Lecture 2: Probability Distribution

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 2, 2023

Today

1. Probability distribution
   - Why probability distributions?
   - Terminology
2. Summary

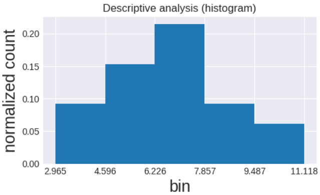
Learning outcome

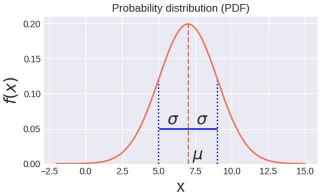
- Be able to explain the following concepts: experiment, event, random variable, probability distribution
- Given the PDF/PMF, be able to describe the probability distribution of a continuous/discrete random variable
- Understand Gaussian distribution and Bernoulli distribution: 1) PDF/PMF; 2) what are the parameters? 3) what happens to the shape of the distribution if we change the parameters?

Why probability distributions?

Histogram vs probability distribution

You need to get a good overview (i.e. **distribution**) of your **1000** ducks' weights without weighing all of them, because, well, data collection is expensive. You weighed **20** ducks and you plotted the histogram of the weights. Your best friend Jack looked at the histogram and suggested that you should use a **Gaussian distribution** to make a better **estimation** of the **distribution**.





Histogram vs probability distribution

- Question 1: Why can't I just use **descriptive analysis**, like the **histogram**, to describe the data distribution? Why should I use **probability distributions**?

  To address this question, let's describe the data distribution using a **histogram** and a **Gaussian distribution** to see the difference.

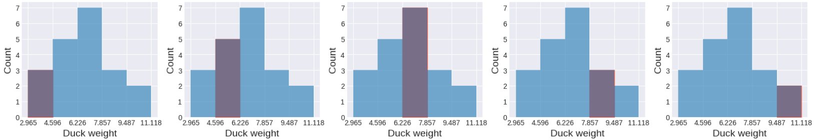
Histogram vs probability distribution

Here are the weights of the 20 ducks in kg

duck id	1	2	3	4	$\cdots$	19	20
weight	6.98	5.43	2.97	7.07	$\cdots$	4.63	7.27

Let's try to describe these ducks using a histogram with 5 bins.

- Divide the range of the data into 5 bins
- There are 3 ducks within the first bin $[2.965, 4.596]$; there are 5 ducks within second bin $[4.596, 6.226]$, etc.



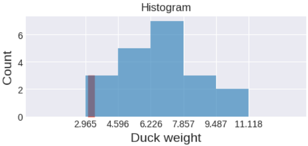
- Now you want to use the histogram to **describe the distribution** of 1000 ducks
- **This** allows you to answer questions such as "what is the chance of a duck weighing between 2.965 kg and 4.596 kg?" $\frac{3}{20}$
  How about between 3.1 kg and 3.4 kg?

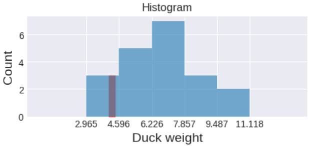
Histogram vs probability distribution

- Resolution: how many bins we use to describe one kilogram (the number of bins per kilogram)

$$\frac{\text{number of bins}}{\textit{range}} = \frac{\text{number of bins}}{\max(\textit{weights}) - \min(\textit{weights})} = \frac{5}{11.118 - 2.965} = 0.61$$

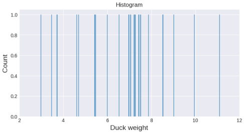
- How do we describe the distribution of a duck?
  - The chance of a duck weighing between 3.1 and 3.4: $(3.4 - 3.1) \times \textit{resolution} \times \frac{3}{20} = 0.028$
  - The chance of a duck weighing between 4.1 and 4.4: $(4.4 - 4.1) \times \textit{resolution} \times \frac{3}{20} = 0.028$





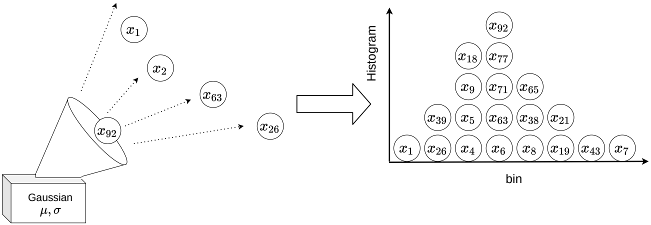
$$\text{“chance”} = \frac{\text{the area of the red rectangle}}{\text{the total area of the blue rectangles}}$$

  - We have “low resolution” due to quantization - you don't know what's going on within each bin
  - And if we increase the number of bins? 5 -> 200



Histogram vs probability distribution

- Descriptive analysis (e.g. histogram) is limited to the existing sample - it is not designed for describing **unseen data**
- Now let's try to use a Gaussian distribution to describe the data
  - First, we **assume** that data is **generated** from a Gaussian distribution (e.g. **np.random.normal** in Python)
  - We can generate as many data points as we like
  - The histogram of these data points will be "bell-shaped"

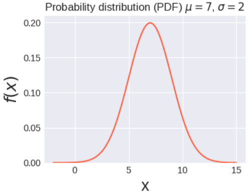


Histogram vs probability distribution

- A Gaussian distribution is described by a **function** that **looks similar** to this "bell-shaped" histogram

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

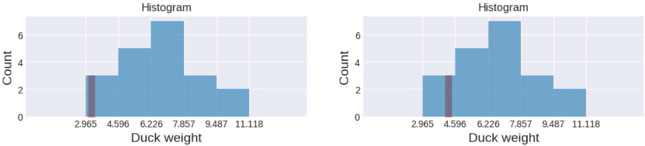
- This **function** is sufficiently defined by two **parameters** μ and σ.
- The **shape** of the **function** (e.g. given μ=7 and σ=2):



- We will try to use this **function** instead of the histogram to describe the data.

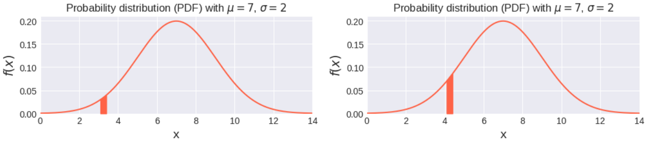
Histogram vs probability distribution

- Describe the distribution:
  - Histogram (using 0.61 bins to describe 1 kg):
    - The chance of $weight \in [3.1, 3.4]$: $(3.4 - 3.1) \times resolution \times \frac{3}{20} = 0.028$
    - The chance of $weight \in [4.1, 4.4]$: $(4.4 - 4.1) \times resolution \times \frac{3}{20} = 0.028$



$$\text{“chance”} = \frac{\text{the area of the red rectangle}}{\text{the total area of the blue rectangles}}$$

  - Gaussian distribution (using **infinite** bins to describe 1 kg):
    - The chance of $weight \in [3.1, 3.4]$: $\int_{3.1}^{3.4} f(t)dt = 0.010$
    - The chance of $weight \in [4.1, 4.4]$: $\int_{4.1}^{4.4} f(t)dt = 0.023$



$$\text{“chance”} = \frac{\text{the red area}}{\text{the total area under the curve}}$$

Histogram vs probability distribution

- Descriptive analysis: a *histogram* with M bins (e.g. $M = 5$)
- Gaussian distribution: a mathematical function $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- Comparison

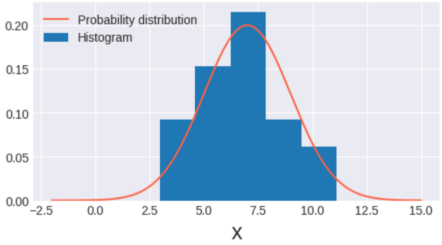
	Histogram		Gaussian distribution	
Representation	M values		mathematical function f	
Number of parameters	M	-	2 (μ and σ)	+
Resolution	$\frac{M}{\max(x)-\min(x)}$	-	infinity	+
Analytical properties	No	-	Yes	+
Assumptions	No	+	Yes	-
Can be directly computed from data	Yes	+	Parameters unknown	-

- For your use case, you want to **estimate** the **distribution** of your 1000 ducks without weighing all of them. It is hard to do that from the histogram. The histogram **describes** the data you have seen, but it is not designed for describing unseen data.
- Statistical modelling using probability distributions (e.g. a Gaussian distribution) can help you with that!
- This concludes the question why we use probability distributions instead of histograms to describe your ducks.

Disclaimer: we used a continuous distribution to illustrate the comparison between a histogram and a probability distribution. A discrete probability distribution differs from a continuous distribution.

Choosing a probability distribution

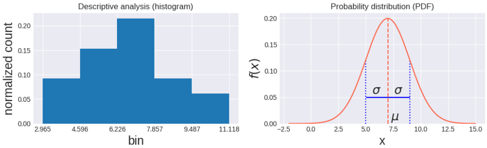
- Question 2: How do I know which probability distribution I should use to describe the data? How do I know that it should be a Gaussian distribution?
  - Short answer: if the probability distribution looks like the histogram, go for it!



  - Long answer will be given in lecture 3.

Parameter estimation and evaluation

- Question 3: Okay fine, let's say we describe the data with a Gaussian distribution. How do I know what the parameters μ and σ are?



  - This is done by **parameter estimation**. In lecture 3 & 4, we will talk about the **maximum likelihood estimation (MLE)** and the **maximum a posteriori estimation (MAP)**.

Terminology

Probability distribution

- **Experiment**: an action that leads to one outcome. For example, we weigh a duck and look at its weight. The outcome is weight = 2 kg.
- **Sample space**: the set of all possible outcomes from the experiment. The sample space of the previous example is any real value between 0 and ∞.
- **Event**: a subset of the sample space, for example, a duck weighs between 5kg and 6kg.
- **Probability distribution**: the probability of the occurrence of *any* event in the sample space, e.g. P(a duck weighs between a kg and b kg) for any $0 < a < b < \infty$ (not only for $a = 5$ and $b = 6$).
- **Random variable** X:
  - Heuristically, X assigns a numerical value to each outcome of the experiment:

$$X : \text{weight} \to \mathbb{R}$$

  - X follows some underlying probability distribution.
  - **Discrete random variable** and **continuous random variable**: depends on the sample space of the experiment; the underlying distributions are called **discrete distribution** and **continuous distribution**, respectively. For example, weights are continuous so X from this example is a continuous random variable.
- **Data** x: a value drawn from the **underlying distribution of** X.
  - We use a **capital letter** (e.g. X) to denote a random variable and the corresponding lower case letter (e.g. x) to denote the data generated from the underlying distribution of X.
  - Discrete random variable: categorical data or discrete numerical data
  - Continuous random variable: continuous numerical data

Probability distribution

A probability distribution describes the probabilities of occurrence of all possible events.

More precisely, a probability distribution is defined by a function f_X (also denoted as f if neglecting X does not cause confusion), where

- for discrete distribution, the **probability mass function (PMF)** is used, where

$$f_X(x_i) = \boxed{P(X = x_i)}$$

  where $0 \leq f_X(x_i) \leq 1$ for all x_i (discrete).

- for continuous distribution, the **probability density function (PDF)** is used, where

$$\boxed{P(a \leq X \leq b)} = \int_a^b f_X(x)dx, \ \forall a, b \in \mathbb{R}, a \leq b$$

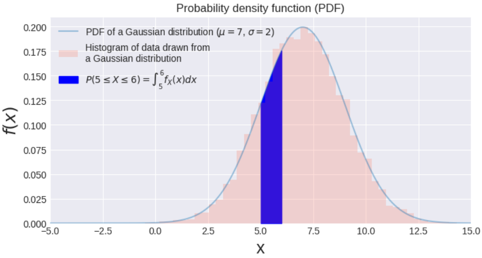
  where $f_X(x) \geq 0$ for all x (continuous).

$\boxed{P(\text{event})}$ is the probability of the **event** occurring.

Example: continuous random variables and PDF

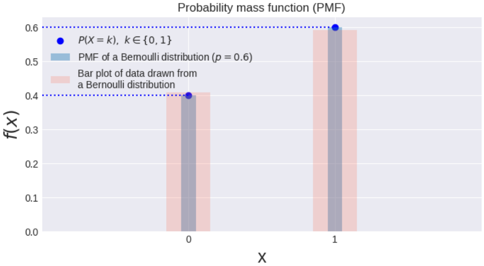
- **Experiment**: you weigh a duck and look at its weight
- **Sample space**: $0 < \textit{weight} < \infty$
- **Random variable** X : $\textit{weight} \rightarrow \mathbb{R}$
  - $X = x$ if the duck weighs x kg for $0 < x < \infty$
  - X follows a **Gaussian distribution** with parameters μ and σ; denoted as $X \sim \mathcal{N}(\mu, \sigma^2)$
- **PDF**: $f_X(x)$

$$P(a \leq X \leq b) = \underbrace{\int_a^b f_X(x)dx}_{\text{Integral = area under the PDF curve}} \qquad \forall a, b \in \mathbb{R}, a \leq b$$



Example: discrete random variables and PMF

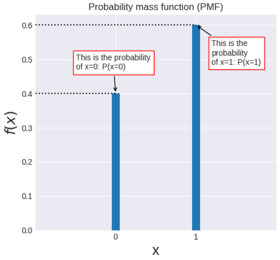
- **Experiment**: you measure the color of the duck.
- **Sample space**: the color can be either red or blue
- **Random variable** X : $color \to \mathbb{Z}$
  - $X = \begin{cases} 0, & \text{if a duck is red} \\ 1, & \text{if a duck is blue} \end{cases}$
  - X follows a **Bernoulli distribution** with parameter p; denoted as $X \sim Bernoulli(p)$
- **PMF**: $f_X(x_i) = P(X = x_i)$

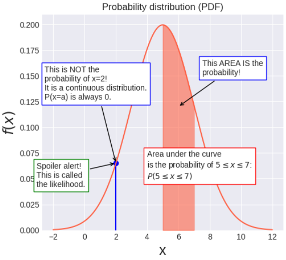


Probability distribution

Differences between PMF and PDF

- Discrete distribution $f_X(x_i) = P(X = x_i)$:
  - y-axis represents the probability itself
- Continuous distribution:
  - $P(a \leq X \leq b) = \int_a^b f_X(x)dx$: **y-axis $f(x)$** DOES NOT represent the probability itself.
  - For continuous distributions, **the probability at any given value is always 0**, i.e. $P(X = a) = P(a \leq X \leq a) = \int_a^a f_X(x)dx \equiv 0$. Example: what is the probability of a duck weighing exactly 4.32028374... kg?





So far:

- Data types, data containers, descriptive statistics (e.g. sample mean, sample variance, data quantile), visualization (e.g. histogram)
- Probability distributions, sample space, events, random variables, PMF, PDF, parameters

Not yet:

- How to choose a probability distribution for a given data set?

Next:

- Comparing two distributions using a Q-Q plot

Before next lecture:

- Quantile
- PMF and PDF

Stay strong (for your ducks)!