Bayesian Methods for Data Science (DATS 6450 - 11)

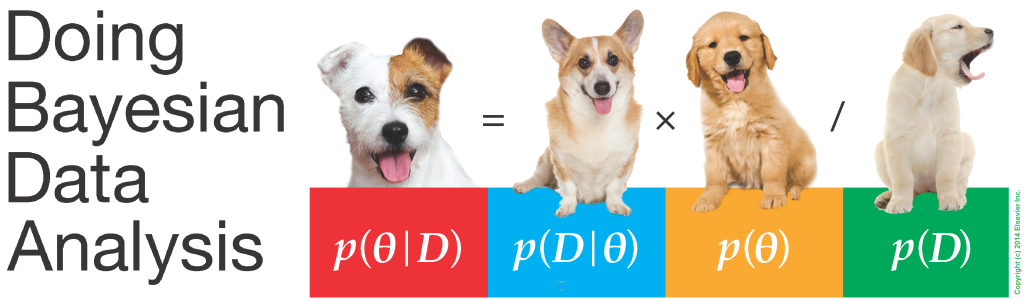
What is This Stuff Called Probability

Yuxiao Huang

Data Science, Columbian College of Arts & Sciences
George Washington University
yuxiaohuang@gwu.edu

September 11, 2019

Reference



Picture courtesy of the book website

- This set of slices is an excerpt of the book by Professor John K. Kruschke, with some trivial changes by the creator of the slides
- Please find the reference to and website of the book below:
  - Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier
  - `https://sites.google.com/site/doingbayesiandataanalysis/`

Overview

1. The set of all possible events
2. Probability: outside or inside the head
3. Probability distributions
4. Two-way distributions

The sample space

- The sample space is a set of events that are exhaustive and mutually exclusive
- The sample space may refer to:
  - data: when we flip a coin, we are sampling from the space of possible outcomes, {heads, tails}
  - parameters: when we grab a coin at random from a sack of coins, we are sampling from the space of possible biases, $[0, 1]$

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?
- **A:** Because coin flips represent every real-life event that has a binary outcome. **Q:** Any examples?

Coin flips: why you should care

- **Q:** Why should we care about coin flips and the statistics underneath?
- **A:** Because coin flips represent every real-life event that has a binary outcome. **Q:** Any examples?
- **A:**
  - for a heart surgery, whether it is successful or not
  - for a drug, whether it has side effect or not
  - for a survey question, whether the answer is correct or not
  - for a two-candidate election, whether one wins or not
  - ...

Two kinds of probabilities

- **Q:** Recall, what are the two kinds of sample space?

Two kinds of probabilities

- **Q:** Recall, what are the two kinds of sample space?
- **A:**
  - the sample space of the data (e.g., heads or tails)
  - the sample space of the parameters (e.g., bias)
- Each kind of sample space corresponds to a kind of probability
  - for data, the probability is over measurable outcomes that are "out there" in the world
  - for parameters, the probability is over unmeasurable beliefs that are "inside the head"

Probabilities assign numbers to possibilities

- A probability, no matter which kind it is, is just a way of assigning numbers to a set of exhaustive and mutually exclusive events
- A probability needs to satisfy three properties (Kolmogorov, 1956):
  - a probability value must be nonnegative
  - the sum of the probabilities across all events in the entire sample space must be 1
  - for any two mutually exclusive events, the probability that one or the other occurs is the sum of their individual probabilities

Probability distributions

- A probability distribution is simply a list of all possible events and their corresponding probabilities
- There are two kinds of probability distribution
  - discrete distribution: e.g., probabilities over heads or tails
  - continuous distribution: e.g., probabilities over people's heights

Discrete distributions: probability mass

- When the sample space consists of discrete outcomes (e.g., heads or tails), the probability distribution is a list of probabilities of the outcomes
- The probability of a discrete outcome is called as a probability mass
- The sum of the probability masses across the sample space must be 1
- Figure 2.1 (see next page) shows the probability masses of four suspects in the Holmes example

Figure 2.1

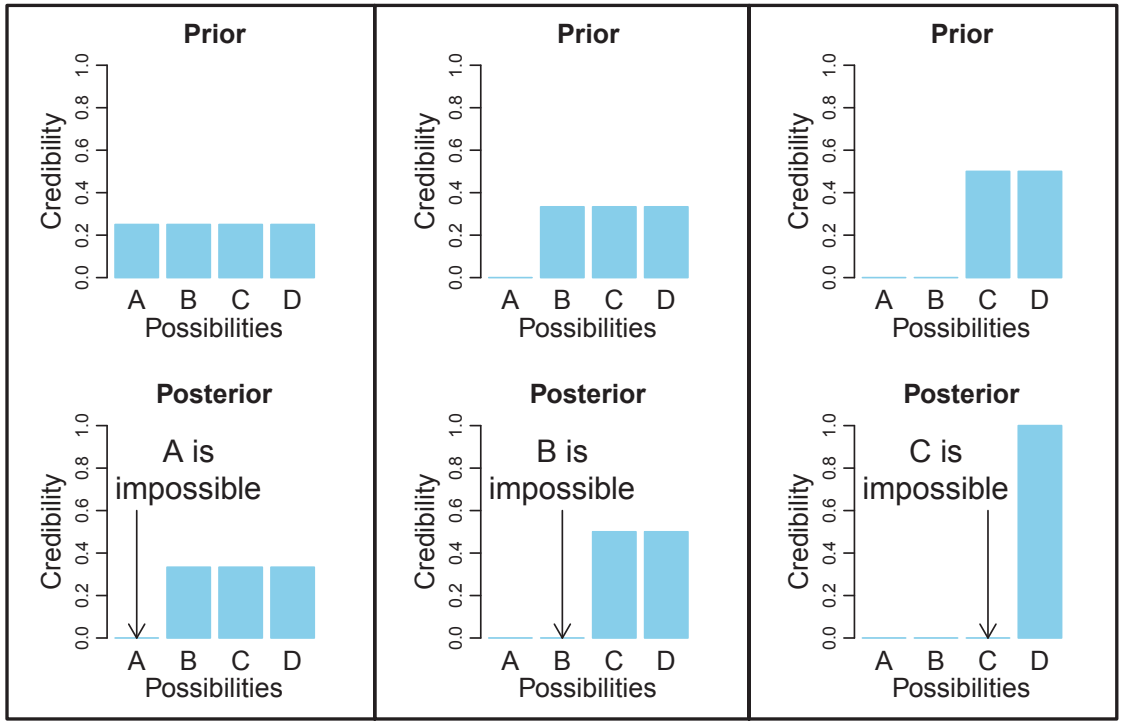


Figure 2.1: The upper-left graph shows the credibilities four possible causes for an outcome. The causes, labeled A, B, C and D, are mutually exclusive and exhaust all possibilities. The causes happen to be equally credible at the outset, hence all have prior credibility of 0.25. The lower-left graph shows the credibilities when one cause is learned to be impossible. The resulting posterior distribution is used as the prior distribution in the middle column, where another cause is learned to be impossible. The posterior distribution from the middle column is used as the prior distribution for the right column. The remaining possible cause is fully implicated by Bayesian re-allocation of credibility. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Continuous distributions: rendezvous with density

- When the sample space consists of continuous outcomes (e.g., people's heights), we cannot use probability mass for a specific outcome. **Q:** Why?

Continuous distributions: rendezvous with density

- When the sample space consists of continuous outcomes (e.g., people's heights), we cannot use probability mass for a specific outcome. **Q:** Why?
- **A:**
  - because the probability mass for a specific outcome will be zero
  - e.g., the probability of someone's height being 67.21413908...
- Instead, we can:
  1. discretize the space into a finite set of mutually exclusive and exhaustive intervals
  2. calculate the probability mass in each interval
  3. use the ratio of probability mass to interval width
  4. this ratio is called the probability density

Probability density

- The top panel of Figure 4.2 (see next page) shows the discretized intervals and probability mass in each interval
- The second panel shows the probability density
- The third panel shows the narrower intervals and probability mass in each interval
- The bottom panel shows the probability density corresponding to the narrower intervals
- Generally, the narrower the intervals are, the more accurate the probability density is

Figure 4.2

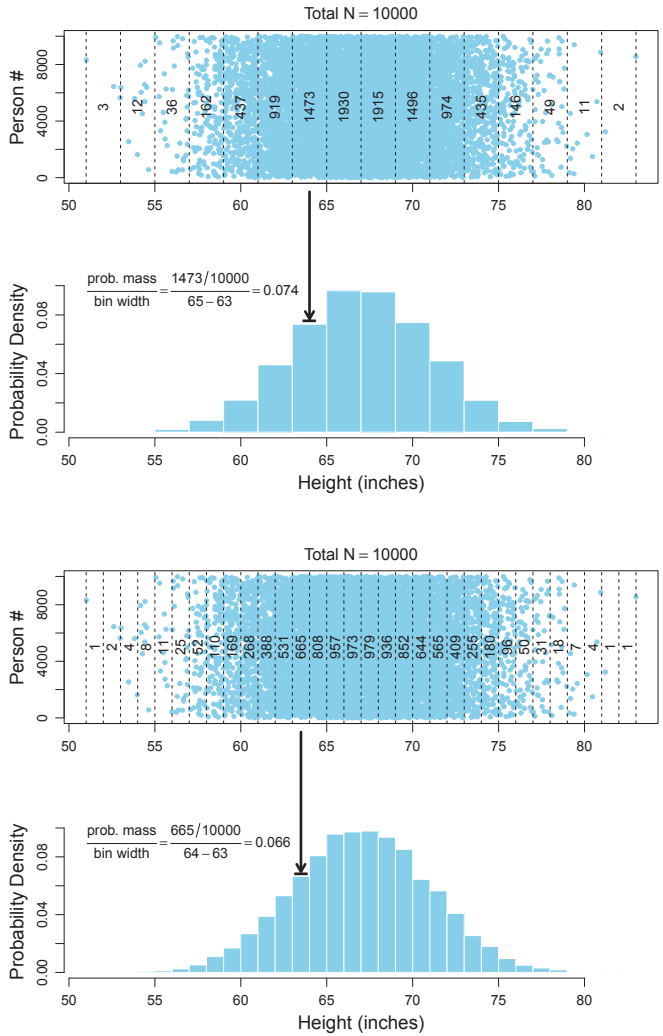


Figure 4.2: Examples of computing probability density. Within each main panel, the upper plot shows a scatter of 10,000 heights of randomly selected people, and the lower plot converts into probability density for the particular selection of bins depicted. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Probability density

- While probability mass cannot exceed 1, probability densities can
- The upper panel of Figure 4.3 (see next page) shows that most of the probability mass is concentrated around 84
- Consequently, the probability density near 84 exceeds 1.0, as shown in the lower panel
- This simply means that there is a high concentration of probability mass relative to the width of the interval

Figure 4.3

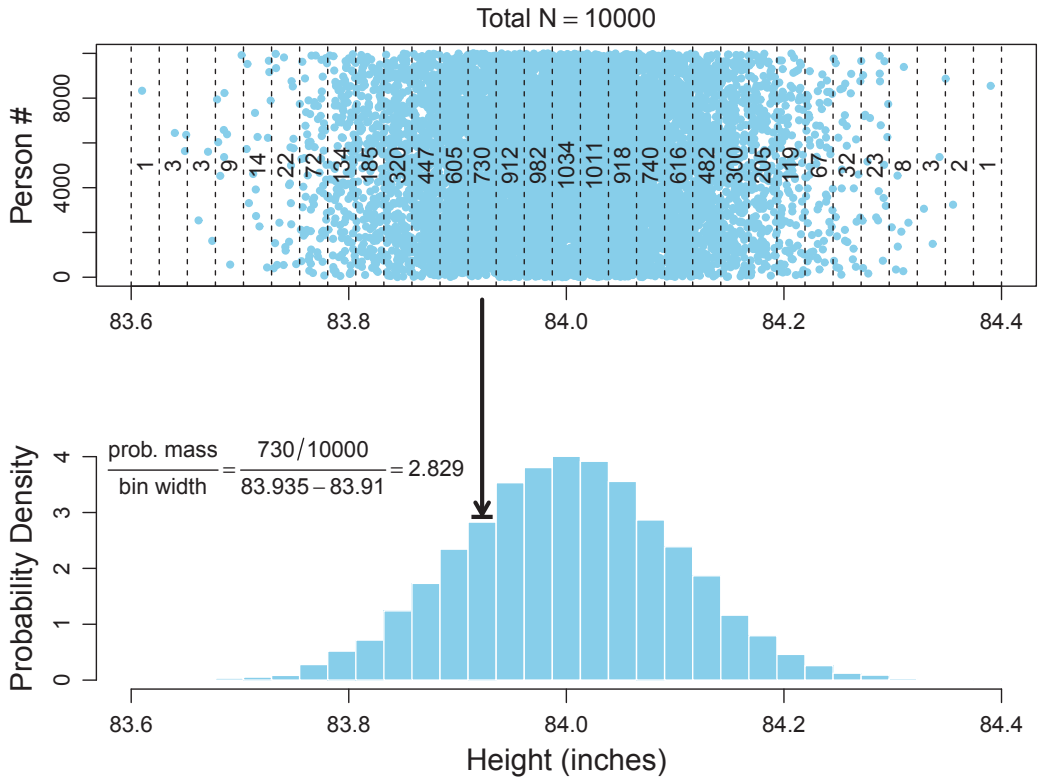


Figure 4.3: Example of probability density greater than 1.0. Here, all the probability mass is concentrated into a small region of the scale, and therefore the density can be high at some values of the scale. The annotated calculation of density uses rounded interval limits for display. (For this example, we can imagine that the points refer to manufactured doors instead of people, and therefore the *y*-axis of the top panel should be labelled "Door" instead of "Person.") Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Properties of probability density functions

- We need to define some notations first. Let:
  - x be the continuous variable
  - Δx be the width of an interval on x
  - i be an index for the intervals
  - $[x_i, x_i + \Delta x]$ be the interval between x_i and $x_i + \Delta x$
  - $p([x_i, x_i + \Delta x])$ be the probability mass of the ith interval
- Then the sum of those probability masses must be 1:

$$\sum_i p([x_i, x_i + \Delta x]) = 1$$

- We can rewrite the equation above in terms of the density of each interval, by dividing and multiplying by Δx:

$$\sum_i \Delta x \frac{p([x_i, x_i + \Delta x])}{\Delta x} = 1$$

Properties of probability density functions

- In the limit, as the interval width becomes infinitesimal, we denote:
  - summation as $\int$ instead of $\sum$
  - the width of the interval around x as $\mathrm{d}x$ instead of Δx
  - the probability density in the infinitesimal interval around x as $p(x)$
- Then the previous equation (in terms of density) can be rewritten as:

$$\sum_i \Delta x \frac{p([x_i, x_i + \Delta x])}{\Delta x} = 1 \qquad \Rightarrow \qquad \int \mathrm{d}x \; p(x) = 1$$

- By an abuse of notation, we use $p(x)$ to represent the probability mass when x is discrete
- Thus, what $p(x)$ represents depends on the context (x being discrete or continuous)

The normal probability density function

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]\right)^2$$

- **Q:** Recall, what are μ and σ? what do they control?

The normal probability density function

- Perhaps the most famous probability density function is the normal distribution, also known as the Gaussian distribution
- The probability density function of normal distribution is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]\right)^2$$

- **Q:** Recall, what are μ and σ? what do they control?
- **A:**
  - μ: the mean/mode/median, location parameter
  - σ: the standard deviation, scale parameter
- An example of the probability density is shown in Figure 4.4 (see next page), where x axis is divided into a dense comb of small intervals
- The figure also shows that the area under the curve is, in fact, 1

Figure 4.4

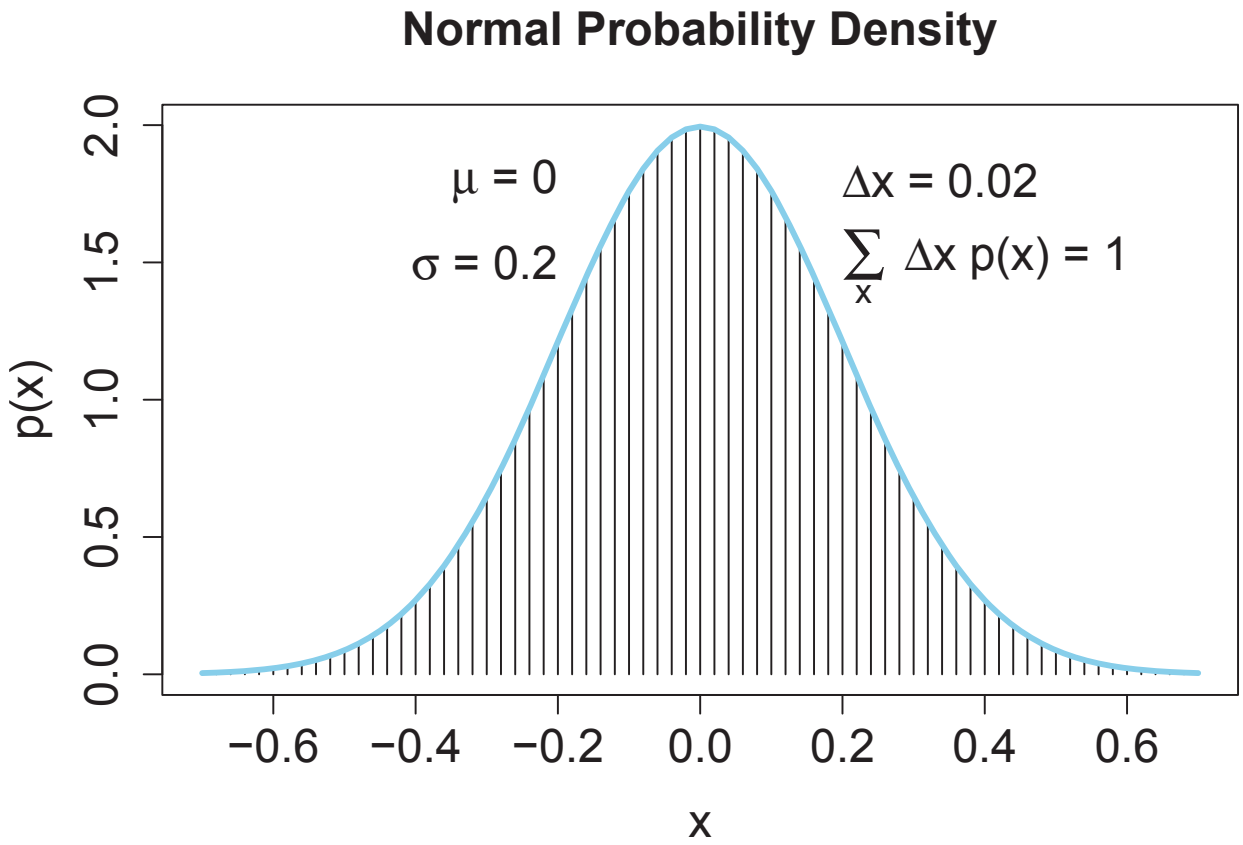


Figure 4.4: A normal probability density function, shown with a comb of narrow intervals. The integral is approximated by summing the width times height of each interval. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Mean of a distribution

- The mean of a probability distribution is the long-run average of the values
- The mean is also called the expected value, denoted by $E[x]$
  - when x is discrete:

$$E[x] = \sum_x p(x)x$$

  - when x is continuous:

$$E[x] = \int \mathrm{d}x \; p(x)x$$

Variance of a distribution

- The variance of a probability distribution is a number that represents the dispersion of the distribution away from its mean
- The definition of variance, var_x, is the mean squared deviation (MSD) of the x values from their mean
  - when x is discrete:

$$\mathrm{var}_x = \sum_x p(x)(x - E[x])^2$$

  - when x is continuous:

$$\mathrm{var}_x = \int \mathrm{d}x \; p(x)(x - E[x])^2$$

- In other words, the variance is just the average value of $(x - E[x])^2$

Highest density interval (HDI)

- **Q:** Recall, what is HDI?

Highest density interval (HDI)

- **Q:** Recall, what is HDI?
- **A:** As shown in Figure 4.5 (see next page), HDI is an interval that spans most of the distribution, say 95% of it, such that every point inside the interval has higher probability than any point outside the interval
- Formally, the values of x in the 95% HDI are those such that $p(x) > W$ where W satisfies

$$\int_{x:p(x)>W} \mathrm{d}x \; p(x) = 0.95$$

- When the distribution refers to probability of values, then the width of the HDI is another way of measuring uncertainty of beliefs
  - if the HDI is wide, then beliefs are uncertain
  - if the HDI is narrow, then beliefs are relatively certain

Figure 4.5

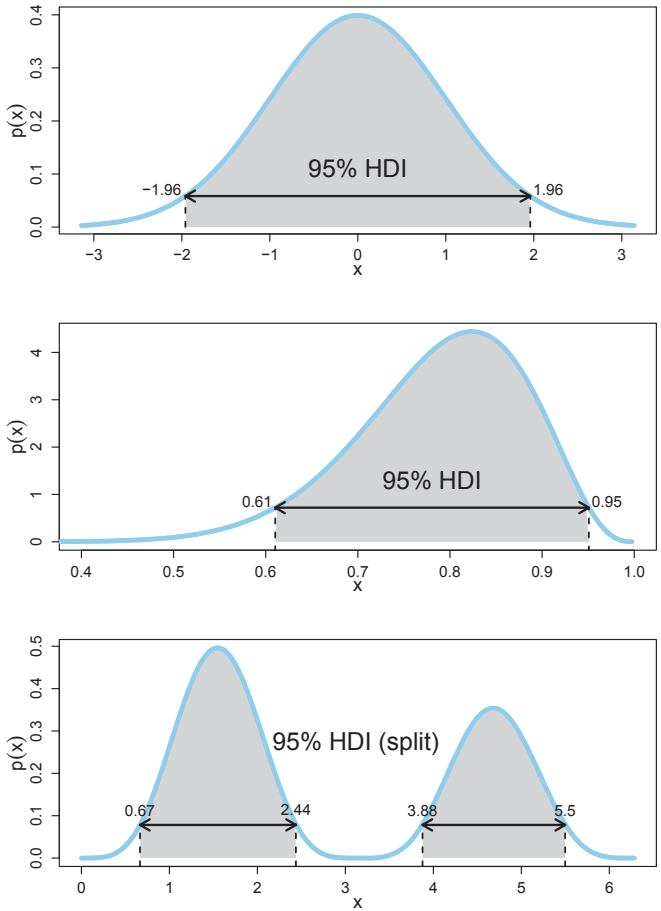


Figure 4.5: Examples of 95% highest density intervals (HDIs). For each example, all the x values inside the interval have higher density than any x value outside the interval, and the total mass of the points inside the interval is 95%. The 95% area is shaded, and it includes the zone below the horizontal arrow. The horizontal arrow indicates the width of the 95% HDI, with its ends annotated by (rounded) x values. The height of the horizontal arrow marks the minimal density exceeded by all x values inside the 95% HDI. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Joint probability and marginal probability

- Table 4.1 (see next page) shows the probabilities of various combinations of people's eye color and hair color
- Each entry indicates the joint probability of particular combinations of eye color (e) and hair color (h), denoted by $p(e, h)$
- The right margin of the table shows the probabilities of the eye colors overall, collapsed across hair colors
- Such probabilities are called marginal probability, denoted by $p(e)$:

$$p(e) = \sum_h p(e, h)$$

- The marginal probabilities of the hair colors, $p(h)$, are indicated on the lower margin of the table:

$$p(h) = \sum_e p(e, h)$$

Table 4.1

Table 4.1: Proportions of combinations of hair color and eye color. Some rows or columns may not sum exactly to their displayed marginals because of rounding error from the original data. Data adapted from Snee (1974). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal (Eye Color)
	Black	Brunette	Red	Blond	
Brown	.11	.20	.04	.01	.37
Blue	.03	.14	.03	.16	.36
Hazel	.03	.09	.02	.02	.16
Green	.01	.05	.02	.03	.11
Marginal (Hair Color)	.18	.48	.12	.21	1.0

Conditional probability

- We may want to know the probability of an outcome y, given some evidence x, $p(y|x)$
- This is sometimes called the conditional probability of y given x, and can be calculated as

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

  that is, the number of instances where both x and y being true, out of the number where x being true
- Table 4.2 (see next page) shows the conditional probability of a hair color, given the eye color being blue

Table 4.2

Table 4.2: Example of conditional probability. Of the blue-eyed people in Table 4.1, what proportion have hair color h? Each cell shows $p(h|\text{blue}) = p(\text{blue}, h)/p(\text{blue})$ rounded to two decimal points. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan. 2nd Edition.* Academic Press / Elsevier.

Eye Color	Hair Color				Marginal (Eye Color)
	Black	Brunette	Red	Blond	
Blue	.03/.36 = .08	.14/.36 = .39	.03/.36 = .08	.16/.36 = .45	.36/.36 = 1.0

Absolute independence

- We say x and y are absolutely independent, if

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

- Examples of absolute independence include:
  - eye color and height
  - hair color and weight
  - ...

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?
- **A:**

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

- **Q:** What is the overall size of the marginal probabilities?

Absolute independence

- Absolute independence is powerful in terms of simplifying inference
- Consider n binary variables, $x_1, \ldots, x_n$
- **Q:** What is the size of joint probability distribution, $p(x_1, \ldots, x_n)$
- **A:** 2^n
- **Q:** When the variables are absolutely independent, can we convert the joint probability to marginal probabilities, $p(x_i)$?
- **A:**

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

- **Q:** What is the overall size of the marginal probabilities?
- **A:** $2n$
- Absolute independence reduces the size from 2^n to $2n$!

Conditional Independence

- While powerful, absolute independence is rare
- It is more common to have conditional independence
- We say x and y are conditionally independent given z, if

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

- Examples of conditional independence include:
  - yellow finger and lung cancer, given the status of smoking
  - eye color and hair color, given the genes
  - ...

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?
- **A:**

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1})$$

- **Q:** What is the overall size of the conditional probabilities?

Conditional Independence

- Conditional independence is also powerful in terms of simplifying inference (although less powerful than absolute independence)
- Again, consider n binary variables, $x_1, \ldots, x_n$
- **Q:** When x_i (where $i > 1$) is conditionally independent with others, given x_{i-1}, can we convert the joint probability to conditional probabilities, $p(x_i|x_{i-1})$?
- **A:**

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i|x_{i-1})$$

- **Q:** What is the overall size of the conditional probabilities?
- **A:** $4n - 2$
- Conditional independence reduces the size from 2^n to $4n - 2$!