Principles and Practices of Data Science

Lecture 8

Melvin Ayala

Lecture 8: Probability, Uncertainty, and Information

Sections

1. Probability
  1.1. Preliminaries
  1.2. Random Events
  1.3. Discrete Random Processes
    1.3.1. The Poisson Process
    1.3.2. The Bernoulli Process
  1.4. Continuous Random Processes
    1.4.1. Preliminaries
    1.4.2. The Normal Distribution
    1.4.3. Other Types of Distributions
      1.4.3.1. The Exponential Distribution
      1.4.3.2. The Gamma Distribution
      1.4.3.3 The Weibull Distribution
  1.5. Central Limit Theorem
  1.6. Histograms
2. Uncertainty
3. Information Theory

1. Probability

1.1. Preliminaries

Brief Definition of Probability of an Event

- the extent to which an event is likely to occur, measured by the ratio of the favorable cases to the whole number of possible cases.

$$\text{Probability of an event} = \frac{\text{Number of favorable cases}}{\text{Number of possible cases}}$$

1.2. Random Events

Random events can be:

- independent (each event is not affected by other events),
- dependent (also called "conditional", where an event is affected by other events)
- mutually exclusive (events can't happen at the same time)

1.3. Discrete Random Variables

Discrete variable:

- is a variable that can "only" take-on certain numbers on the number line.
- the sum of the probabilities of each of its possible values is equal to 1.

Examples:

- count of events in a different interval
- rolling a dice and annotating the number we get (1, 2, 3, 4, 5, 6)
- the number of rain drops that fall over a square kilometer on a particular month in a particular country

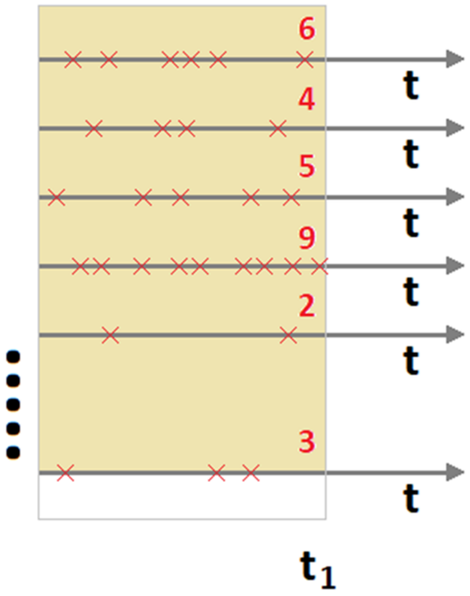
Outcomes when rolling a dice:

- roll a single dice → possible outcomes: 1, 2, 3, 4, 5, 6
- probability of each outcome = 1/6.
- this is a discrete random variable since: 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1

1.3.1. The Poisson Process

Poisson process:

- one of the most widely-used counting processes.
- used to count the occurrences of certain events that appear to happen at a certain rate, but completely at random (without a certain structure).
- a model for a series of discrete events where the average time between events is known, but the exact timing of events is random.
- the arrival of an event is independent of the event before (waiting time between events is memoryless).



Events counted in interval $(0, t_1)$:

$X = \{6, 4, 5, 9, 2, \ldots, 3\}$

$$P(\text{k events in interval t}) = P(X = k, t) \\ = ?$$

Examples:

- the number of car accidents at a site or in an area
- the location of users in a wireless network
- the requests for individual documents on a web server
- the outbreak of wars
- photons landing on a photodiode
- the number of meteorites greater than 1 meter diameter that strike Earth in a year
- the number of laser photons hitting a detector in a particular time interval
- the number of students achieving a low and high mark in an exam

Poisson Distribution

- a discrete probability distribution used for modelling Poisson processes
- expresses the probability of a given number of events occurring in a fixed interval of time or space
- these events occur with a known constant mean rate and independently of the time since the last event.

Probability density function

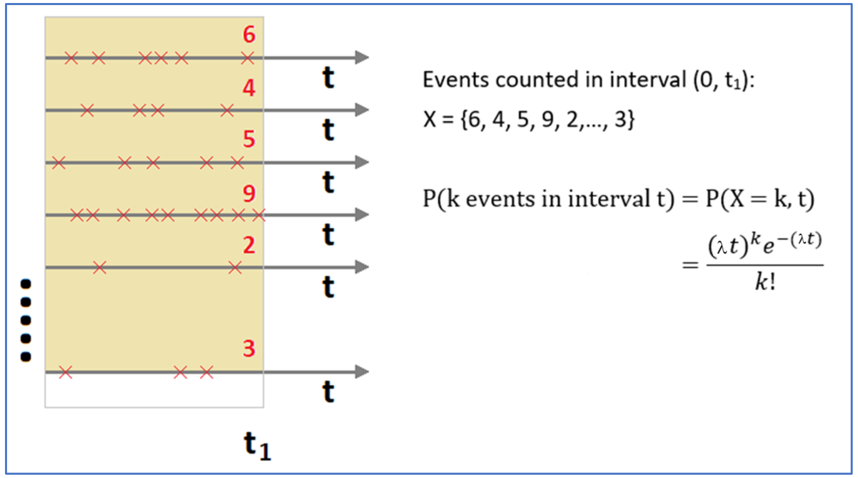
- a discrete random variable X is said to have a Poisson distribution, with parameter λ > 0, if it has a probability distribution function given by:

$$\mathrm{f(k, \lambda){=}P(X{=}k){=}}\ \frac{\lambda^k e^{-\lambda}}{k!}$$

- when k depends on time:

$$\mathrm{f(k(t), \lambda){=}P(X(t){=}k){=}}\ \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

where:
k is the number of occurrences (= 0, 1, 2, 3, …)
e is Euler's number (e = 2.71828)
! is the factorial function.



Events counted in interval (0, t_1):
X = {6, 4, 5, 9, 2,…, 3}

$$\mathrm{P(k\ events\ in\ interval\ t) = P(X = k, t)} = \frac{(\lambda t)^k e^{-(\lambda t)}}{k!}$$

Poisson Distribution: A Fishing Example

Suppose that people catch fish in a certain river according to a Poisson process at a rate of 3 per hour.
What is the probability that you catch exactly 4 fish in a given 2-hour period in the same river?

Solution:

We have:

λ = 3/h

t = 2 h

K = 4

then:

$$\mathrm{P(X(t){=}k)} = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

$$\mathrm{P(X(t){=}k)} = \frac{(3 * 2)^4 e^{-3*2}}{4!}$$

P(X(2) = 4) = 0.134

1.3.2. The Bernoulli Process

Bernoulli process:

- named after Jacob Bernoulli
- a finite or infinite sequence of binary random variables
- a discrete-time stochastic process that takes only two values, canonically 0 and 1
- the component Bernoulli variables Xi are identically distributed and independent.
- a Bernoulli process is a repeated coin flipping, possibly with an unfair coin (but with consistent unfairness).
- every variable Xi in the sequence is associated with a Bernoulli trial or experiment. They all have the same Bernoulli distribution.

Jacob Bernoulli
Swiss mathematician
(1655 – 1705)

Definition:

- a Bernoulli process is a finite or infinite sequence of independent random variables X1, X2, X3, ..., such that:
  - for each i, the value of Xi is either 0 or 1
  - for all values of i, the probability p that Xi = 1 is the same

- a Bernoulli process is a sequence of independent identically distributed Bernoulli trials.

Binomial Distribution

- the discrete probability distribution of the number of successes in a sequence of n independent experiments
- each asking a yes–no question
- and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability q = 1 – p).
- with parameters n and p
- a single success/failure experiment is also called a Bernoulli trial or Bernoulli experiment
- a sequence of outcomes is called a Bernoulli process
- for a single trial, i.e., n = 1, the binomial distribution is a Bernoulli distribution
- the binomial distribution is the basis for the popular binomial test of statistical significance

Usage:

- frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N.
- If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one.

Binomial Experiments:

- number of successes in a specific number of trials.
- may be imagined as the probability distribution of a number of heads that appear on a coin flip in a specific experiment comprising of a fixed number of coin flips.

Binomial Distribution (Cont.)

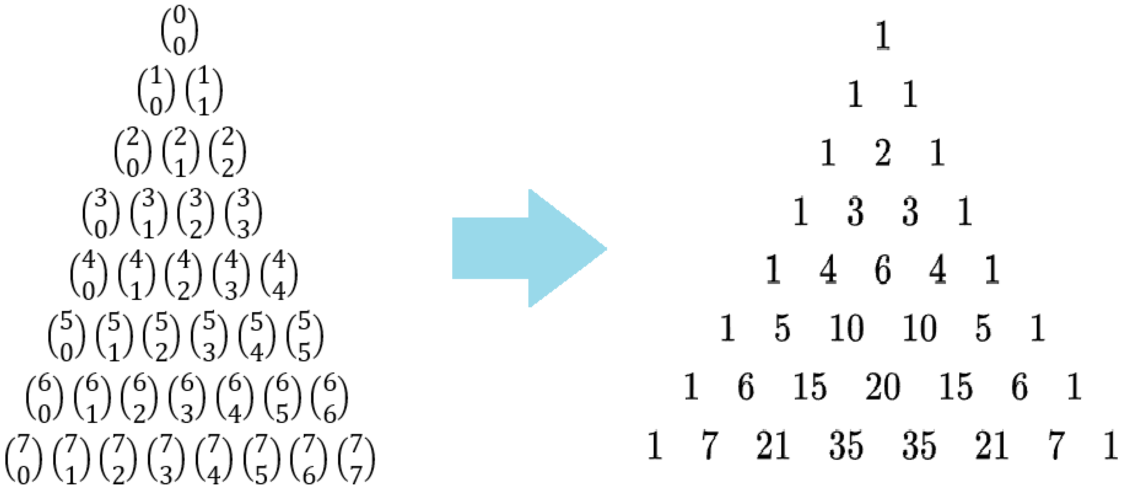
Probability Density Function:

$$\text{f(k, n, p)=P(k, n, p)=P(X=k)=}\binom{n}{k} p^k \, (1-p)^{n-k}$$

for k = 0, 1, 2, ..., n, where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

The coefficients of the binomial distribution can also be obtained from the Pascal Triangle



$$\begin{array}{c}
\binom{0}{0} \\
\binom{1}{0}\binom{1}{1} \\
\binom{2}{0}\binom{2}{1}\binom{2}{2} \\
\binom{3}{0}\binom{3}{1}\binom{3}{2}\binom{3}{3} \\
\binom{4}{0}\binom{4}{1}\binom{4}{2}\binom{4}{3}\binom{4}{4} \\
\binom{5}{0}\binom{5}{1}\binom{5}{2}\binom{5}{3}\binom{5}{4}\binom{5}{5} \\
\binom{6}{0}\binom{6}{1}\binom{6}{2}\binom{6}{3}\binom{6}{4}\binom{6}{5}\binom{6}{6} \\
\binom{7}{0}\binom{7}{1}\binom{7}{2}\binom{7}{3}\binom{7}{4}\binom{7}{5}\binom{7}{6}\binom{7}{7}
\end{array}$$

$$\begin{array}{c}
1 \\
1 \quad 1 \\
1 \quad 2 \quad 1 \\
1 \quad 3 \quad 3 \quad 1 \\
1 \quad 4 \quad 6 \quad 4 \quad 1 \\
1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\
1 \quad 6 \quad 15 \quad 20 \quad 15 \quad 6 \quad 1 \\
1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1
\end{array}$$

Binomial Distribution: Example 1

A coin is tossed 5 times. What is the probability of getting exactly 2 heads?

Solution:

n = 5

k = 2

Set head = 1, tail = 0

$$P(2, 5, 0.5) = \binom{5}{2} (0.5)^2 \, (1 - 0.5)^3$$

$$= \frac{5!}{2!\,(5-2)!} (0.5)^2 \, (1 - 0.5)^3$$

$$= 10 \, (0.5)^2 \, (0.5)^3 = 10 \, (0.5)^5$$

$$= 0.3125$$

Possibility #	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
1	1	1	0	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	0	1
5	0	1	1	0	0
6	0	1	0	1	0
7	0	1	0	0	1
8	0	0	1	1	0
9	0	0	1	0	1
10	0	0	0	1	1

10

$$\begin{array}{c}
\binom{0}{0} \\
\binom{1}{0} \; \binom{1}{1} \\
\binom{2}{0} \; \binom{2}{1} \; \binom{2}{2} \\
\binom{3}{0} \; \binom{3}{1} \; \binom{3}{2} \; \binom{3}{3} \\
\binom{4}{0} \; \binom{4}{1} \; \binom{4}{2} \; \binom{4}{3} \; \binom{4}{4} \\
\binom{5}{0} \; \binom{5}{1} \; \boxed{\binom{5}{2}} \; \binom{5}{3} \; \binom{5}{4} \; \binom{5}{5} \\
\binom{6}{0} \; \binom{6}{1} \; \binom{6}{2} \; \binom{6}{3} \; \binom{6}{4} \; \binom{6}{5} \; \binom{6}{6} \\
\binom{7}{0} \; \binom{7}{1} \; \binom{7}{2} \; \binom{7}{3} \; \binom{7}{4} \; \binom{7}{5} \; \binom{7}{6} \; \binom{7}{7}
\end{array}$$

Binomial Distribution: Example 2

A coin is tossed 10 times. What is the probability of getting exactly 6 heads?

Solution:

n = 10
k = 6

Use this formula:

$$\mathrm{P}(6, 10, 0.5) = \binom{10}{6} (0.5)^6 \, (1 - 0.5)^4$$

$$\mathrm{P}(6, 10, 0.5) = \frac{10!}{6!\,(10-6)!} (0.5)^6 \, (1 - 0.5)^4$$

P(6, 10, 0.5) = 210 $(0.5)^6$ $(0.5)^4$ = 0.2051

1.4. Continuous Random Processes

1.4.1. Preliminaries

Continuous Random Variable

- takes an infinite number of possible values. Continuous random variables are usually measurements.
- not defined at specific values. Instead, it is defined over an interval of values and is represented by the area under a curve.
- the probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.

A particular example:

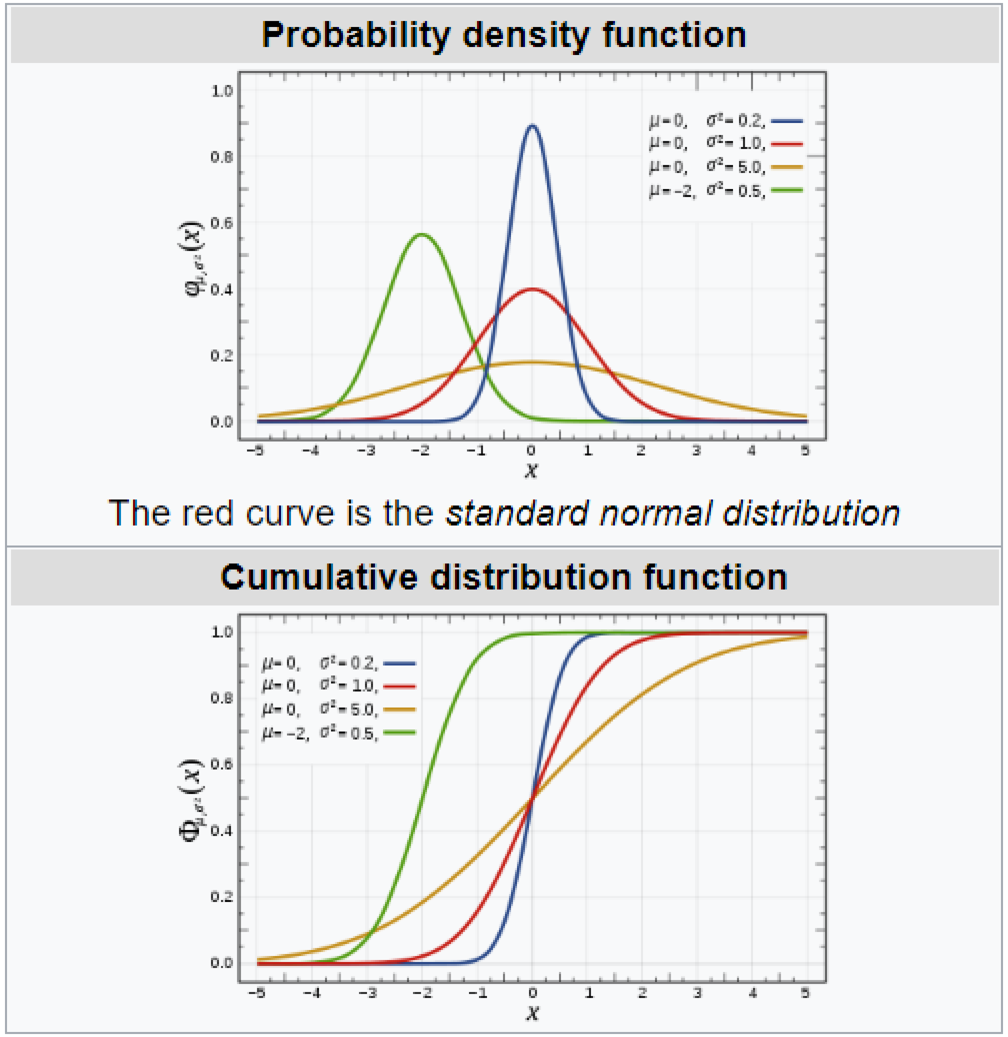
In a study of the ecology of a lake, with X being depth measurements at randomly chosen locations:

- X is a continuous random variable.
- range for X is the minimum depth possible to the maximum depth possible

1.4.2. The Normal Distribution

What Is a Normal Distribution?

- also known as the Gaussian distribution
- a probability distribution that is symmetric about the mean
- shows that data near the mean are more frequent in occurrence than data far from the mean
- in graphical form, the normal distribution appears as a "bell curve“



Probability density function

The red curve is the *standard normal distribution*

Cumulative distribution function

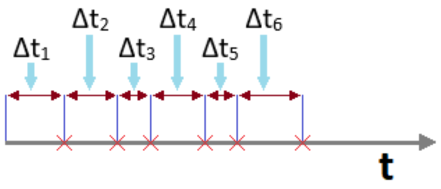
Density distribution function:

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

1.4.3. Other Types of Distributions

1.4.3.1. The Exponential Distribution

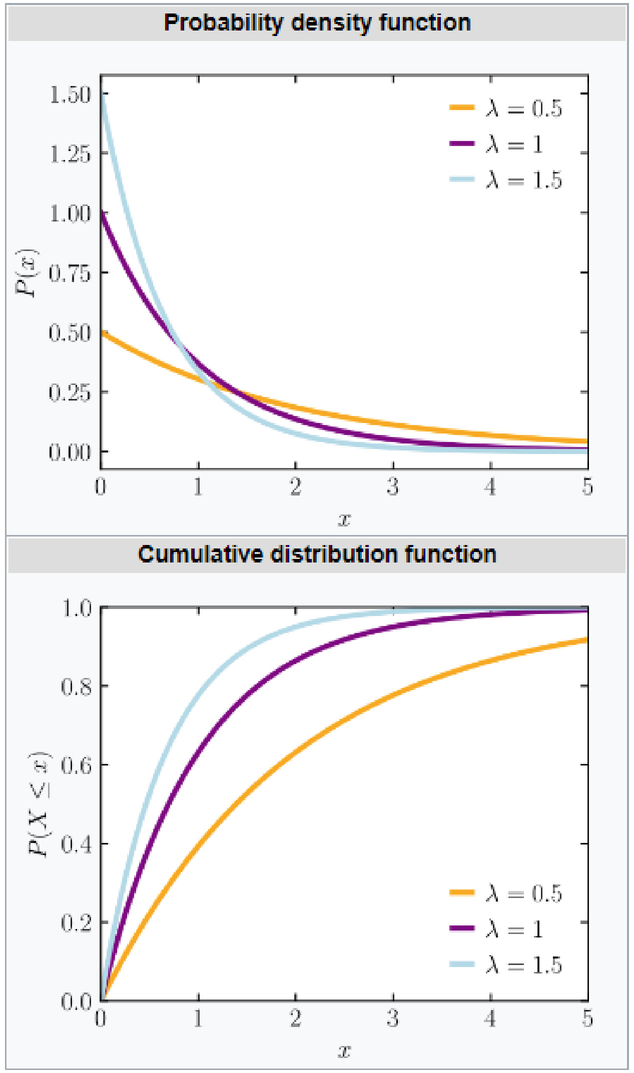
- related to the Poisson distribution.
- where the Poisson distribution describes the number of events per unit time, the exponential distribution describes the waiting time between events.
- it takes the same parameter as the Poisson distribution: the event rate.



Times between events:
T = { Δt_1, Δt_2, Δt_3, …}

$$\mathrm{P(next\ event\ happens\ at\ t)} = \mathrm{P}(\mathrm{T} = \mathrm{t}, \lambda)$$

$$= \begin{cases} \lambda\, e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$


Probability density function

Cumulative distribution function

Density distribution function:

$$f(x, \lambda) = \begin{cases} \lambda\, e^{-\lambda x} & \mathrm{x} \geq 0 \\ 0 & \mathrm{x} < 0 \end{cases}$$

1.4.3.2. The Gamma Distribution

- a variation on the exponential distribution.
- rather than describing the time between events, it describes the time to wait for a fixed number of events.
- two parameters: the lambda parameter of the exponential distribution, plus a k parameter for the number of events to wait for.

Example:

- waiting time for launching an event based on the number of visitors

Density distribution function:

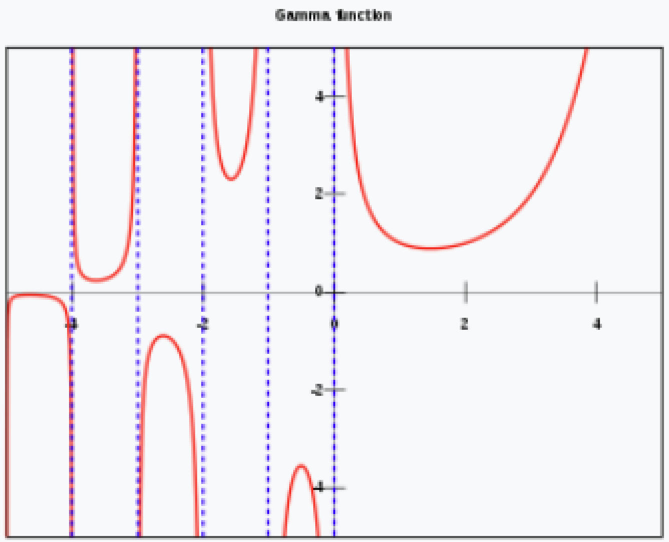
- for integers:

$$\Gamma(n) = (n-1)!$$

- for complex numbers with a positive real part:

$$\Gamma(z) = \int_0^{\infty} t^{z-1}\, e^{-t}\, dt$$

Gamma



The gamma function along part of the real axis

1.4.3.3. The Weibull Distribution

- a variation of the waiting time problem
- describes a waiting time for one event, if that event becomes more or less likely with time.

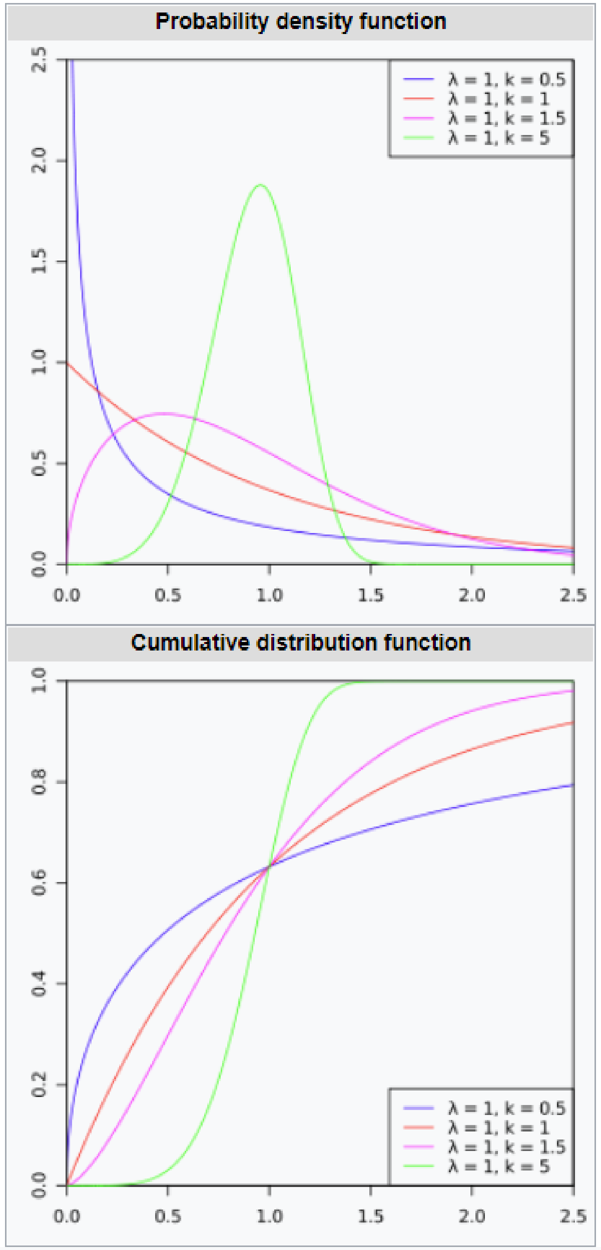
Example:

the life-time of a computer:

- the longer you have your computer, the more likely it becomes that it will break.
- Thus, the probability of failure, or the rate, is not constant.

Density distribution function:

$$f(x,\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



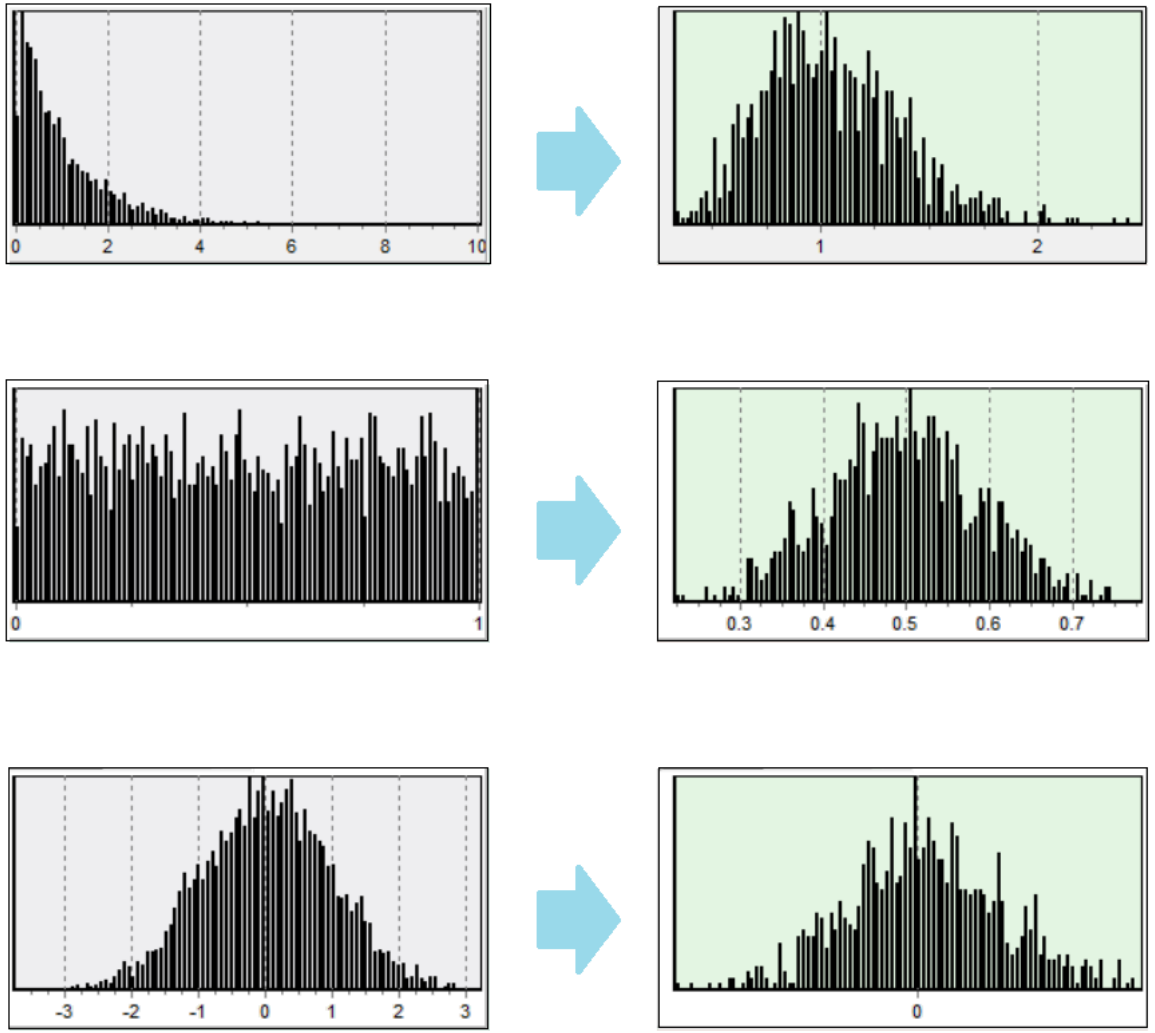
1.5. Central Limit Theorem

Central Limit Theorem:

- states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then:
  - the distribution of the sample means will be approximately normally distributed.
  - the average of the sample means approaches the theoretical mean.
  - **standard deviation** of the distribution of the sample means ($\boldsymbol{\sigma_s}$) approaches a fraction of the standard deviation of the population ($\boldsymbol{\sigma_p}$) (divided by the square root of the sample size).

Population Distribution

Sampling Distribution (e.g. sample means)



(view demo)

Implications:

- The mean of the sampling distribution is the mean of the population.

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution is the standard deviation of the population divided by the square root of the sample size.

$$\mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- We can describe the sampling distribution of the mean using this notation:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

1.6. Histograms

Histogram:

- is the most commonly used graph to show frequency distributions.
- it looks very much like a bar chart, but there are important differences between them.
- when the bin size is sufficiently small and the sample size is big, the histograms represent the probability density function of a random variable.

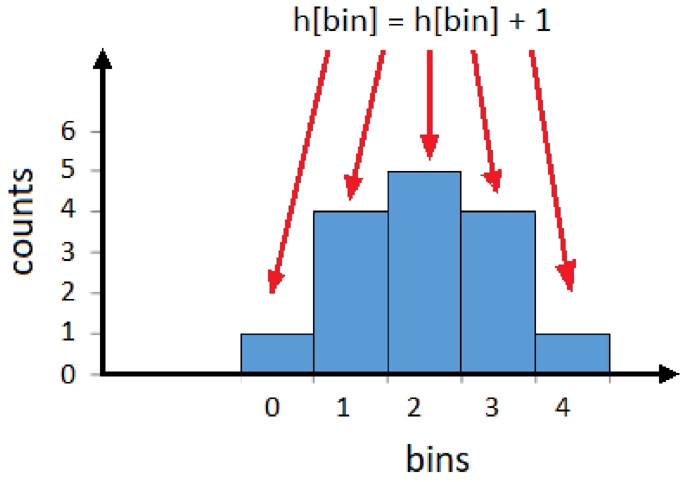
Use a histogram when:

- the data are numerical
- you want to see the shape of the data's distribution, especially when determining whether the output of a process is distributed approximately normally
- analyzing whether a process can meet the customer's requirements
- analyzing what the output from a supplier's process looks like
- seeing whether a process change has occurred from one time period to another
- determining whether the outputs of two or more processes are different
- you wish to communicate the distribution of data quickly and easily to others

How to Populate a Histogram

Steps:

1. Find x_{min} and x_{max} of data
2. Decide the number of bins n_{bins}
3. Compute binsize = $(x_{max} - x_{min})/n_{bins}$
4. Initialize a histogram h as an integer array of size n_{bins}. Set all elements to zero: h[i] = 0 for all i.
5. For each data point x[i], do the following:
   a. compute the index: index = $(x[i] - x_{min})$ div binsize
   b. add the following exception when $x[i] = x_{max}$: index = $n_{bins} - 1$
   c. accumulate the corresponding bar: h[index] = H[index] + 1

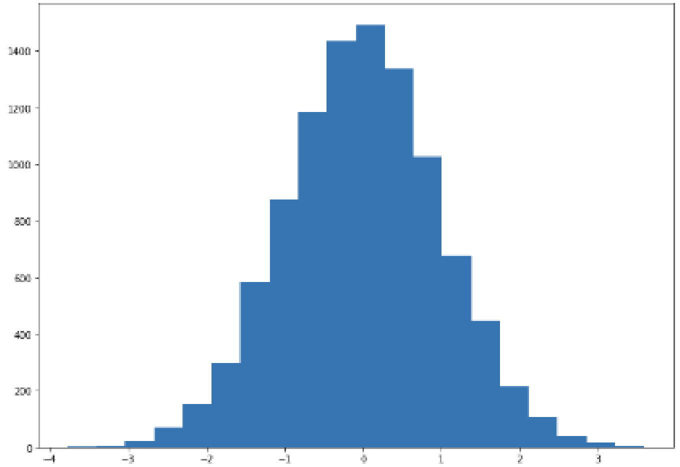


Now your histogram will have values for h[0], h[1], h[2], …, h[n_{bins} - 1]

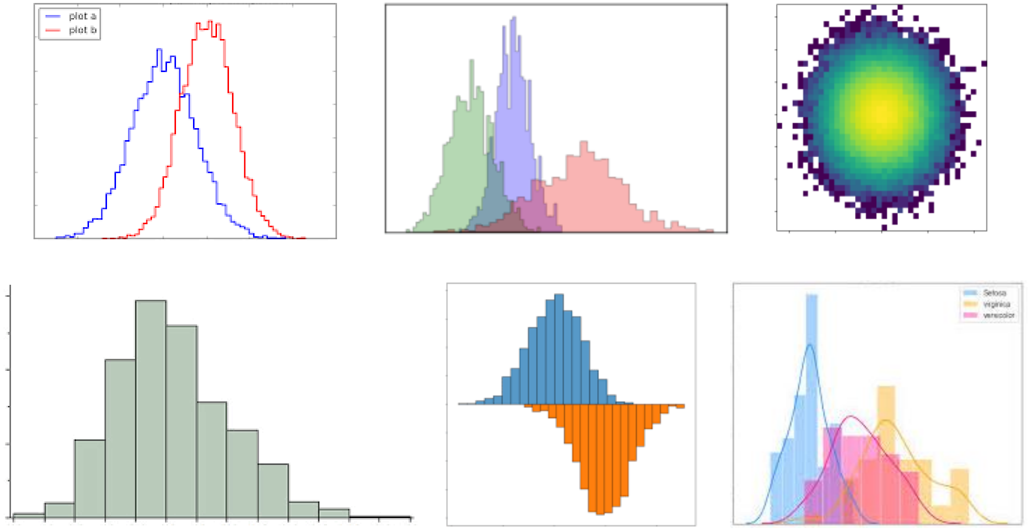
Note: Variations are possible (e.g., starting at zero, including negative indices, etc.)

How to Plot a Histogram in Python

```
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import colors
from matplotlib.ticker import PercentFormatter

# Creating dataset
np.random.seed(23685752)
N_points = 10000
n_bins = 20

# Creating distribution
x = np.random.randn(N_points)
y = .8 ** x + np.random.randn(10000) + 25

# Creating histogram
fig, axs = plt.subplots(1, 1,
                        figsize =(10, 7),
                        tight_layout = True)

axs.hist(x, bins = n_bins)

# Show plot
plt.show()
```



Histogram codes in Python can produce more visually appealing histograms such as the ones below:



2. Uncertainty

What is Uncertainty?

refers to situations involving imperfect or unknown information
applies to predictions of future events, to physical measurements that are already made, or to the unknown

measured with probability,
Measured with a set of possible states or outcomes where probabilities are assigned to each possible state or outcome

The Uncertainty Principle in Quantum Mechanics

- formulated by the German physicist Werner Heisenberg in 1927.
- states that we cannot know both the position and speed of a particle, such as a photon or electron, with perfect accuracy.
- the more we nail down the particle's position, the less we know about its speed and vice versa.

In general:

- in the quantum world it is impossible to simultaneously know two quantities, such as a particle's location and its momentum, with complete accuracy.
- the more you know about one, the less you can know about the other.
- has been popularized in many ways.

3. Information Theory

What Is Information Theory?

- mathematical treatment of the concepts, parameters and rules governing the transmission of messages
- information theory is the science of quantifying information for communication.
- proposed and developed by Claude Shannon while working at the US telephone company Bell Labs.
- quantifying the amount of information in events, random variables, etc.: fundamental concept in information theory
- requires knowledge of probabilities, therefore the relationship of information theory to probability.

Shannon’s Entropy

- used to quantity information
- intuition behind it: idea of measuring how much surprise there is in an event
  - rare events (low probability) → more surprising
  - common events (high probability) → less surprising
- therefore, rare events have more information than common ones, i.e.:
  - low probability event: high Information (surprising).
  - high probability event: low Information (unsurprising).
- rare events are more uncertain or more surprising → require more information to represent them than common events.

How to Calculate the Information of an Event

- we can calculate the amount of information there is in an event using the probability of that event.
- called “Shannon information”
- calculated for a discrete event x as follows:

information(x) = -log(p(x))

where:

log() is the base-2 logarithm

p(x) is the probability of the event x

- base-2 logarithm means:
  - the units of the information measure is in bits (binary digits).
  - the number of bits required to represent the event

- calculation of information is often written as h():

h(x) = -log(p(x))

How to Calculate the Entropy of a Random Process

- the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes.
- given a discrete random variable X, with p(x_i) being the probabilities of its individual outcomes i (i = 1, 2, …, n), its entropy can be calculated as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Example: How to Calculate the Entropy of a Random Process

- suppose we have a random process represented by a sequence of numbers: 1 0 3 5 8 3 0 7 0 1
- each distinct character has a different probability associated with it occurring:

  p(1)=2/10p(1)=2/10

  p(0)=3/10p(0)=3/10

  p(3)=2/10p(3)=2/10

  p(5)=1/10p(5)=1/10

  p(8)=1/10p(8)=1/10

  p(7)=1/10p(7)=1/10

- **the Shannon entropy of this process is given by:**

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$$

H(X) = - p(1) * log2(1/p(1)) - p(0) * log2(1/p(0)) - p(3) * log2(1/p(3))
- p(5) * log2(1/p(5)) - p(8) * log2(1/p(8)) - p(7) * log2(1/p(7))
= - 0.2 * log2(1/0.2) - 0.3 * log2(1/0.3) - 0.2 * log2(1/0.2) - 0.1 * log2(1/0.1) - 0.1 * log2(1/0.1) - 0.1 * log2(1/0.1)
= 2.44644

Tossing a Coin: How fair is it?

- can be modelled as a Bernoulli process
- entropy is maximized if the coin is fair (heads and tails with equal probability)
- situation of maximum uncertainty → it is most difficult to predict the outcome of the next toss

$$\mathrm{H(X)} = -\sum_{i=1}^{n} p(x_i)\log p(x_i) = -(0.5)log_2(0.5) - (0.5)log_2(0.5) = -2(0.5)(-1) = 1$$

If we know the coin is not fair:

- heads or tails with probabilities p and q, where $p \neq q$
- there is less uncertainty
- reduced uncertainty is quantified in a lower entropy
- for example, if p = 0.7, then:

$$\mathrm{H(Xunfair)} = -(0.7)\log_2(0.7) - (0.3)\log_2(0.3) = 0.8816 < 1$$

Uniform probability → maximum uncertainty → maximum entropy.

The extreme case:

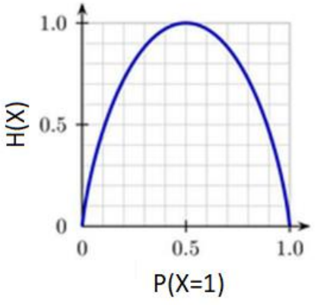
double-headed coin that never comes up tails (p=1, q=0) → there is no uncertainty → entropy is zero

$$\mathrm{H}(X_{TotallyUnfair}) = -(1.0)\log_2(1.0) - 0.0 = 0$$

No uncertainty (outcome is always certain)

Graph of entropy vs. P(X=1)

for a Bernoulli trial X = {0, 1}

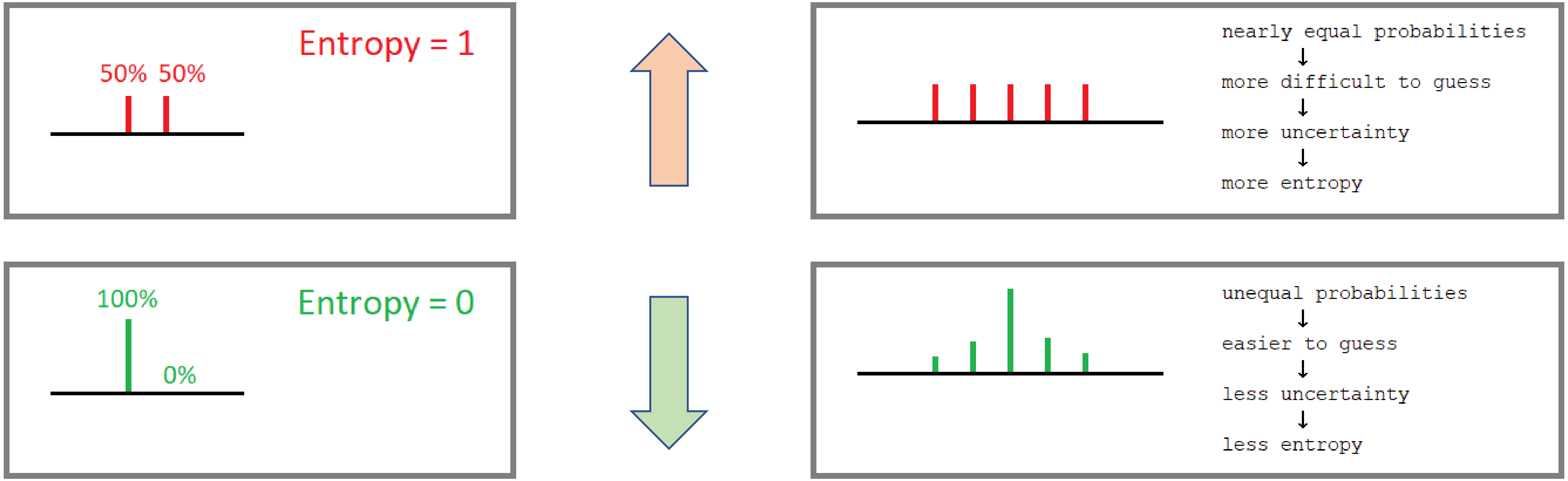


The highest entropy H(X) = 1 = $-\log_2(1/2)$

occurs when P(X=1)=0.5

Concluding Remarks

- entropy is higher when there is more uncertainty in the outcome of a random variable
- entropy is higher when the outcome is most difficult to predict
- a fair game should have a maximum entropy (one where there are only two equally probable outcomes)
- entropy =1 → total uncertainty (outcome is highly difficult to predict)
- entropy between 0 and 1 → some uncertainty (outcome is somehow uncertain)
- entropy = 0 → no uncertainty (outcome is always certain)



nearly equal probabilities
↓
more difficult to guess
↓
more uncertainty
↓
more entropy

unequal probabilities
↓
easier to guess
↓
less uncertainty
↓
less entropy

End of Lecture