Doing Bayesian Data Analysis §5.1~5.2

Chao-En Yu

National Tsing Hua University

Reallocation of Credibility

- On a typical day at your location, what is the probability that it is cloudy?

- Suppose you are told it is raining, now what is the probability that it is cloudy?

- We can be pretty sure that

$$p(\text{cloudy}) < p(\text{cloudy}|\text{raining}).$$

Human Reasoning

- Suppose instead you are told that everyone outside is wearing sunglasses, what is the probability that it is cloudy?

- Most likely, it is true that

$$p(\text{cloudy}) > p(\text{cloudy}|\text{sunglass}).$$

- Notice how we human beings have reasoned in this example.

Bayes' Rule

- Bayes' rule is merely the mathematical relation between the prior allocation of credibility and the posterior reallocation of credibility conditional on data.

- From the definition of conditional probability,

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

i.e., the probability that r and c happen together relative to the probability that r happens at all.

Only a condition

Derivation

- Notice that by definition we also have

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

← Marginal probability

- Multiplying both sides by $p(r)$ and $p(c)$, respectively, gives

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

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

which is called *Bayes' rule*.

Bayes' Rule

- We can further re-write the denominator in terms of $p(r|c)$:

$$p(c|r) = \frac{p(r|c)p(c)}{\sum_{c^*} p(r|c^*)p(c^*)}$$

- The c in the numerator is a specific fixed value, whereas the c^* in the denominator is a variable that takes on all possible values.

- $p(c|r)$ is called posterior; $p(c)$ is prior; $p(r|c)$ is likelihood; $p(r)$ is marginal likelihood (evidence).

Intuition of Bayes Rule

- Without knowing anything about a person's eye color, all we believe about hair colors is expressed by the marginal probabilities at the bottom.

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal (Eye color)
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

Intuition of Bayes Rule

- If we are told that the selected person's eyes are blue, we can focus out attention on the Blue row.

- Notice that we have gone from the "prior" (marginal) beliefs about hair color before knowing eye color, to the "posterior" (conditional) beliefs.

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal (Eye color)
Blue	$0.03/0.36 = 0.08$	$0.14/0.36 = 0.39$	$0.03/0.36 = 0.08$	$0.16/0.36 = 0.45$	$0.36/0.36 = 1.0$

Prosecutor's Fallacy

- A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. 85% of the cabs in the city are Green and 15% are Blue.
- A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.
- What is the probability that the cab involved in the accident was blue rather green?

$$p(\theta = \text{☹} \mid T = +) = ?$$

Test result	Disease: $\theta =$ ☹ (present)	Disease: $\theta =$ ☺ (absent)	Marginal (test result)
$T = +$	$p(+\mid \text{☹})\, p(\text{☹}) = 0.99 \cdot 0.001$	$p(+\mid \text{☺})\, p(\text{☺}) = 0.05 \cdot (1 - 0.001)$	$\sum_\theta p(+\mid\theta)\, p(\theta)$
$T = -$	$p(-\mid \text{☹})\, p(\text{☹}) = (1 - 0.99) \cdot 0.001$	$p(-\mid \text{☺})\, p(\text{☺}) = (1 - 0.05) \cdot (1 - 0.001)$	$\sum_\theta p(-\mid\theta)\, p(\theta)$
Marginal (disease)	$p(\text{☹}) = 0.001$	$p(\text{☺}) = 1 - 0.001$	1.0

Hit Rate vs False Alarm

	Disease	
Test result	$\theta =$ ☹ (present)	$\theta =$ ☺ (absent)
$T = +$	True Positive (Hit Rate)	False Positive (False Alarm)
$T = -$	False Negative	True Negative

Parametric Model

- The key application of Bayes' rule is when the row variable represents data values and the column variable represents parameter values.

- A Bayesian model specifies
  $p(\text{data values}|\text{parameter values})$
  along with the prior, $p(parameter\ values)$

- We use Bayes' rule to convert that to know
  $p(\text{parameter values}|\text{data values})$ ← How strongly we should believe in the various parameter values given the data.

Parameter Inference

- The posterior $p(\theta|D)$ is obtained by conditionalizing on the row with the observed data value, and that operation is Baye's rule.

	Model parameter			
Data	...	**θ value**	...	**Marginal**
⋮		⋮		⋮
***D* value**	...	$p(D,\theta) = p(D\|\theta)\, p(\theta)$	...	$p(D) = \sum_{\theta^*} p(D\|\theta^*)\, p(\theta^*)$
⋮		⋮		⋮
Marginal	...	$p(\theta)$	...	

Caveats

- For continuous parameters, the only change in Bayes' rule is that the marginal likelihood changes from the sum to an integral:

$$p(D) = \int d\theta^* \, p(D|\theta^*)p(\theta^*)$$

- Suppose we observe some data D and D', does our final belief depend on whether we update with D first and D' second, or update with D' first and D second?

Statistical Independence

- When $p(x,\ y) = p(x) * p(y), \forall x, y$, we say that x and y are statistically independent.

- By Bayes' rule, x and y are statistically independent if $p(y \mid x) = p(y), \forall x, y$.

- Bayesian prefers the second definition for it relates probability to an addition of information, though independence is often a convenient assumption.

Data-Order Invariance

- The answer is: It depends!

- When the data probabilities are independent, $p(D, D'|\theta) = p(D|\theta) \cdot p(D'|\theta)$, then the order of updating has no effect of the final posterior.

★ Intuition: If the likelihood function has no dependence on data ordering, then the posterior shouldn't have any dependence on data ordering.

Proof of Order-Invariance

- Equivalent to assume that datum is equally representative of the underlying process, regardless of when the datum was observed, and regardless of any data observed before or after.

$$p(\theta|D', D) = \frac{p(D', D|\theta)\, p(\theta)}{\sum_{\theta^*} p(D', D|\theta^*)\, p(\theta^*)} \qquad \text{Bayes' rule}$$

$$= \frac{p(D'|\theta)p(D|\theta)\, p(\theta)}{\sum_{\theta^*} p(D'|\theta^*)p(D|\theta^*)\, p(\theta^*)} \qquad \text{by assumption of independence}$$

$$= \frac{p(D|\theta)p(D'|\theta)\, p(\theta)}{\sum_{\theta^*} p(D|\theta^*)p(D'|\theta^*)\, p(\theta^*)} \qquad \text{multiplication is commutative}$$

$$= p(\theta|D, D') \qquad \text{Bayes' rule}$$

Doing Bayesian Data Analysis §4.3~4.4

Chao-En Yu

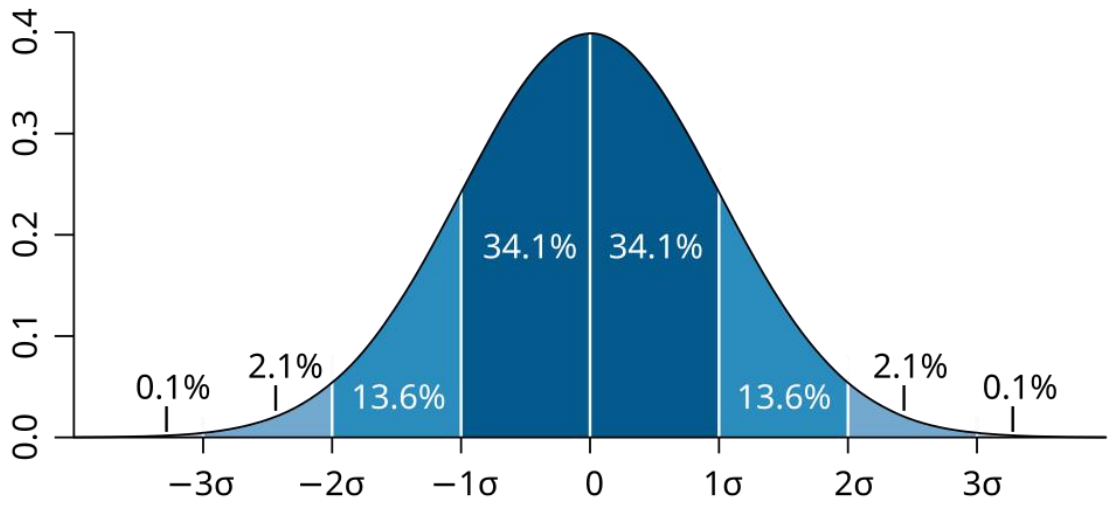
National Tsing Hua University

Probability Distributions

- A probability distribution is a theoretical and mathematical function or table that describes the likelihood of all possible outcomes for a *random variable*.

Can be discrete or continuous.

- A probability, whether it's outside the head or inside the head, is just a way of assigning numbers to a set of mutually exclusive possibilities.



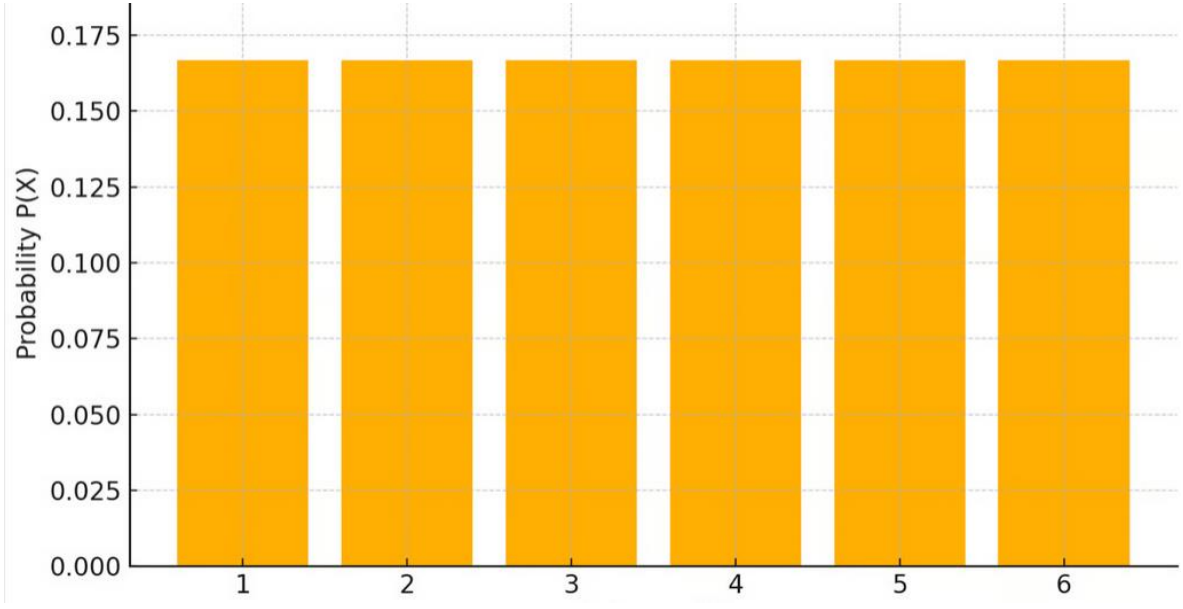
Probability Mass Function

- The probability mass function (PMF) is a relationship that defines the probabilities associated with each possible outcome of a discrete random variable.

the horizontal axis

- PMFs $P(x)$ are visualized using bar charts (長條圖).

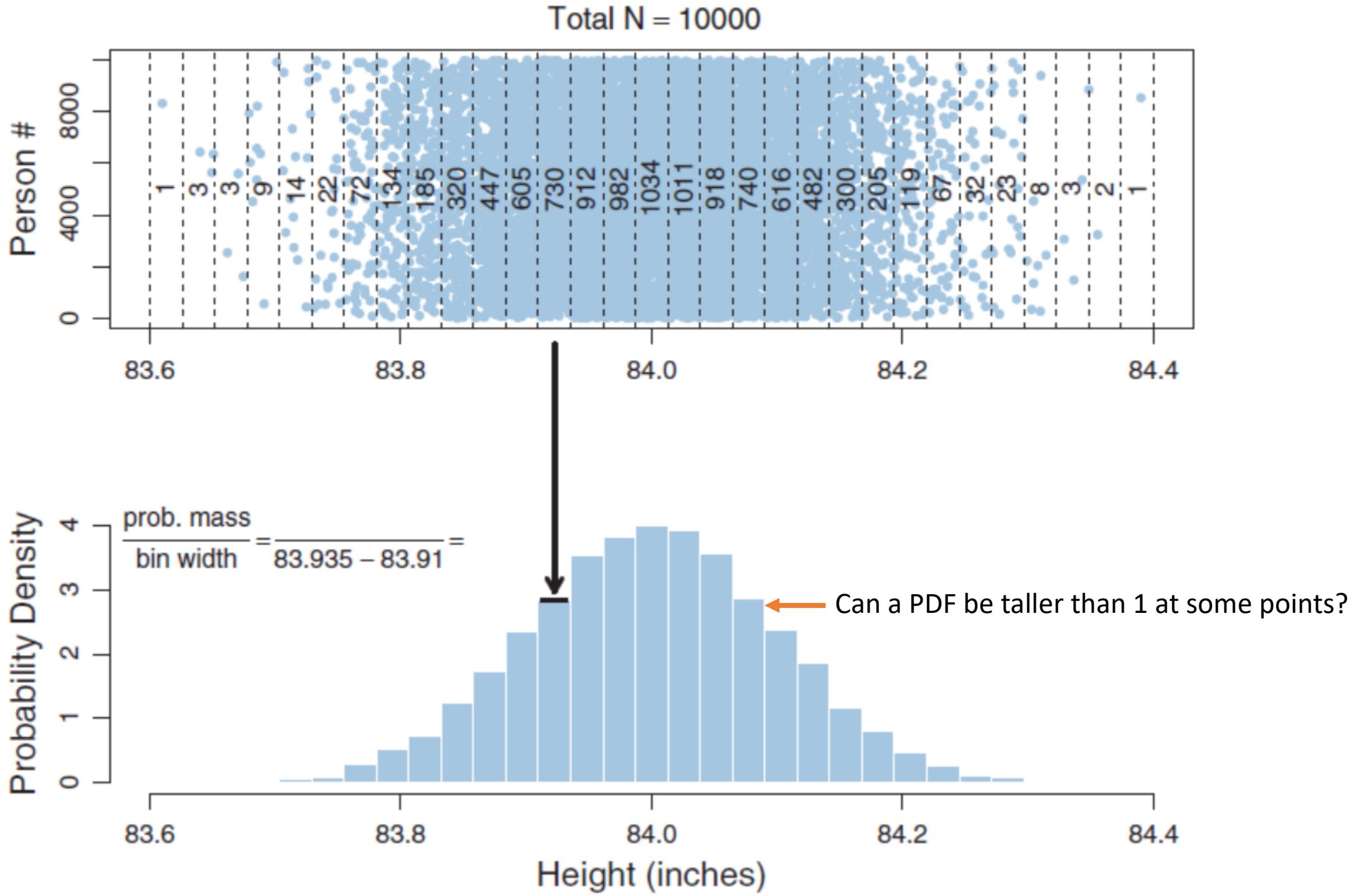
- The vertical axis represents the probability assigned by the PMF.



Probability Density Function

- The probability density function (PDF) defines the probability of a continuous random variable lying between a specific range of values.
- PDFs $f(x)$ are visualized using histograms (直方圖).
- We make the intervals infinitesimally narrow, and talk about the ratio of the probability mass to the interval width. That ratio is called the *probability density*.

Is the density a probability? Why?


Total N = 10000
Person #
0
4000
8000
1
3
3
9
14
22
72
134
185
320
447
605
730
912
982
1034
1011
918
740
616
482
300
205
119
67
32
23
8
3
2
1
83.6
83.8
84.0
84.2
84.4
Probability Density
0
1
2
3
4
prob. mass / bin width = / 83.935 – 83.91 =
Can a PDF be taller than 1 at some points?
83.6
83.8
84.0
84.2
84.4
Height (inches)

Properties of PDF

- The probability mass of the ith interval is denoted $p([x_i, x_i + \Delta x])$. Then the sum of those probability masses must be 1:

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

- The probability density is the ratio of probability mass over interval width, so we have

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

Integral of PDF

- Do not confuse the probability density $p(x)$ with $p([x_i, x_i + \Delta x])$, which is the probability mass in an interval.

Thus, the density is often denoted by $f(x)$

- The above summation equation becomes an integral:

$$\sum_i \Delta x \, \frac{p([x_i, x_i + \Delta x])}{\Delta x} = 1 \text{ that is, } \int dx \, p(x) = 1.$$

$\int$ dx $p(x)$

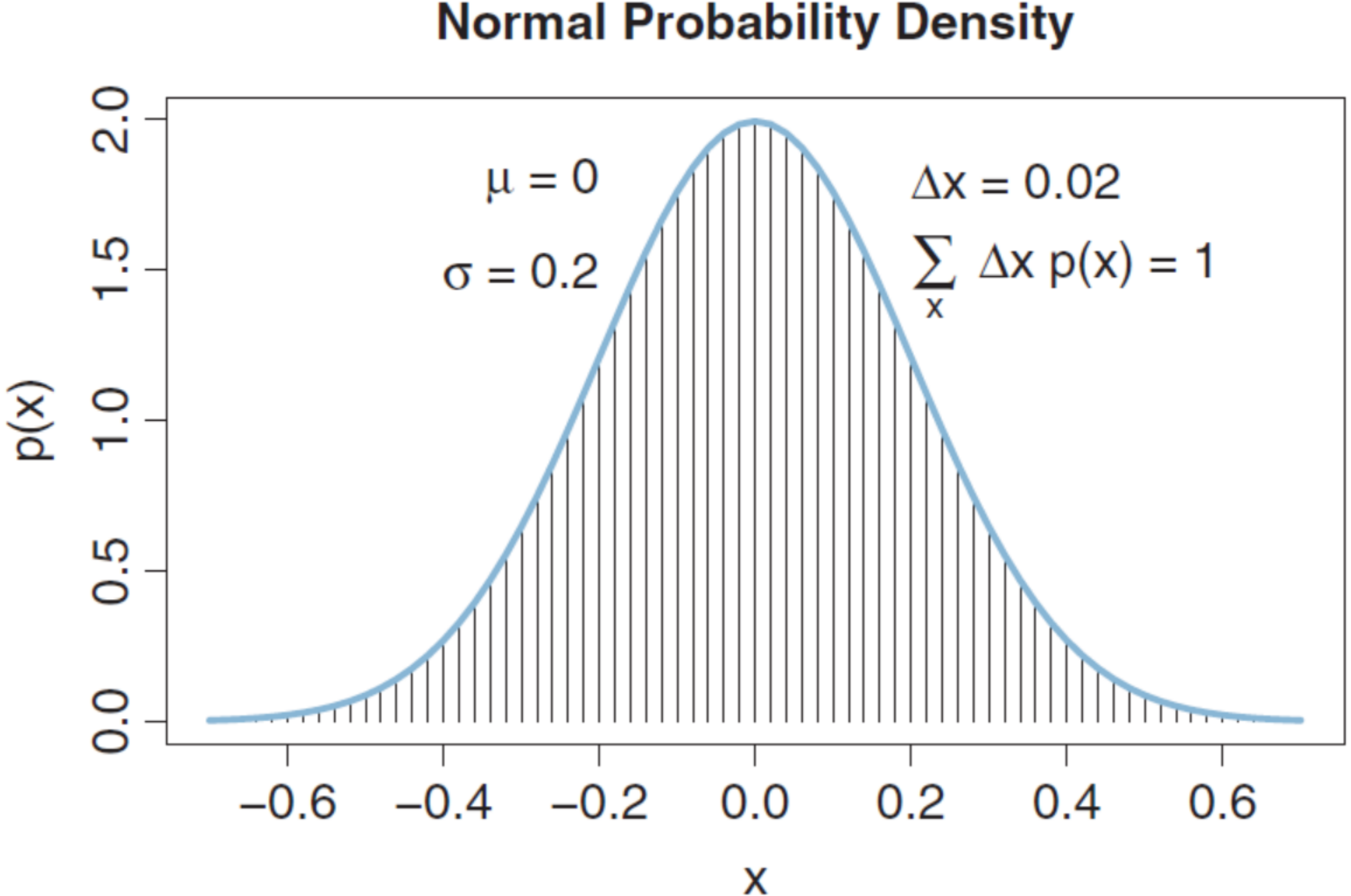
The Normal Distribution

- Any function that has only nonnegative values and integrates to 1 can be a probability density function.

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

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

- Two parameters: The mean (μ) and the standard deviation (σ governs how wide the bell is).


Normal Probability Density
μ = 0
σ = 0.2
Δx = 0.02
∑ Δx p(x) = 1
x
p(x)
0.0
0.5
1.0
1.5
2.0
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
x

Mean of a Distribution

- The *mean* of a probability distribution (also called the expected value) is an average value in the long run, weighted by the probability of each outcome:

$$E[x] = \sum_x p(x)\, x \quad \text{or} \quad E[x] = \int dx\, p(x)\, x$$

- Ex 1. $p(x) = 6x(1-x)$ for $x \in [0,1]$.

- Ex 2. Can you verify that $E[x] = \mu$ for the standard normal distribution ($\mu = 0, \sigma = 1$)?

Variance of a Distribution

- The *variance* of a probability distribution σ^2 is a number that represents the dispersion of the distribution away from its mean.

- The definition of variance is the mean squared deviation (MSD) of the x values from their mean:

$$\mathrm{Var}[x] = \int dx\, p(x)\, (x - E[x])^2$$

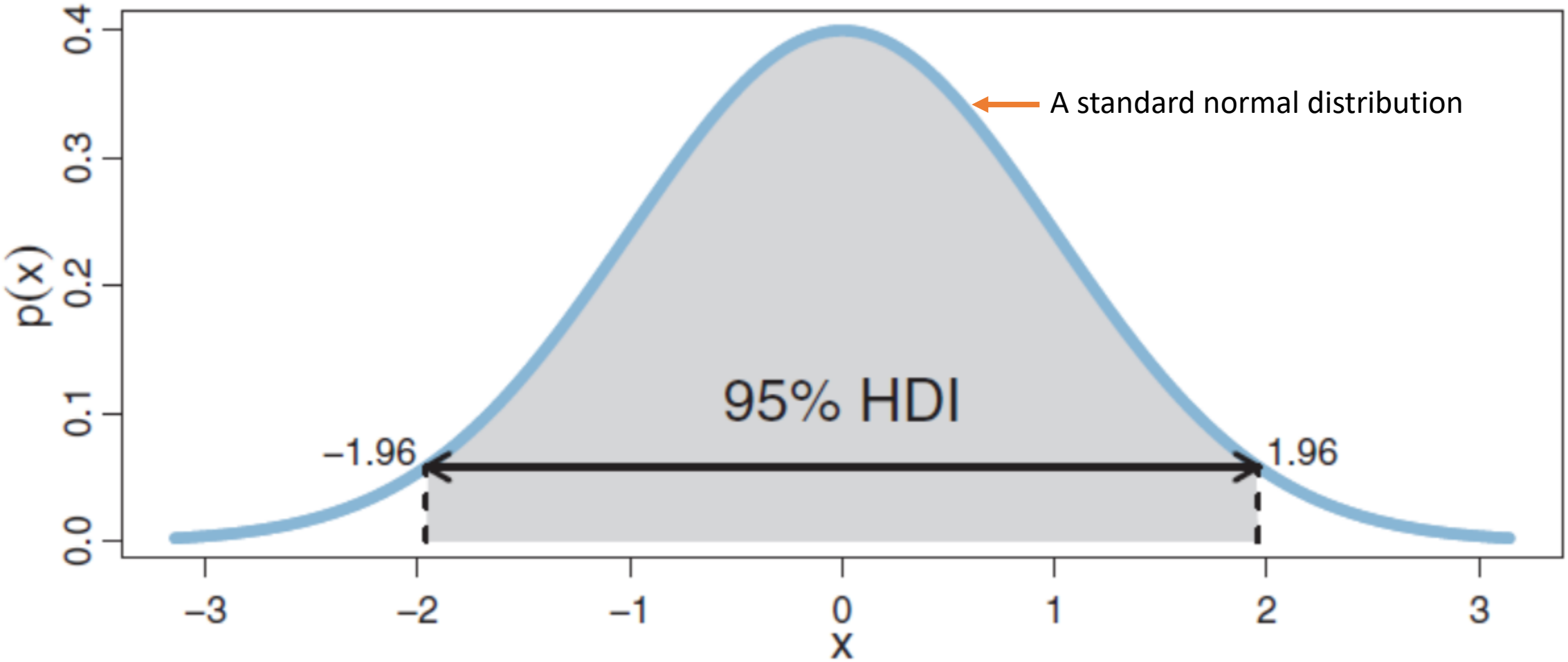
- Exercise 4.5. What is the probability mass under the standard normal curve from $x = \mu - \sigma$ to $x = \mu + \sigma$?

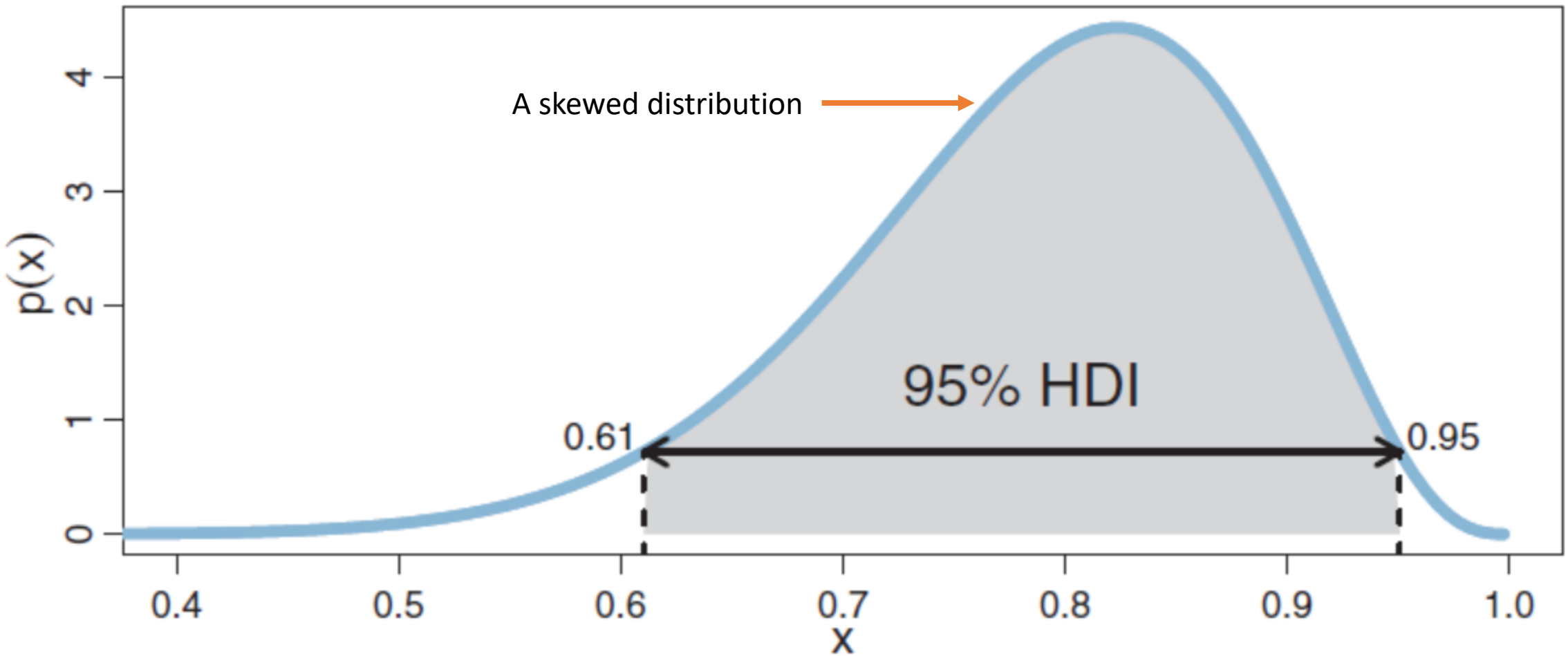
Standard Deviation

- The square root of the variance, sometimes referred to as root mean squared deviation (RMSD), is called the *standard deviation* of the distribution..
- A probability distribution can refer to probability of measurement values or of parameter 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 (HDI)

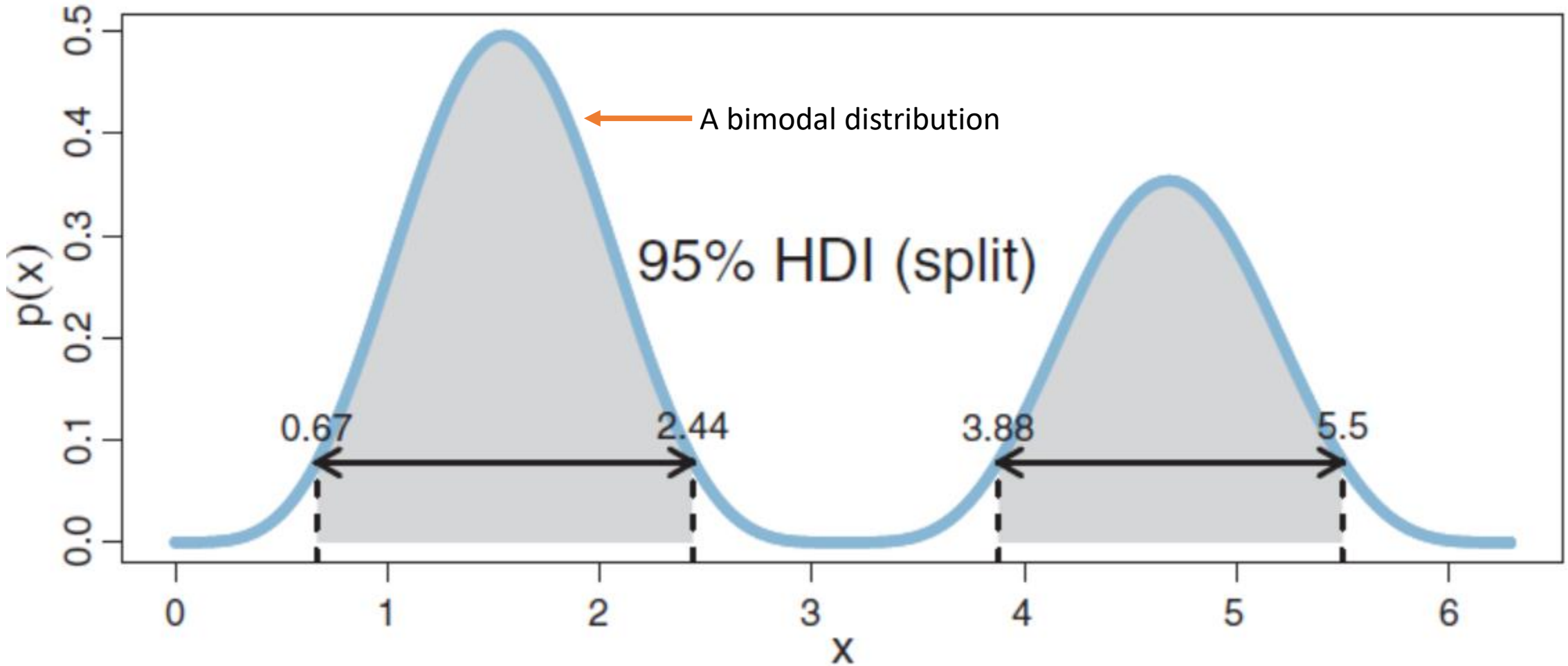
- Another way of summarizing a distribution is the *highest density interval* (HDI).
- The HDI indicates which points of a distribution are most *credible*. The area under the curve between the HDI limits is, say 0.95 (or ≥ 0.95 for p.m.f.s).
- Moreover, every point inside the HDI has higher probability density than any point outside the interval.


A standard normal distribution
95% HDI
−1.96
1.96
p(x)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x


A skewed distribution
95% HDI
0.61
0.95
p(x)
0
1
2
3
4
0.4
0.5
0.6
0.7
0.8
0.9
1.0
x

Caveats

- The HDI does not necessarily produce equal-area tails outside the HDI (but confidence interval does).
- With a multimodal distribution, the HDI is split into two subintervals, not good for statistical estimation.
- If the HDI is wide, then beliefs are uncertain. If the HDI is narrow, then beliefs are relatively certain.


A bimodal distribution
95% HDI (split)
0.67
2.44
3.88
5.5
p(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6