Module 2 - Basics of Probability

MATH2269 Applied Bayesian Statistics

Dr. Haydar Demirhan

Introduction

In this module, we will focus on basics of probability. A basic knowledge of probability is required to understand what is happening behind the MCMC samplers running the Bayesian analysis in practice. Actually, Bayesian thinking is yet another (modern) definition of probability. So, we will try to find an answer to the question:

What is the probability?

Our main goal in this module is to understand basic concepts of probability. This will enable us to go further easily. In detail, we will study,

- Definitions of probability,

- Probability distributions,

- Expected value and variance,

- Two-way distributions,

- Conditional probability, and

- Independence.

Sample Space

The sample space is the key concept of probability. We should define our sample space before going on with building other context.

The set of all possible events

The sample space is simply defined as the set that contains all possible events. So what is an event?

An event is a subset of outcomes of an experiment. For example, If I flip a coin (experiment), what would I expect to observe (event)?

- A head, tail, or torso? Or combinations of these?

It is not possible to observe any combination of events in a single flip of ONE coin at a time. So, the outcomes are mutually exclusive. Also, torso is not observable under standard circumstances.Thus, there are only two events: H and T that compose the {}.

Let $A_1, A_2, \ldots, A_n$ be the possible outcomes of an experiment, then the resulting sample space, Ω, is written as $\Omega = \{A_1, A_2, \ldots, A_n\}$

Sampling

Sampling is the job of selecting units from a population of interest according to a predetermined rule. Sampling is the most important part of a scientific research. It is directly related with the scientific quality of the research project. Mainly, there are two types of sampling that {} and {}.

- *Probabilistic* sampling techniques include simple random sampling, stratified random sampling, cluster sampling, etc.

- *Non-probabilistic* sampling techniques include quota sampling, snowball sampling, expert sampling, etc.

There is no selection bias in *probabilistic* sampling; and hence, it is possible to make statistical inferences such as modelling over the random samples observed by *probabilistic* sampling techniques. However, it is very very dangerous to make inferences with the samples those did not collected under probabilistic sampling plans. You cannot know the amount of bias that is caused by the sampling bias of the *non-probabilistic* sampling scheme. It is only possible to give descriptive statistics over the samples gathered by the *non-probabilistic* sampling schemes!

Probability

The degree of belief

The degree of belief represents our personal judgement of possibilities of the event in a sample space. As an example let us have a look at the question "Is the coin fair?" If the coin was manufactured by a government mint, we would have a high degree of belief in fairness of the coin. However, if the coin was manufactured by Acme Magic and Novelty Company, we would have a high degree of belief in unfairness of the coin.

The degree of belief in a parameter θ can be denoted $p(\theta)$. And it represents our personal judgement of possibilities of the event in a sample space.

If θ is defined as "the probability of having a H in a flip of the coin," $\theta = 0.5$ means that the coin is fair. If the coin was manufactured by a government mint, we can write $p(\theta = 0.5) = 0.9$ to represent our high degree of belief. If the coin was manufactured by Acme Magic and Novelty Company, we can assign $p(\theta = 0.5) = 0.1$ or $p(\theta = 0.9) = 0.9$ to show our disbelief in fairness of the coin. As you see, it is not possible to represent the result ``torso" or our belief in by this way. Why?
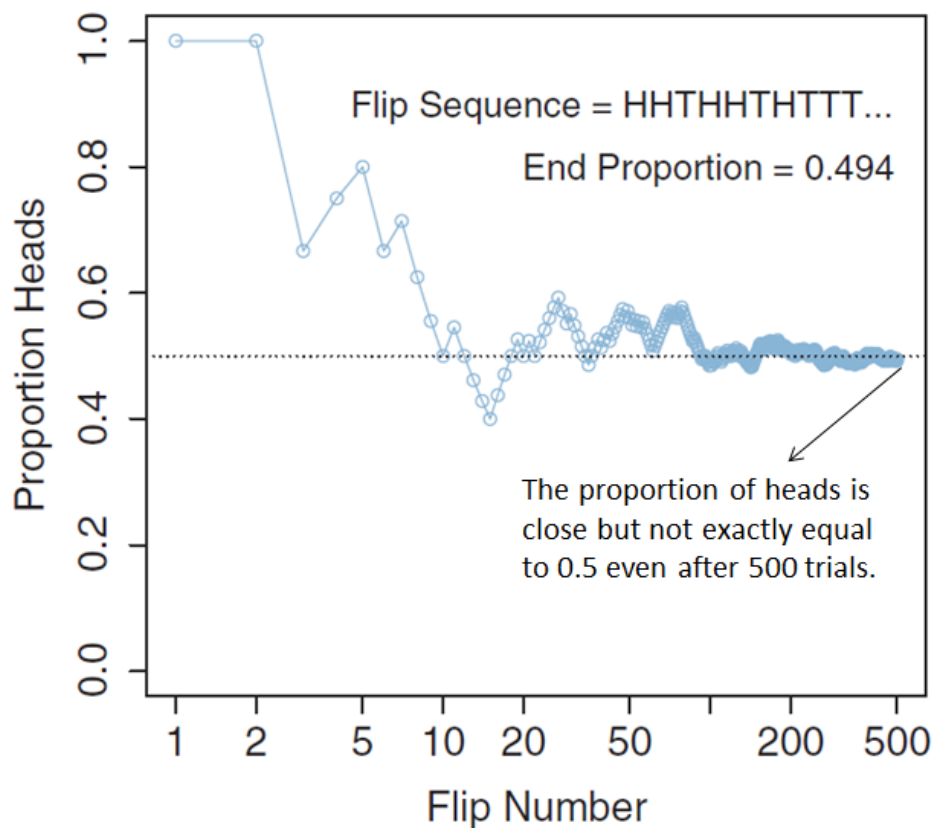
OUTSIDE the head

In the nature, probabilities are defined over sample spaces generated by well-defined experiments. The concept of probability is mathematically defined in four ways: classical definition, frequency definition, axiomatic definition, and Bayesian definition.

Relative frequency definition of probability

Relative frequency definition of probability is closely related with long-runs of experiments. The define the probability, we run the experiment for many times and observe the frequency of each outcome within total number of replications.

When we repeat an experiment so many times, eventually, relative frequencies of events that compose the sample space approaches o the corresponding probabilities. For example, if we flip the coin infinitely many times, the number of heads approaches to half of the total number of flips if the coin is fair.

Running Proportion of Heads

Flip Sequence = HHTHHTHTTT...

End Proportion = 0.494

The proportion of heads is close but not exactly equal to 0.5 even after 500 trials.

Adopted from: J. Kruschke. 2014. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Ed., Elsevier.

Classical definition of probability

Classical definition of probability is directly related to the proportion of outcomes of interest to the total number of outcomes over the sample space.

In our example, the sample space is $\Omega = \{H, T\}$. For a fair coin, both outcomes are equally likely. Therefore, the probability of getting a H is the same as that of getting a T and equals to 1/2.

Consider a fair six-side dice. So, $\Omega = \{1, 2, \ldots, 6\}$. What is the probability of getting an even number in a toss of this dice? Because the number of sides with even numbers in the sample space is 3 and the total number of possible outcomes is 6, the desired probability is $3/6 = 0.5$.

Axiomatic definition of probability

The numbers that satisfy the following properties are called probabilities (Kruschke, 2014; Miller and Miller, 2004):

1. Probability of an event A is a non-negative real number: $P(A) \geq 0$ for all $A \in \Omega$ (i.e., zero or positive),

2. $P(\Omega) = 1$; the sum of the probabilities across all events in the sample space must be equal to 1,

3. If $A_1, A_2, \ldots$ is an finite or infinite sequence of mutually exclusive event of Ω, then
   $P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$ For any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities.

These axioms were introduced by the famous Russian mathematician Andrey Kolmogorov (1903-1987) in 1933. You can find the story here (http://www.probabilityandfinance.com/articles/04.pdf).

INSIDE the head

Our judgments about the chance of occurrence of events, in fact, represents the perception of probability inside our brains. Our beliefs refer to a mutually exclusive and exhaustive set of possibilities those based on a (sample) space composed by our perception of the experiment. 'The mathematical properties of probabilities outside the head and beliefs inside the head are the same in their essentials.'

Our subjective beliefs determines how strongly do you agree on a statement. To transform our subjective beliefs into mathematics, we should specify how likely we think each possible outcome is. To include all available information as objective as possible, we should calibrate our subjective beliefs.

Relative to the other events with clear probabilities, you should choose a number between 0 and 1 that reflects your belief probability. You can use a graphic rating scale to represent the degree of your belief. See here (https://www.questionpro.com/blog/4-measurement-scales-every-researcher-should-remember/) for more information. In some cases, the number of outcomes are infinitely many. So, we need to describe our in a more general manner that considers all of these infinite number of possible outcomes. The way of doing this is to use a probability distribution to represent our beliefs.

Probability Distributions

In order to study probability distributions, we should study the concept of random variable.

A random variable (r.v.) is a function that assigns a number to each element of the associated sample space. When our sample space contains qualitative characteristics as elements, we need to convert these elements to numbers to work on them mathematically. So, the context of random variable does it for us!

Let us roll a dice once. You can do it ! How does a random variable converts the picture on the upper side of a dice to a number? Let X be number shown in the picture on the upper side of the dice.

A probability distribution is the list of all possible outcomes of an experiment and probabilities that correspond to the outcomes.

Discrete probability distributions

If the number of mutually exclusive and exhaustive elements contained in a samples space is finite $(\Omega = \{A_1, A_2, \ldots, A_n\}, 1 < n < \infty)$ or countably infinite $(\Omega = \{A_1, A_2, \ldots\})$, we have a *discrete sample space* and the probability distribution associated with that sample space is discrete.

The mathematical function that represents a discrete probability distribution is called {} (p.m.f) or {} (p.f.). 'The term {} refers the amount of stuff in an object.'

Outcomes	Probabilities
•	1/6
• •	1/6
• • •	1/6
•• ••	1/6
••• •••	1/6
••• •••	1/6

Probability distribution

$\longrightarrow$

Probability function

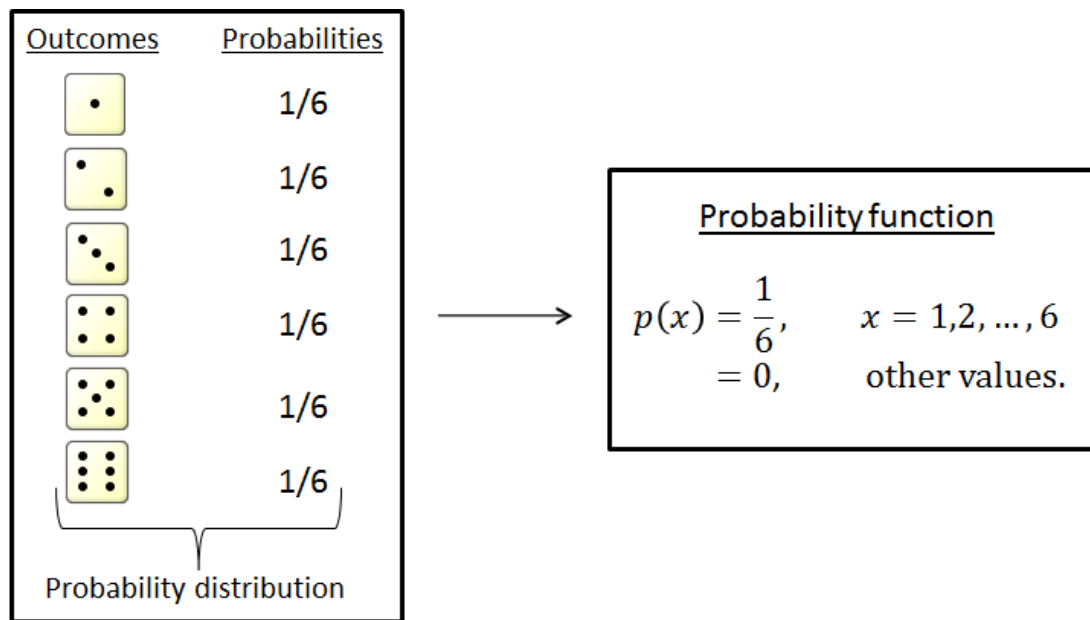$$p(x) = \frac{1}{6}, \qquad x = 1,2,\ldots,6$$
$$= 0, \qquad \text{other values.}$$

Continuous probability distributions

If the samples space contains an uncountably infinite number of elements, then we have a {} and the probability distribution associated with that sample space is continuous.

The mathematical function that represents a continuous probability distribution is called {} (p.d.f). Note that the role of a continuous random variable is the same with continuous probability distributions.

When working with continuous random variables, it is impossible to talk on specific values that the r.v. takes. Instead a continuous r.v. takes interval values. In this case, we can consider probability mass of intervals or probability densities.

It is impossible to measure height of a person. The accuracy of the measurement depends on that of the measurement tool. Height of a randomly selected person would be 188.83212342... cm. therefore, if an r.v. X is defined as the height of a person the domain that supports X is $x > 0$. You can refer here (https://ourworldindata.org/data/food-agriculture/human-height/) for some data on the human height ;)
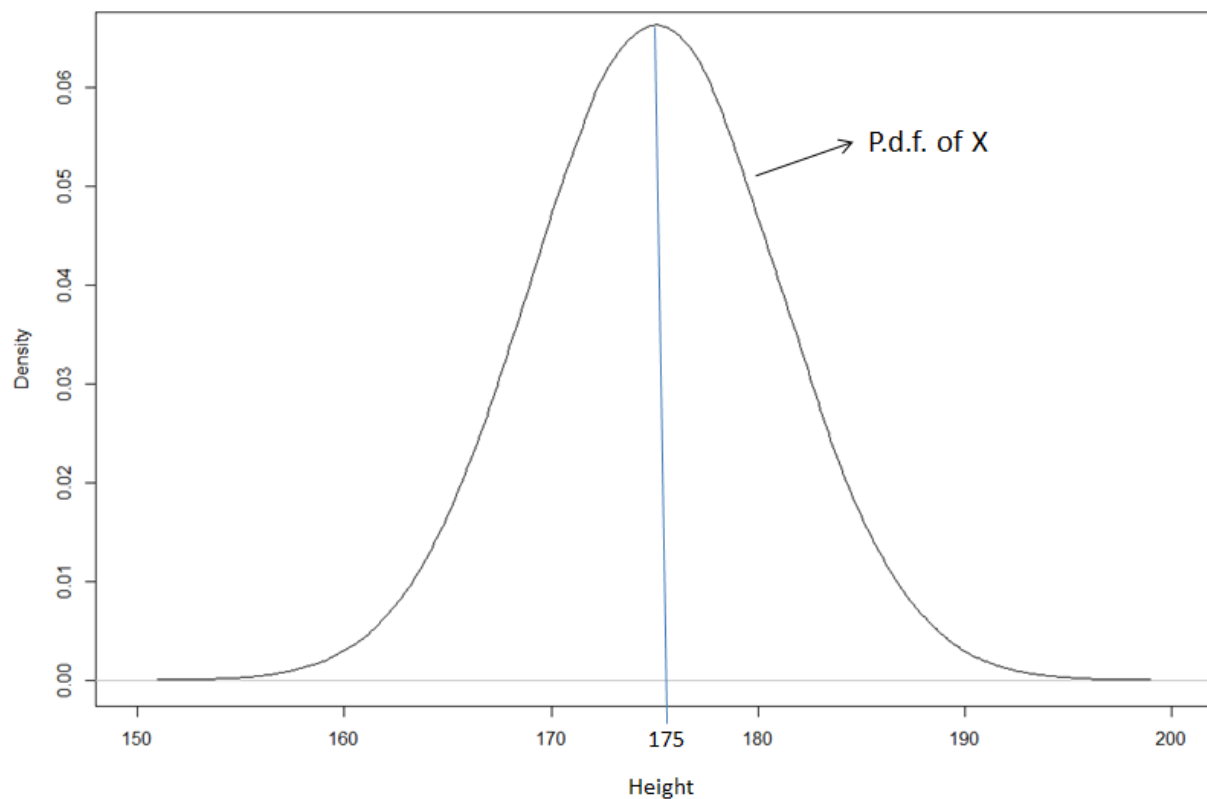
It is possible to *discretize* the domain of a continuous variable to study its distribution. However, in this case, we should note that the discretisation decreases the resolution!

The heights of a population can be discretized into intervals. This is equivalent to putting heights into a finite set of mutually exclusive and exhaustive ``bins." Then we can find the probability that a randomly selected person falls into any of those intervals.

Adopted from: J. Kruschke. 2014. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Ed., Elsevier.

Adopted from: J. Kruschke. 2014. Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, Second Ed., Elsevier.
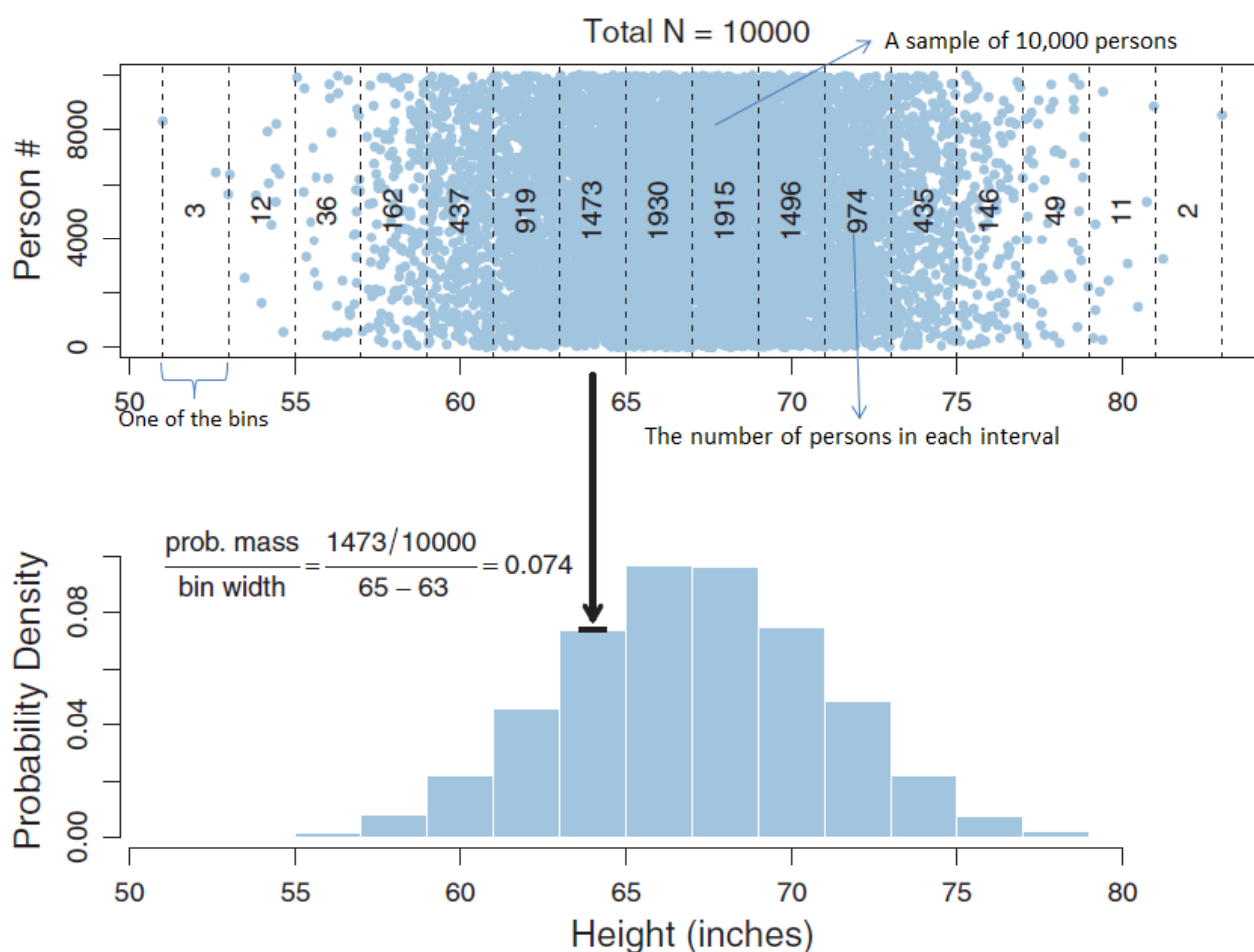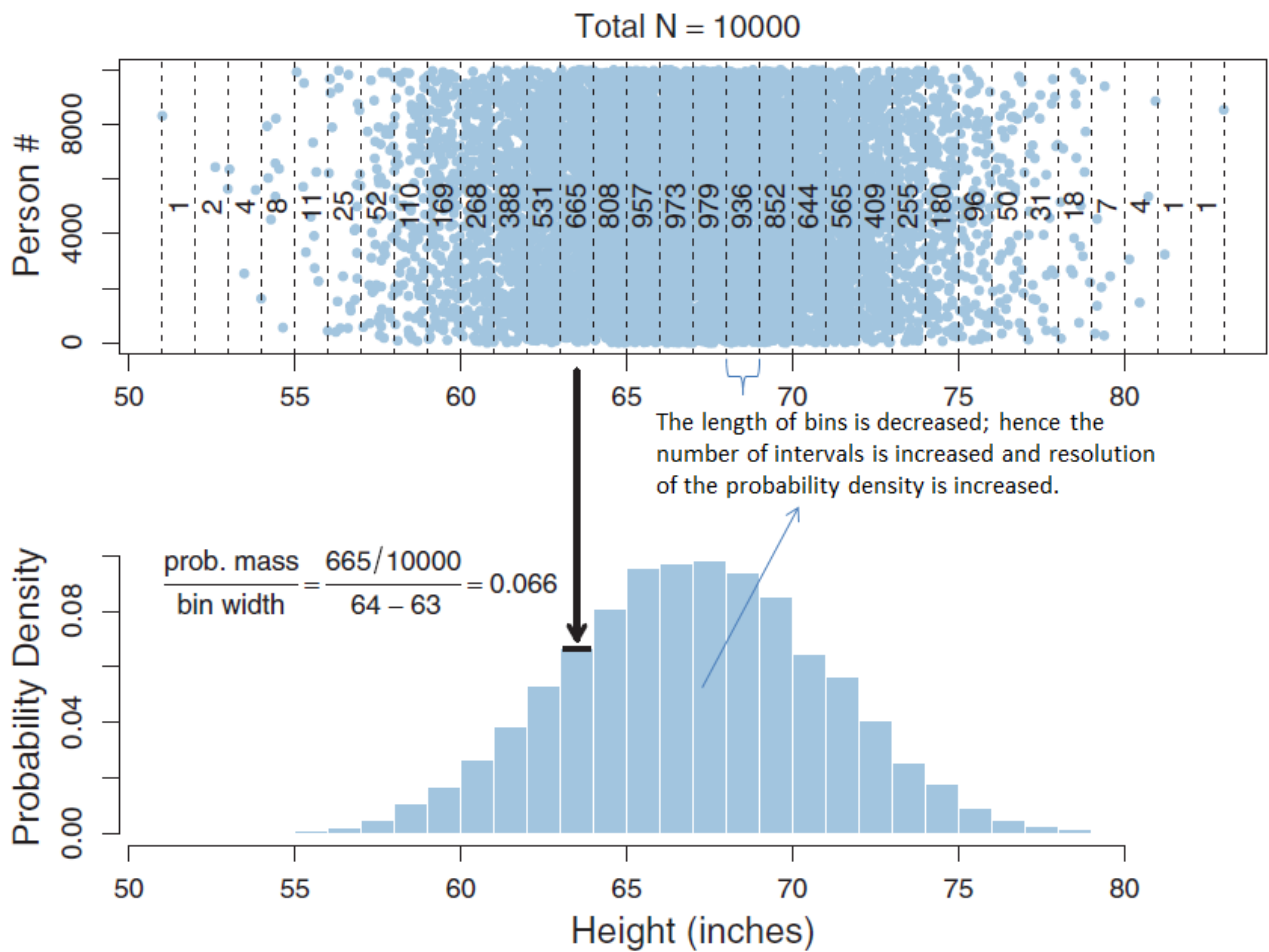
Properties of probability distributions

What makes a mathematical function a probability function?

A function that satisfy the following properties is called probability function for a discrete r.v. X (Miller and Miller, 2004):

- $p(x) \geq 0$ for all values of X in its domain R_X,
- $\sum_{x \in R_X} p(x) = 1$.

What makes a mathematical function a probability density function?

A function that satisfy the following properties is called probability density function for a continuous r.v. Y (Miller and Miller, 2004):

- $p(y) \geq 0$ for $-\infty < x < \infty$,
- $\int_{-\infty}^{\infty} p(y)\mathrm{d}y = 1$.

May be the most frequently used distribution in practice is normal (Gaussian) distribution. If $X \sim N(\mu, \sigma^2)$, then X has the following p.d.f.:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}, \quad -\infty < x < \infty.$$

The pdf of normal distribution for several parameters are shown in the following plot:
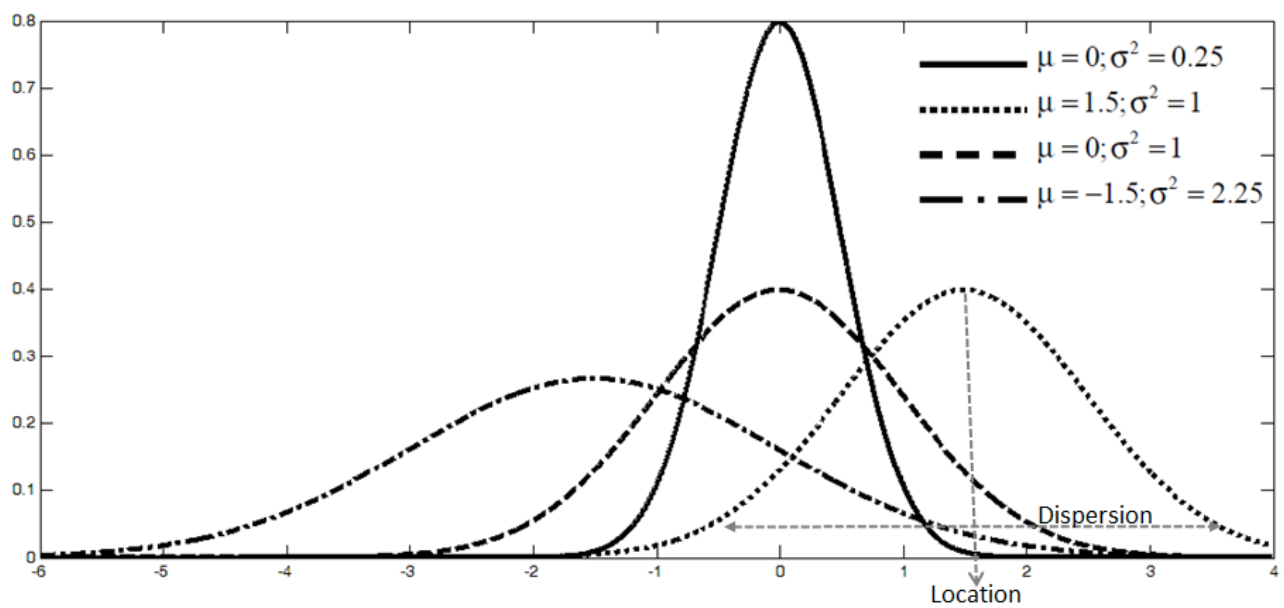
Use the following app to generate various normal distributions and try to observe the effect of parameters on the location and shape of the distribution.

Normal Distribution

```
##
## **********************************************************************
## Prepared by Dr. Haydar Demirhan
## e@mail: haydar.demirhan@rmit.edu.au
## **********************************************************************
```

```
## Loading required package: ggplot2
```

Mean for the first normal distribution

-10   0   10

-10 -8 -6 -4 -2   0   2   4   6   8   10
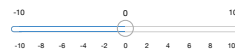
Variance for the first normal distribution

0 1   20

0   2   4   6   8   10 12 14 16 18 20

Mean for the second normal distribution

-10   2   10

-10 -8 -6 -4 -2   0   2   4   6   8   10

Variance for the second normal distribution

0   5   20

0   2   4   6   8   10 12 14 16 18 20

Pdf's of normal distribution
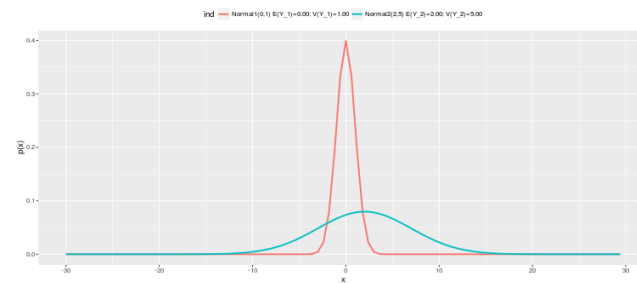
Mean and variance of a distribution

The mean of a distribution is also called *expected value $E(X)$* of a random variable. It shows what we expected to see as the value of the relevant random variable for most of the time.

If X is a discrete r.v., $E(X)$ is calculated as follows:

$$E(X) = \sum_{x \in R_X} x \cdot p(x)$$

If X is a continuous r.v., $E(X)$ is calculated as follows:

$$E(X) = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x.$$

The variance of a random variable $V(X)$ shows dispersion of the distribution away from its mean. For discrete and continuous random variables, $V(X)$ is calculated as follows:

$$V(X) = E(X^2) - E(X).$$

The square root of the variance is called *standard deviation* of an rv.

The relation between mean and variance of standard normal and normal distributions is shown below:
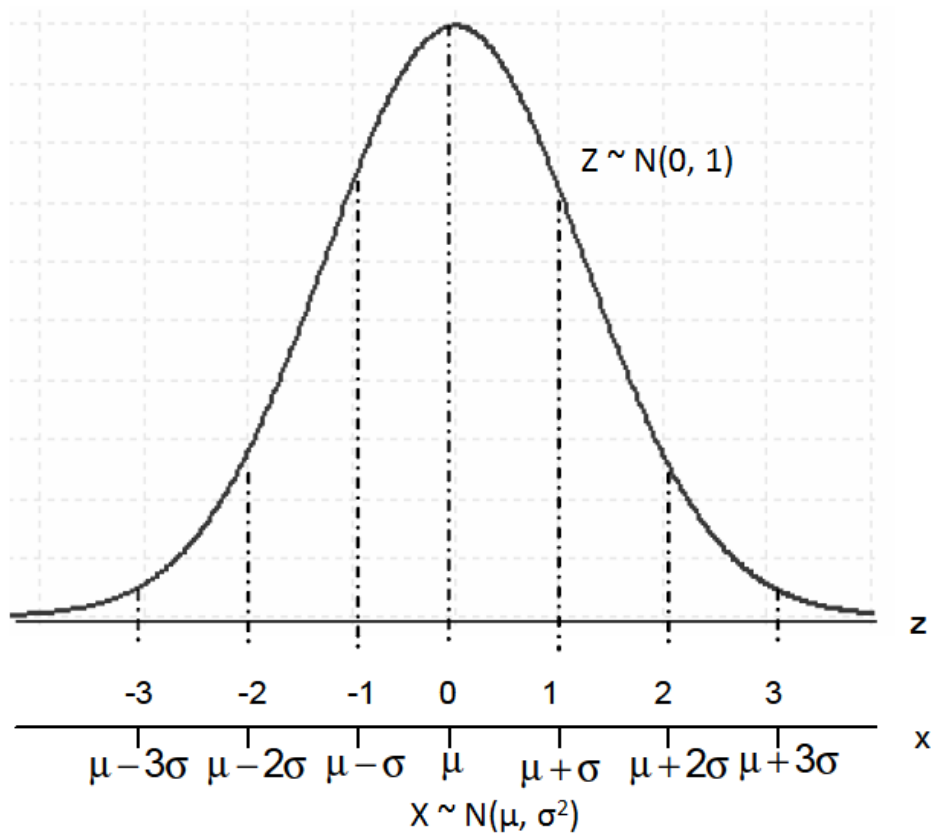
Example

Find the mean and variance of the rv X that has the following pdf.

$$
\begin{aligned}
p(x) \quad &= 6x(1-x), \quad 0 < x < 1, \\
&= 0, \qquad\qquad elsewhere.
\end{aligned}
$$

Pdf of the rv X is given below. The location of the distribution is shown by the down arrow.

p(x)

1.4
1.2
1.0
0.8
0.6
0.4
0.2

0.2   0.4   0.6   0.8   1.0   x

Location of the distribution is around here!

Variance and expected value of X are superimposed on the pdf below:

p(x)

1.4
1.2
1.0
0.8
0.6
0.4
0.2

V(X) = 0.05

0.2   0.4   0.6   0.8   1.0   x

E(X) = 0.5

What do mean and variance mean in terms of Bayesian statistics?

Role of mean in Bayesian thinking

When $p(\theta)$ represents credibility values of θ, instead of the probability of sampling θ, then the mean of $p(\theta)$ can be thought of as a value of ? that represents a typical credible value.

Role of variance in Bayesian thinking
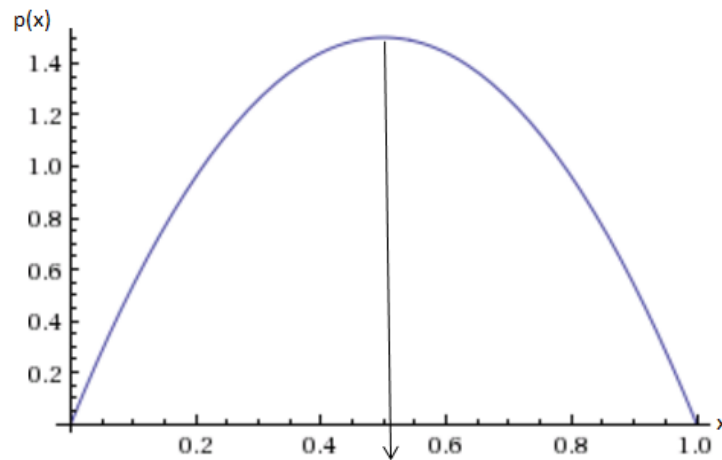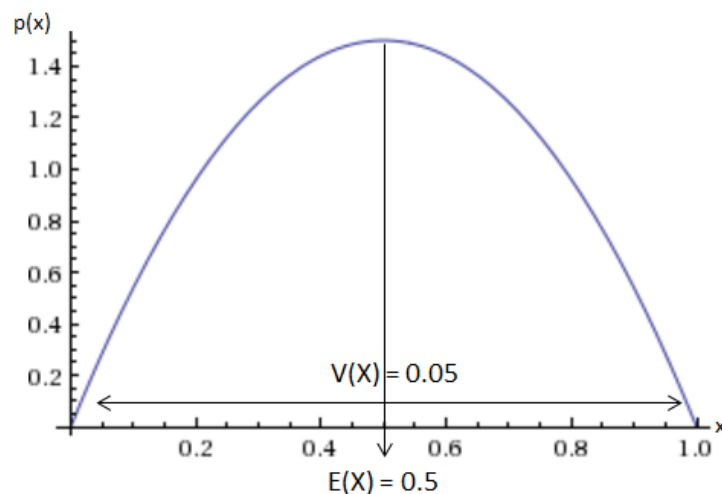
The standard deviation of θ, which measures how wide the distribution is, can be thought of as a measure of uncertainty across candidate values. The standard deviation of θ, which measures how wide the distribution is, can be thought of as a measure of uncertainty across candidate values.'

Highest density interval

The Highest density interval (HDI) indicates which points of a distribution are most credible, and the interval that covers most of the distribution. The main characteristic of HDI intervals is that every point inside the interval has higher credibility than any point outside the interval.

The values of x in the $(1 - \alpha)\%$ HDI are those such that $p(x) > W$ where W satisfies the following equation:

$$\int_{x:p(x)>W} p(x)\mathrm{d}x = (1 - \alpha).$$

An example for a symmetric HDI interval is given below:
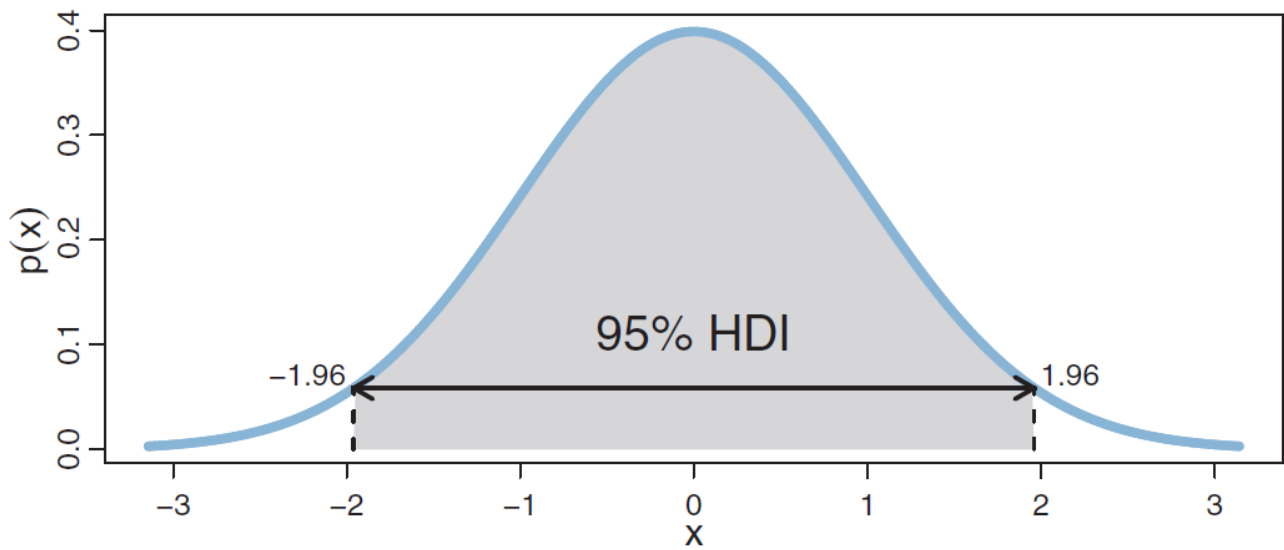
An asymmetric HDI interval is given below:
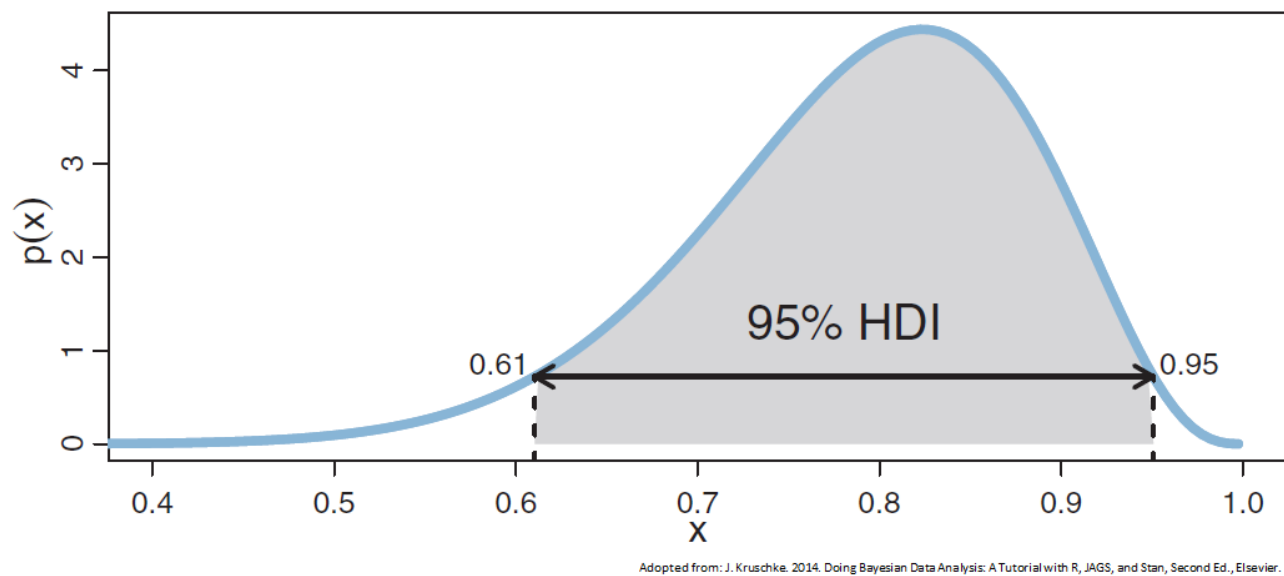
An HDI interval for a multi-modal distribution is shown below:
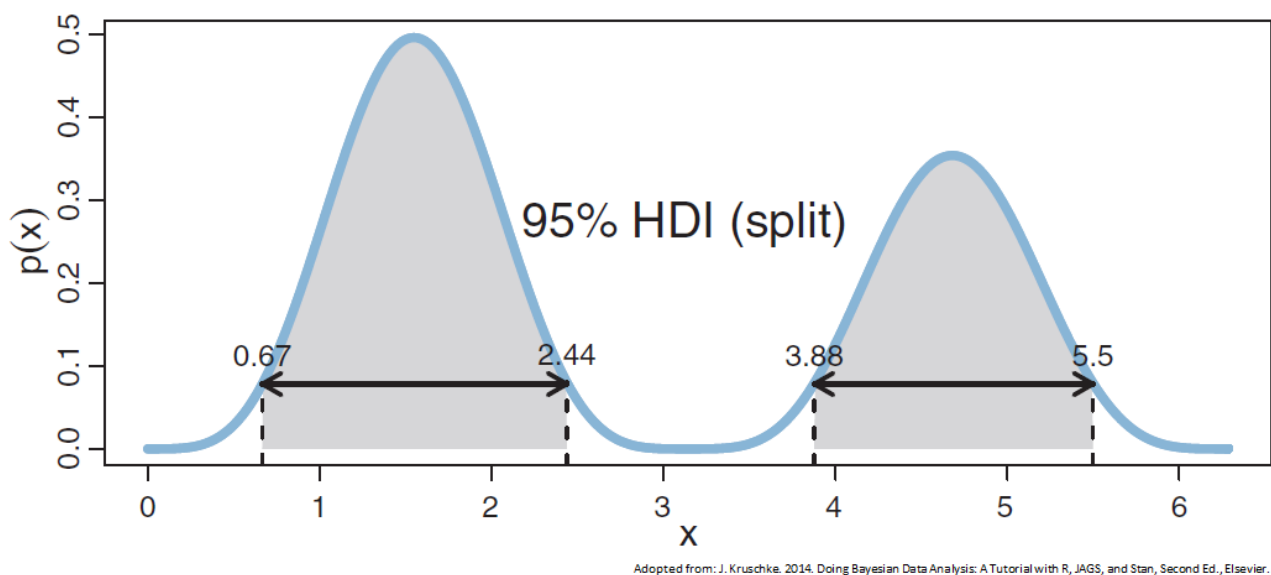
Two-way distributions

In some studies, we work with two variables that are interrelated with each other. For example, the temperature is just related with humidity. So, one can consider temperature and humidity of a site, simultaneously. In this case, we have a two-way distribution.

The types of considered variables are important here. The variables constituting the joint distribution can be both discrete or continuous. Although it is possible to have a mixed two-way distributions with one discrete and one continuous distribution, this case is outside of our current focus.

Example

Consider eye color and hair color of a population composed of 100 students. When we consider eye color and hair color of students simultaneously, we get the following two-way distribution.

	Hair color				
Eye color	Black	Brunette	Red	Blond	Total
Brown	11	20	4	1	36
Blue	3	14	3	16	36
Hazel	3	9	3	2	17
Green	1	5	2	3	11
Total	18	48	12	22	100

Two-way distributions represent joint distribution of two variables. It is possible to built up joint distributions with more than two variables. For two discrete variables, a two-way distribution contains, joint probabilities associated with the combinations of levels of two variables. If X is row variable and Y is column variable, then $p(x, y)$ shows that probability than an individual is in the level x of row variable and in the level y of column variable, and it is call *joint probability function*.

Example

Joint probability distribution of Eye Color and Hair Color is seen in the table below. Because we have two variables, it is practical to show joint pf, which is composed of (x, y) pairs and corresponding probabilities $p(x, y) = P(X = x, Y = y)$ in a table structure.

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal
Brown	0.11	0.2	0.04	0.01	0.36
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.03	0.02	0.17
Green	0.01	0.05	0.02	0.03	0.11
Marginal	0.18	0.48	0.12	0.22	1

For example, the top-left cell indicates that the joint probability of brown eyes and black hair is 0.11, the joint probability of blue eyes and black hair is only 0.03.

Please be aware!

Although our text book writes probabilities in percentages such as '… the joint probability of brown eyes and black hair is 0.11 (i.e., 11%),' it is not a convenient way of inferring probabilities.

When we say, a probability is equal to 11%, we only consider (approximate) relative frequency definition of the probability! However, in our example, we have joint population distribution of eye color and hair color; hence, the probabilities at hand are exact!

Marginal distributions

When we have given a joint distribution, we one of the potential interest is the distributions of components themselves. The probability distribution of each component is called *marginal distribution*. Each marginal distribution is probability distribution itself.

Marginal probability function of each margin is obtained by summing out the rest of variables:

$$p(x) = \sum_{R_Y} p(x, y).$$

$$p(y) = \sum_{R_X} p(x, y).$$

Marginal probability *distribution* function of each margin is obtained by integrating out the rest of variables:

$$p(x) = \int_{R_Y} p(x, y) \mathrm{d}y.$$

$$p(y) = \int_{R_X} p(x, y) \mathrm{d}x.$$

Marginal probability distributions of Eye Color and Hair Color is given below:

	Hair color					Marginal distribution of Eye color
Eye color	Black	Brunette	Red	Blond	Marginal	
Brown	0.11	0.2	0.04	0.01	0.36	
Blue	0.03	0.14	0.03	0.16	0.36	
Hazel	0.03	0.09	0.03	0.02	0.17	
Green	0.01	0.05	0.02	0.03	0.11	
Marginal	0.18	0.48	0.12	0.22	1	

Marginal distribution of Hair color

Conditional distributions

In a joint distribution, to observe the effect of one variable on the other one we can use conditional distributions. In the probability context, we infer the probability of one outcome, given another outcome. The function of conditioning is to narrow down the space of interest by putting some information into the process.

Example

Suppose if you told that a person of interest has blue eyes. Conditional on this information, what would you say about the probability that the person has blond hair? At the beginning you have not known any thing about the person's both eye color and hair color. But now, you now that the person has blue eyes! The space is narrowed down and you have more information!

Suppose that the value of X is given as x, the probability that Y has the value of y is written as $P(y|x)$ and read as *the probability of y given x*. The corresponding conditional distribution is $p(X = x|Y = y)$. Simply, we write the conditional pdf or pf of X given y as $p(x|y)$.

Given joint probability distribution $(P(A, B))$ of events A and B,

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

Given joint pf or pdf $(p(x, y))$ of rv's X and Y,

$$p(x|y) = \frac{p(x, y)}{p(y)} \text{ or } p(y|x) = \frac{p(x, y)}{p(x)}.$$

It is also possible to write for discrete variables:

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x, y)}{\sum_{R_X} p(x, y)},$$

and

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\sum_{R_Y} p(x, y)},$$

and for continuous variables:

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(x, y)}{\int_{R_X} p(x, y)\mathrm{d}x},$$

and

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\int_{R_Y} p(x, y)\mathrm{d}y}.$$

It is inferred from the previous equations that

$$p(x, y) = p(x|y)p(y) \text{ and } p(x, y) = p(y|x)p(x).$$

Please be aware!

There is no temporal order in conditional probabilities. When we say "the probability of event A given B we do not mean that B has already happened and A has yet to happen. All we mean is that we are restricting our calculations of probability to a particular subset of possible outcomes.

A better gloss of $p(A|B)$ is, "among all joint outcomes with value B, this proportion of them also has value A." So, for example, we can talk about the conditional probability that it rained the previous night given that there are clouds the next morning. This is simply referring to the proportion of all cloudy mornings that had rain the night before.

Example

Conditional probabilities given the eye colour is as follows:

Eye color	Hair color				Marginal
	Black	Brunette	Red	Blond	
Hazel	0.03/0.17 = 0.18	0.09/0.17 = 0.52	0.03/0.17 = 0.18	0.02/0.17 = 0.12	1

Independence

When the occurrence of event A has no influence on the occurrence of event B, these events are thought to be independent. If the rv's X and Y are independent, the following equation must be satisfied:

$$p(x, y) = p(x) \cdot p(y).$$

This implies that

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

and

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x)p(y)}{p(x)} = p(y).$$

$$X \perp Y \Longrightarrow$$

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

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

Because the rv's X and Y are independent, the conditional distributions are equal to the marginal distributions: hence, rv's do not contain any information about each other. We do not have any reduction in our uncertainty due to the independence.

In our example, as you can easily check, because

$$P(blue) \cdot P(blond) = 0.36 \cdot 0.21 = 0.08 \neq 0.16 = P(blue, blond),$$

eye color and hair color are not independent.

Derivation of the Bayes' rule

The Bayes rule is itself a conditional probability. It can be derived by the use of following definition of conditional probability of the Cause given the Reason:

$$p(c|r) = \frac{p(r, c)}{p(r)},$$

which means that the conditional probability of c given r is equal to probability of observing both c and r at the same time relative to the probability of r.

When we multiply both sides of above equation by $p(r)$, we get

$$p(c|r)p(r) = p(r, c).$$

We also have the following analogously,

$$p(r|c)p(c) = p(r, c).$$

Thus, we have

$$p(c|r)p(r) = p(r|c)p(c),$$

and this gives us the Bayes rule:

$$p(c|r) = \frac{p(r|c)p(c)}{p(r)}.$$

When C is a discrete random variable, we have a complete system over the domain of C; hence, we re-write the above equation as

$$p(c|r) = \frac{p(r|c)p(c)}{\sum_{c \in R_c} p(r|c)p(c)}.$$

Here, we are restricting our attention only to the given value of result R, and then we normalize the probabilities in that result by dividing by the result's total probability.

You should note that we are working with random variables here. So, each variable has a known or an unknown probability distribution. And most of the probability distributions have parameters related with summary statistics of that distribution. In the frequentist class of statistics, these parameters are accepted and treated as unknown CONSTANTS. However, in the Bayesian class of statistics those parameters are also treated as RANDOM VARIABLES!

Example:

Suppose that in a general population, probability of having a disease is 0.001 and presence and absence of the disease is denoted by the parameter θ. So, our parameter θ is a random variable here and it has the support set $\{0, 1\}$; hence, the case $\theta = 0$ indicates that the disease is absent and $\theta = 1$ indicates that the disease is present. So, we have $P(\theta) = 0.001$. This represents our prior belief that the a randomly chosen person has the disease.

Suppose the test for this disease gives positive result for the person who has the disease 99% of the time. We define R as the random variable representing the test result and the domain of R is composed of $P : \{\text{Test result is positive}\}$ and $N : \{\text{Test result is negative}\}$. Now we can formally define the hit rate of the test as $P(R = P|\theta = 1) = 0.99$. Also, we suppose that the test has 5% false alarm rate which means that when the disease is absent, the test falsely indicates that the disease is present: $P(R = P|\theta = 0) = 0.05$.

Now, suppose that I have chosen a person randomly from the related population and applied the test and the outcome is positive. What about the probability that the person has the disease given the test result is positive. Formally, we are asking that $P(\theta = 1 | R = P) =$?

What is your guess about the value of this probability?

We apply the Bayes' rule to find this probability:

$$p(\theta = 1 | R = P) = \frac{p(R = P | \theta = 1)p(\theta = 1)}{p(R = P | \theta = 1)p(\theta = 1) + p(R = P | \theta = 0)p(\theta = 0)}$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.99 \cdot (1 - 0.001))} = 0.019$$

The interpretation of this result is that when the test result is positive, probability of having the disease is 0.019 with a 0.99 hit rate, or equivalently, posterior probability of having the disease with a positive test result is 0.019 with a 0.99 hit rate. This value is lower than our expectation. The cause of this situation is the low-prior probability of the disease. Because the person is randomly selected from a population and there is no other symptoms of the disease with the person. If there were some symptoms of the disease, we would define higher prior probabilities. Note that here we are re-allocating our credibilities according to the given data!

To make it clear notice that

With classical statistics, we can only say that

- given a person has the disease probability of getting a "positive" test result is 0.99! Say, Probability I to this.

However, in the Bayesian setting we can find

- the probability of having a disease given that the test result is positive! Say, Probability II to this.

We cannot find the Probability II without the application of Bayes theorem. But we can find Probability I (the value of which is 0.99) by finding 100 persons having the disease (we can easily observe the illness if someone has) and then applying the test to each person. If we get a positive result in 99 of 100 ill persons, we say that

- given a person has the disease probability of getting a positive test result is 0.99.

Because this probability is calculated over a sample of 100 persons having the disease, this is a classical inference.

However, we cannot find the value of

- probability of having a disease given that the test result is positive

by selecting 100 positive test results and observing the number of ill persons. Because we are not sure whether each person has the disease or not; that's why we are applying the test! We need to apply Bayes theorem to find this probability. This is the difference between conditional probability corresponding the classical and Bayesian inferences!

Summary

In this module, we worked through the basic context of probability. We discussed perception of probability for the representation of our degree of belief. In order to work with practical data, we defined random variable (recall we will use rv to refer a random variable throughout the course). We have mainly two kinds of rv's: discrete and continuous.

Each random variable has a population distribution that can be represented by a probability function (pf) or probability density function (pdf) for discrete and continuous rv's, respectively.

We studied expected value and variance of random variable associated with its pf or pdf and discussed interpretation of mean and variance in terms of specification of prior distribution.

Then we moved on to the joint distributions to represent interaction of two random variables on the same subject of interest. After talking on joint distributions, we mentioned marginal distributions of components of joint distributions. Based on joint distributions we worked our the context of conditional probability and independence.