty, b]\}) \end{aligned}$$

- As for one single random variable, it can be proved that there exists a nonnegative function, the joint probability density function (p.d.f) $p_{Y_1,Y_2}(y_1, y_2)$, such that:

$$F_{Y_1,Y_2}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2$$

# Joint densities: continuous random variables

- As for continuous random variables, from $p_{Y_1,Y_2}(y_1, y_2)$ we can derive the marginal density $p_{Y_1}(y_1)$ of $Y_1$:

$$p_{Y_1}(y_1) = \int_{-\infty}^{\infty} p_{Y_1,Y_2}(y_1, y_2) dy_2$$

- From the joint density $p_{Y_1,Y_2}(y_1, y_2)$ and the marginal density $p_{Y_1}(y_1)$, we can derive the conditional density $p_{Y_2|Y_1}(y_2|y_1)$ of $Y_2$ **given** that $Y_1 = y_1$:

$$p_{Y_2|Y_1}(y_2|y_1) = \frac{p_{Y_1,Y_2}(y_1,y_2)}{p_{Y_1}(y_1)}$$

- The conditional density $p_{Y_2|Y_1}(y_2|y_1)$ of $Y2$ **given** that $Y_1 = y_1$ is a probability density function (pdf). Hence,

$$\int_{y_2 \in \mathbb{R}} p_{Y_2|Y_1}(y_2|y_1) dy_2 = 1.$$

# Joint densities: discrete & continuous random variables

- If $Y_1$ is a discrete random variable with sample space $\mathcal{Y}_1$ and $Y_2$ is a continuous random variable with sample space $\mathcal{Y}_2$ , we can define a joint density function $p_{Y_1,Y_2}(y_1, y_2)$ through:
  - marginal density $p_{Y_1}(y_1) = P(Y_1 = y_1)$
  - conditional density $p_{Y_2|Y_1}(y_2|y_1)$

  Then:

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_2|Y_1}(y_2|y_1)p_{Y_1}(y_1).$$

  For any event $A \times B \subset \mathcal{Y}_1 \times \mathcal{Y}_2$:

$$\begin{aligned} P_{Y_1,Y_2}(A \times B) &= P_{Y_1,Y_2}(\{Y_1 \in A\} \cap \{Y_2 \in B\}) \\ &= \int_{y_2 \in B} \Big( \sum_{y_1 \in A} p_{Y_1,Y_2}(y_1, y_2) \Big) dy_2 \end{aligned}$$

- Notation: We usually denote $p_{Y_1,Y_2}(y_1, y_2)$,$p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$ simply with $p(y_1, y_2)$,$p(y_1)$ and $p(y_2)$. Similarly, we denote $p_{Y_1|Y_2}(y_1|y_2)$ and $p_{Y_2|Y_1}(y_2|y_1)$ simply with $p(y_1|y_2)$ and $p(y_2|y_1)$.

# Bayes' theorem for density functions

- Bayes' theorem for events:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{P(D|A)P(A)}{\sum_{i=1}^{k} P(D|E_i)P(E_i)}$$

- When working with random variables, the probability of any event can be obtained through the density function $p(y)$, whether a p.m.f or a p.d.f $\Longrightarrow$ we can imagine how to re-formulate Bayes' theorem for density functions.
- If $\theta$ and $Y$ are two random variables. Bayes' theorem states that the conditional density of $\theta$ given that we have observed $Y = y$ is given by:

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

  - $p(\theta)$ is called the prior probability of $\theta$: it expresses the beliefs regarding $\theta$ prior to observing the data.
  - $p(y|\theta)$ is the sampling model for the data. (It is what yields the likelihood)
  - $p(y)$ is the marginal density of the data.
- The conditional density $p(\theta|y)$ is called the posterior density of $\theta$.

# Bayes' theorem for density functions

- Using the results for joint densities, the marginal density $p(y)$ of the data can be expressed as:

$$p(y) = \int_{\theta \in \Theta} p(y, \theta) d\theta = \int_{\theta \in \Theta} p(y|\theta) p(\theta) d\theta$$

  where $\Theta$ is the parameter space.
- Substituting the expression for $p(y)$, Bayes' theorem for density functions can be rewritten as:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_{\theta \in \Theta} p(y|\theta)p(\theta)d\theta}$$

- Since $p(\theta|y)$ is a function of $\theta$, the denominator is a constant and thus, the posterior density of $\theta$ is often expressed up to a proportionality constant as:

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

  where $\propto$ stands for "proportional to".

# Bayes' theorem for density functions

- The proportionality constant in the posterior density $p(\theta|y)$ can be determined using the fact that the posterior density $p(\theta|y)$ is a p.d.f.
- Suppose that we have observed $Y = y$, and we want to evaluate and compare the odds of two possible numerical values for $\theta$, let them be $\theta_a$ and $\theta_b$. Given the data observed, the posterior probability (density) of $\theta_a$ relative to $\theta_b$ is:

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

  - $\frac{p(\theta_a)}{p(\theta_b)}$ is the prior odds
  - $\frac{p(y|\theta_a)}{p(y|\theta_b)}$ is the likelihood ratio
- In comparing the posterior probability of $\theta_a$ and $\theta_b$, it is **NOT** necessary to compute $p(y)$!

# Independent random variables

- We have seen earlier the definition of independent events $A$ and $B$: $A$ and $B$ are independent if $P(A|B) = P(A)$, i.e., if knowing that $B$ has occurred does not provide any information on $A$.
- If $A$ and $B$ are independent, then: $P(A \cap B) = P(A)P(B)$.
- If $Y_1, \ldots, Y_n$ are $n$ random variables with sample spaces $\mathcal{Y}_1, ..., \mathcal{Y}_n$ and $\theta$ is a parameter of the joint density function $p(y_1, \ldots, y_n|\theta)$, we say that $Y_1, \ldots, Y_n$ are conditionally independent given $\theta$ if for any collection of events $A_1 \subset \mathcal{Y}_1, \ldots, A_n \subset \mathcal{Y}_n$:

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n|\theta) = P(Y_1 \in A_1|\theta) \cdot ... \cdot P(Y_n \in A_n|\theta)$$

- In terms of joint density functions, $Y_1, \ldots, Y_n$ are conditionally independent given $\theta$ if:

$$p(y_1, \ldots, y_n|\theta) = p_{Y_1}(y_1|\theta) \cdot ... \cdot p_{Y_n}(y_n|\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i|\theta)$$

# Independent random variables

- In terms of conditional density functions, $Y_1, ..., Y_n$ are conditionally independent given $\theta$ if for any $i \neq j \in \{1, ..., n\}$ and subsets $A_i \subset \mathcal{Y}_i, A_j \subset \mathcal{Y}_j$:

$$P(Y_i \in A_i | \theta, Y_j \in A_j) = P(Y_i \in A_i | \theta)$$

- If $Y_1, ..., Y_n$ are random variables that are generated in the same way (e.g. by repeating the same experiment several times, or by sampling a population with replacement), then they can be modeled as having a common density function $p(y_i|\theta), i = 1, ..., n$. In this case, $Y_1, ..., Y_n$ are called conditionally independent and identically distributed and

$$p(y_1, ..., y_n|\theta) = p(y_1|\theta) \cdot ... \cdot p(y_n|\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

- We express this by writing: $Y_1, ..., Y_n|\theta \overset{i.i.d.}{\sim} p(y|\theta)$

# Exchangeable random variables

- Suppose $Y_1, ..., Y_n$ are $n$ random variables. $Y_1, ..., Y_n$ are called exchangeable if their joint marginal density $p(y_1, ..., y_n)$ is invariant to permutations of the indices.
  $Y_1, ..., Y_n$ are exchangeable if for all permutations $\pi$ of the indices set $\{1, 2, ..., n\}$:
  $$p(y_{\pi(1)}, ..., y_{\pi(n)}) = p(y_1, ..., y_n)$$
- Example: Consider a survey conducted in 1998 that asked respondents whether they were happy or not. Let $Y_i$ be the answer from the $i$-th sampled individual. Then:
  $$Y_i = \begin{cases} 1 & \text{if person } i \text{ is happy} \\ 0 & \text{otherwise} \end{cases}$$
  Now consider the answer for 10 respondents and suppose that of these 10 people, 6 responded that they were happy and 4 that they were not. Is there any reason to consider
  $$p(1, 1, 1, 1, 1, 1, 0, 0, 0, 0) \neq p(1, 0, 0, 1, 0, 1, 1, 0, 1, 1)?$$

# Exchangeable random variables

- How can we compute the marginal density $p(y_1, ..., y_n)$?
- Let's consider the random variable $Y_i$: we can model $Y_i$ as a Bernoulli random variable with probability of success (i.e., of responding 'Yes') equal to $\theta$. Then,

$$p(y_i|\theta) = \text{Bernoulli}(\theta) = \theta^{y_i}(1-\theta)^{1-y_i}.$$

- The parameter $\theta$ in the sampling model $p(y_i|\theta)$ is a random variable itself and we assume it has density $p(\theta)$.
- Now let's consider the variables: $Y_1, ..., Y_n$. Can we model them as conditionally independent given $\theta$? If the population is much larger than the sample size, sampling without replacement can be approximately considered as sampling with replacement and thus, $Y_1, ..., Y_n$ can be considered to be i.i.d. given $\theta$. Thus,

$$\begin{aligned} p(y_1, ..., y_n|\theta) &= \prod_{i=1}^n p(y_i|\theta) = \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \end{aligned}$$

# Exchangeable random variables

- We want to derive the marginal density $p(y_1, ..., y_n)$. We know:
  - the (marginal) density $p(\theta)$ for the random variable $\theta$
  - the conditional density $p(y_1, ..., y_n|\theta)$ for the random variables $Y_1, ..., Y_n$ **given** $\theta$: $p(y_1, ..., y_n|\theta) = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}$
- Then:

$$p(y_1, ..., y_n) = \int_{\theta\in\Theta} p(y_1, ..., y_n, \theta)d\theta = \int_{\theta\in\Theta} p(y_1, ..., y_n|\theta)p(\theta)d\theta$$
$$= \int_{\theta\in\Theta} \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}p(\theta)d\theta$$

- Let's now apply the formula above to our example with 10 respondents, 6 of which are happy and 4 are not:

$$p(1,1,1,1,1,1,0,0,0,0) = \int_{\theta\in\Theta} \theta^6(1-\theta)^{10-6}p(\theta)d\theta$$
$$p(1,0,0,1,0,1,1,0,1,1) = \int_{\theta\in\Theta} \theta^6(1-\theta)^{10-6}p(\theta)d\theta$$

- Under this model, $Y_1, ..., Y_n$ are exchangeable.

# Exchangeable random variables

- The result we have seen for the previous example holds in general.
- Theorem: If $\theta$ is a random variable with (marginal) density $p(\theta)$ and $Y_1, ..., Y_n$ are $n$ random variables that are conditionally independent and identically distributed given $\theta$, then **marginally** $Y_1, ..., Y_n$ are exchangeable.
- Proof: Let $\pi$ be a permutation of the indices $\{1, 2, ..., n\}$. We need to show that $p(y_1, ..., y_n) = p(y_{\pi(1)}, ..., y_{\pi(n)})$.
  Since $Y_1, ..., Y_n$ are conditionally independent and identically distributed given $\theta$, $p(y_1, ..., y_n|\theta) = \prod_{i=1}^n p(y_i|\theta)$.
  Hence,

$$\begin{aligned} p(y_1, ..., y_n) &= \int_{\theta\in\Theta} p(y_1, \ldots, y_n|\theta)p(\theta)d\theta \\ &= \int_{\theta\in\Theta} \{\prod_{i=1}^n p(y_i|\theta)\}p(\theta)d\theta \\ &= \int_{\theta\in\Theta} \{\prod_{i=1}^n p(y_{\pi(i)}|\theta)\}p(\theta)d\theta \\ &= p(y_{\pi(1)}, ..., y_{\pi(n)}). \end{aligned}$$

# De Finetti's theorem

- We have shown that:

$$\left.\begin{array}{l} \theta \sim p(\theta) \\ Y_1, ..., Y_n | \theta \overset{i.i.d.}{\sim} p(y|\theta) \end{array}\right\} \Longrightarrow Y_1, ..., Y_n \text{ are exchangeable}$$

- Is the opposite true? De Finetti states that the opposite holds as well.
- Theorem: Let $Y_1, Y_2, ...$ be random variables all with sample space $\mathcal{Y}$. If, for any $n$, $Y_1, \ldots, Y_n$ are exchangeable random variables with marginal density $p(y_1, ..., y_n)$, then it is possible to express $p(y_1, ..., y_n)$ as:

$$p(y_1, ..., y_n) = \int_{\theta \in \Theta} p(y_1, ..., y_n|\theta) p(\theta) d\theta = \int_{\theta \in \Theta} \prod_{i=1}^{n} p(y_i|\theta) p(\theta) d\theta$$

for some parameter $\theta$ provided with a marginal density $p(\theta)$ (prior distribution of $\theta$) and some sampling model $p(y|\theta)$. The form of the prior distribution $p(\theta)$ and of the sampling model $p(y|\theta)$ depends on the form of $p(y_1, ..., y_n)$.

# De Finetti's theorem

- De Finetti's theorem states that if we have random variables that are (marginally) exchangeable, we can always set up a Bayesian model for our data.
- De Finetti's theorem + the previous theorem yield:

$$\left.\begin{array}{l}\theta \sim p(\theta) \\ Y_1, ..., Y_n | \theta \overset{i.i.d.}{\sim} p(y|\theta)\end{array}\right\} \Longleftrightarrow Y_1, ..., Y_n \text{ are exchangeable for all } n$$

- When can we assume that random variables $Y_1, ..., Y_n$ are exchangeable for any $n$?
  - The label/order of the variables does not convey any information and we have repeatability
- Some examples include:
  - $Y_1, ..., Y_n$ are the outcomes of a repeatable experiment
  - $Y_1, ..., Y_n$ are sampled from a **finite** population **with** replacement
  - $Y_1, ..., Y_n$ are sampled from a **infinite** population **without** replacement

# Prior distributions

- We have seen that the following holds:

$$\left.\begin{array}{l}\theta \sim p(\theta) \\ Y_1, ..., Y_n|\theta \overset{i.i.d.}{\sim} p(y|\theta)\end{array}\right\} \Longleftrightarrow Y_1, ..., Y_n \text{ are exchangeable for all } n$$

- The questions now are:
  - What is $p(\theta)$?
  - How do we choose $p(\theta)$?

- If $Y_1, ..., Y_n$ are binary random variables, then $p(\theta)$ represents our beliefs, before observing the data, of $\lim_{n\to\infty} \sum_{i=1}^{n} \frac{Y_i}{n}$.

# Prior distributions

- How to choose a prior distribution $p(\theta)$? There are various type of prior distributions:
  - Non-informative priors: a prior distributions $p(\theta)$ is non-informative if it has minimal impact on the posterior distribution $p(\theta|y)$ of $\theta$.
    In general, a non-informative prior is one which is dominated by the likelihood, that is, it is a prior that does not change very much over the region in which the likelihood is appreciable and does not assume large values outside that range. A prior that has these properties is said to be locally uniform.
    Non-informative priors are also called reference priors, vague priors or flat priors.
  - Example: if $\theta \in \Theta = [0, 1]$, then $p(\theta) = \text{Uniform}(0, 1)$ is a non-informative prior and $p(\theta) = 1$ for $\theta \in \Theta = [0, 1]$.
  - Example: if $\theta \in \Theta = \mathbb{R}$, then if $p(\theta) = \text{Normal}(\mu_0, \sigma_0^2)$ with $\sigma_0^2 \to \infty$, we get a non-informative prior. That is, we can pick $\sigma_0^2$ large enough so that we obtain a non-informative prior.

# Prior distributions

- How to choose a prior distribution $p(\theta)$? There are various type of prior distributions:
  - Improper priors: a prior distribution $p(\theta)$ on $\theta \in \Theta$ is said to be improper if
$$\int_{\theta \in \Theta} p(\theta) d\theta = \infty$$
  - Improper priors are often used in Bayesian inference because they include non-informative priors.
  - Example: if $\theta \in \Theta = \mathbb{R}$ and $p(\theta) \propto 1$, then $p(\theta)$ is an improper prior since clearly
$$\int_{\theta \in \Theta} p(\theta) d\theta = \int_{-\infty}^{\infty} d\theta = \infty$$
  - An improper prior $p(\theta)$ may result in an improper posterior distribution $p(\theta|y)$. One **cannot** make inference with improper posterior distributions!
  - An improper prior $p(\theta)$ may still lead to a proper posterior distribution.

# Prior distributions

- If $p(\theta) = c$ for some $c > 0$, then

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_{\theta\in\Theta} p(y|\theta)p(\theta)d\theta} \\ &= \frac{p(y|\theta)c}{\int_{\theta\in\Theta} p(y|\theta)cd\theta} = \frac{p(y|\theta)}{\int_{\theta\in\Theta} p(y|\theta)d\theta} \\ &= \frac{1}{K}p(y|\theta) \end{aligned}$$

  where $K = \int_{\theta\in\Theta} p(y|\theta)d\theta$.

  The posterior distribution $p(\theta|y)$ is just the renormalized likelihood.

# Prior distributions

- The choice of the prior distribution $p(\theta)$ is a delicate issue and the main criticism to Bayesian inference.
- Two people analyzing the same data but using different priors could obtain different answers.
- When carrying out a Bayesian analysis it is important to examine the prior sensitivity, that is, determine how sensitive are the results to the choice of the prior distribution.

# Binomial model

- Consider data from a randomized controlled clinical trial evaluating a drug to reperfuse blocked blood vessels in patients with ischemic stroke.
- We indicate with $Y_i$ whether the $i$-th patient treated with the drug had a favorable outcome (no disability at 90 days). Then:

$$Y_i = \begin{cases} 1 & \text{if patient } i \text{ has a favorable outcome} \\ 0 & \text{otherwise} \end{cases}$$

- Then $Y_i$ is a binary random variable.
- Suppose that data from $n = 129$ patients were available. Then, since the size $N$ of the general population is much greater than the size $n$ of the sample, $Y_1, ..., Y_n$ can be considered to be exchangeable random variables.

# Binomial model

- Based on De Finetti's theorem, we can set up the following model for our data:
  - a prior probability $p(\theta)$ for a parameter $\theta$, representing the proportion of 1's in the population ($\sum_{i=1}^{N} \frac{Y_i}{N}$ )
  - given $\theta$, the $Y_1, ..., Y_n$ are i.i.d. with $p(y|\theta) = \theta^y(1-\theta)^{1-y}$
- The sampling model adopted leads to:

$$p(y_1, ..., y_n|\theta) = p(y_1, ..., y_{129}|\theta) = \prod_{i=1}^{129} \theta^{y_i}(1-\theta)^{1-y_i}$$
$$= \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}$$

- Now, we need to choose a prior distribution $p(\theta)$ for $\theta$.

# Binomial model

- The parameter $\theta$ is a random variable with parameter space $\Theta = [0, 1]$.
- Let's use a non-informative prior, e.g., $p(\theta) = \text{Uniform}([0, 1])$, i.e., $p(\theta) = 1$ for $\theta \in [0, 1]$.
- Then, we obtain the following posterior distribution:

$$p(\theta|y_1, ..., y_{129}) = \frac{p(y_1, ..., y_{129}|\theta)p(\theta)}{p(y_1, ..., y_{129})} = \frac{1}{p(y_1, ..., y_{129})} p(y_1, ..., y_{129}|\theta)$$
$$\propto p(y_1, ..., y_{129}|\theta)$$

  since $p(y_1, ..., y_{129})$ does not depend on $\theta$.

- Therefore, the posterior distribution $p(\theta|y_1, ..., y_{129})$ of $\theta$ has the same shape as $p(y_1, ..., y_{129}|\theta)$ but they are on different scales.

# Binomial model

- Suppose that our data is as follows:
  - $n = 129$ patients received the drug
  - $118$ had a favorable outcome
  - $11$ did not

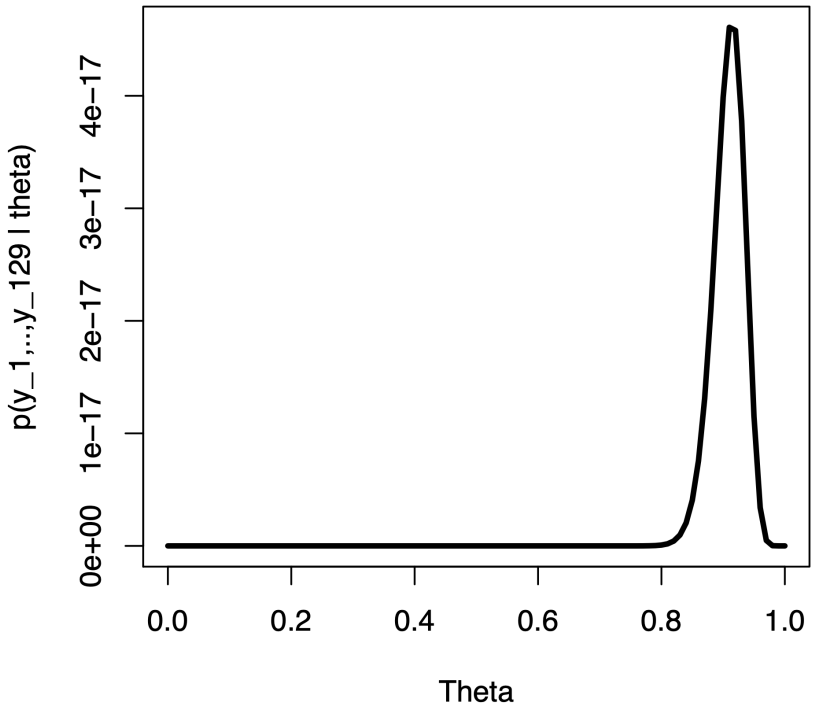
- Then:
  - $\sum_{i=1}^{129} y_i = 118$
  - $129 - \sum_{i=1}^{129} y_i = 11$

  which implies

$$p(y_1, ..., y_{129}|\theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i} = \theta^{118}(1-\theta)^{11}$$

# Binomial model

- The following plot shows $p(y_1, ..., y_{129}|\theta) = \theta^{118}(1-\theta)^{11}$ as a function of $\theta$.




- The posterior distribution $p(\theta|y_1, ..., y_{129})$ has the same shape but simply a different scale, so $p(\theta|y_1, ..., y_{129})$ puts mass on values of $\theta$ that are greater than $0.8$.
- Note the scale on the vertical axis: the values are of the order of $10^{-17}$.

# Binomial model

- To determine the scale for the posterior distribution $p(\theta|y_1, ..., y_n)$, we use the fact that this is a conditional density and since it is a probability density function, it needs to integrate to 1, that is:

$$1 = \int_{\theta \in \Theta} p(\theta|y_1, ..., y_{129}) d\theta = \int_0^1 \frac{\theta^{118}(1-\theta)^{11}}{p(y_1, ..., y_{129})} d\theta$$

- The following result from calculus is useful:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1} d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where $a, b > 0$, and $\Gamma(x)$ with $x > 0$ is the Gamma function which can be looked up in `R` using the command `gamma()`.

- Thus, $p(\theta|y_1, ..., y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)} \theta^{118}(1-\theta)^{11}$.

# Binomial model

- Definition: A random variable $Y$ with sample space $\mathcal{Y} = [0, 1]$ is said to have a Beta distribution$(a, b)$ if its p.d.f. is given by

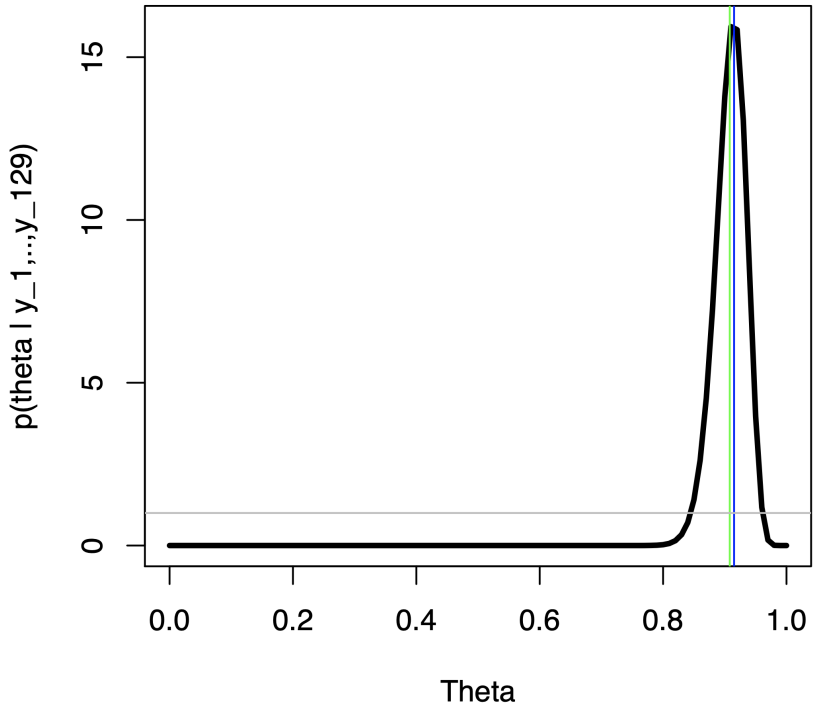
$$p(y|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} \quad \text{with} \quad a, b > 0.$$

  For such a random variable
  - $E[Y] = \frac{a}{a+b}$
  - $\text{Mode}[Y] = \frac{a-1}{(a-1)+(b-1)}$ if $a > 1$ and $b > 1$
  - $\text{Var}[Y] = \frac{ab}{(a+b+1)(a+b)^2} = \frac{E[Y](1-E[Y])}{(a+b+1)}$

- Thus, the posterior distribution $p(\theta|y_1, ..., y_{129}) = \frac{\Gamma(131)}{\Gamma(119)\Gamma(12)} \theta^{118}(1-\theta)^{11}$ is a Beta distribution with $a = 119$ and $b = 12$.

# Binomial model

- The posterior distribution $p(\theta|y_1, ..., y_{129}) = \text{Beta}(119, 12)$ has $\text{Mean} = 0.908$, $\text{Mode} = 0.915$, and $\text{Variance} = 6.25 \times 10^{-4}$.
- Prior (in gray) and posterior (in black) distributions of $\theta$:




- The shape is the same as $p(y_1, ..., y_{129}|\theta)$, but the scale is much different.
- Note that in Bayesian inference:

  $$\text{Posterior information} \geq \text{Prior information} \geq 0$$

  with the 2nd "$\geq$" replaced by "$=$" only if the prior is non-informative.