Lecture 3: Comparing distributions

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 6, 2023

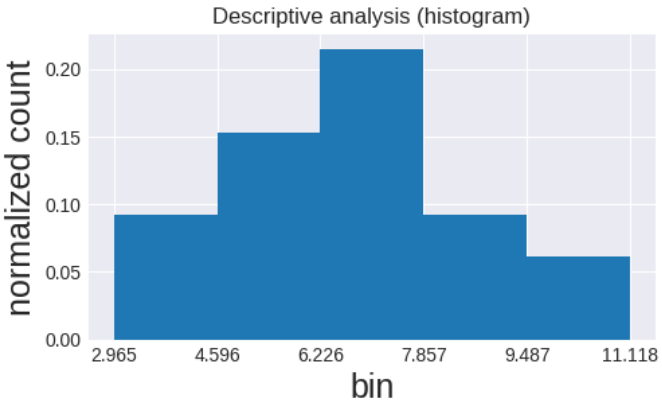
Today

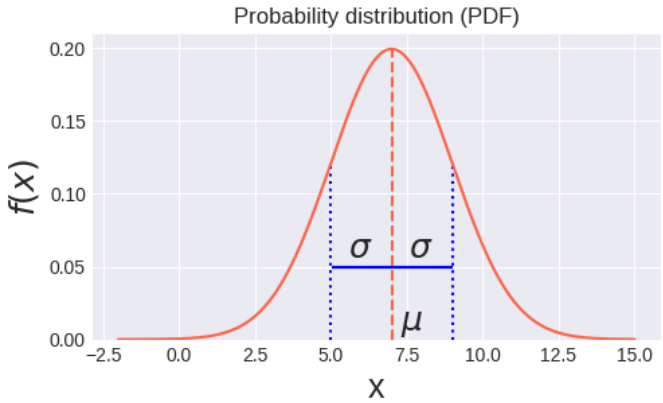
Learning outcome

- Be able to apply the learning routine to study a new probability distribution
- Be able to compute conditional probability
- Be able to explain the following terminology: Cumulative distribution function (CDF), Q-Q plot, one-sample/two-sample tests
- Be able to compute quantiles in Python for a given theoretical probability distribution
- Understand the relation between quantile and CDF
- Be able to construct a Q-Q plot

Recap: three questions from lecture 2

Jack suggested to use a Gaussian distribution to model your data.





- ✅ Question 1: Why should I use probability distributions instead of histograms?
- ? Question 2: How do you know if my data follows a Gaussian distribution?
- ? Question 3: How do I find the unknown parameters?

In today's lecture, we are going to address question 2.

Today

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Recall

- Discrete distribution is described by the probability mass function (PMF)
- Continuous distribution is described by the probability density function (PDF)

For each distribution, you need to know:

- its PMF or PDF: the equation and the shape
- its parameters
- its applications
- how to estimate the parameters (next lecture)

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Bernoulli distribution

Bernoulli distribution

In your town everybody has ducks (of course). Ducks in this town ONLY have TWO colors: blue and red. What is the probability distribution we use to describe the duck colors in your town?

- Let X be a discrete random variable $X = \begin{cases} 0 & \text{a duck is red} \\ 1 & \text{a duck is blue} \end{cases}$
- Given p the probability of a duck being blue, we can express the probability distribution as follows:

  $$P(\text{a duck is red}) = P(X = 0) = 1 - p$$
  $$P(\text{a duck is blue}) = P(X = 1) = p$$

- What is the PMF?
  Merge these two equations:

$$P(X = k) = f_X(k) \equiv f_X(k \mid p) = pk + (1 - p)(1 - k), \ k \in \{0, 1\}, p \in [0, 1]$$

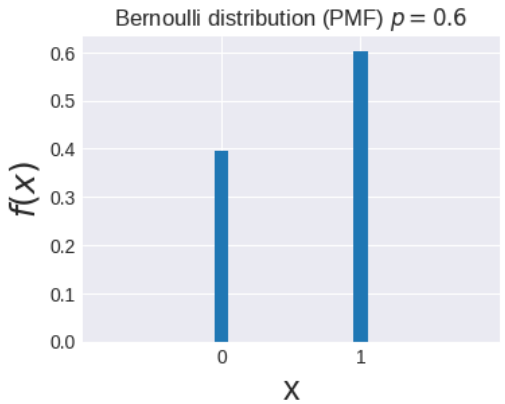
Note: here we use a $\mid$ to indicate that the parameter p is given.

Bernoulli distribution

- Discrete distribution
- Applies to nominal data with 2 categories
- PMF:
  - Equation

$$f_X(k \mid p) = pk + (1-p)(1-k), k \in \{0,1\}, p \in [0,1]$$

  - Shape



- Parameters: p

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Categorical distribution

Categorical distribution

In Jack's town, ducks have FOUR colors: blue, red, green and gray. What is the probability distribution of duck colors in Jack's town?

- Given p_1 the probability of a duck being blue, p_2 the probability of a duck being red, p_3 the probability of a duck being green and p_4 the probability of a duck being gray. Note that $p_1 + p_2 + p_3 + p_4 = 1$.
- Let X be a discrete random variable $X = \begin{cases} 1 & \text{a duck is blue} \\ 2 & \text{a duck is red} \\ 3 & \text{a duck is green} \\ 4 & \text{a duck is gray} \end{cases}$
- Now we can express the probability distribution as follows:

  $P(\text{a duck is blue}) = P(X = 1) = p_1$
  $P(\text{a duck is red}) = P(X = 2) = p_2$
  $P(\text{a duck is green}) = P(X = 3) = p_3$
  $P(\text{a duck is gray}) = P(X = 4) = p_4$
- What is the PMF?

$$P(X = k) = f_X(k) \equiv f_X(k \mid p_1, p_2, p_3, p_4) = p_k, \quad \sum_{i=1}^{4} p_i = 1, p_i \geq 0, \quad k \in \{1, \cdots, 4\}$$

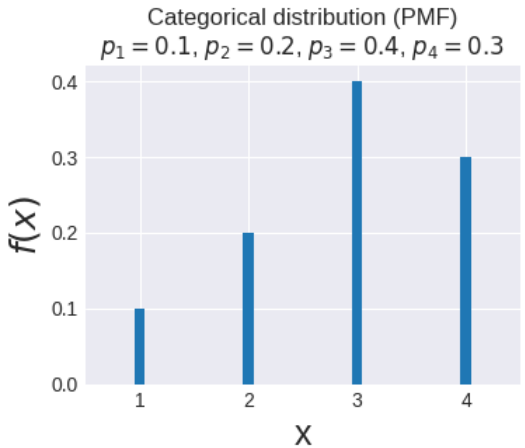
Note: categorical distribution is also called multinoulli distribution. It is a generalization of the Bernoulli distribution.

Categorical distribution

- Discrete distribution
- Applies to nominal data with $n > 0$ categories
- PMF:
  - Equation

$$f_X(k \mid p_1, p_2, \cdots, p_n) = p_k, \quad \sum_{i=1}^{n} p_i = 1, p_i \geq 0, \quad k \in \{1, \cdots, n\}$$

  - Shape



- Parameters: p_k, $k \in \{1, \cdots, n\}$ for given n; $n - 1$ parameters ($\sum_{i=1}^{n} p_i = 1$).

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Discrete uniform

Discrete uniform distribution

Meanwhile, back to your town, a team of scientists crunched some numbers and they stated that the number of ducks that each person has follows a uniform distribution between 1 and 1000. What does that mean?

- Given integers a and b with $a \leq b$ (here we have $a = 1$ and $b = 1000$)
- Let X be a discrete random variable: $X = k$ if a person has k ducks
- We can express the probability distribution as follows:

$$P(\text{a person has k ducks}) = P(X = k) = \frac{1}{b-a+1}$$

- What is the PMF of a discrete uniform distribution?

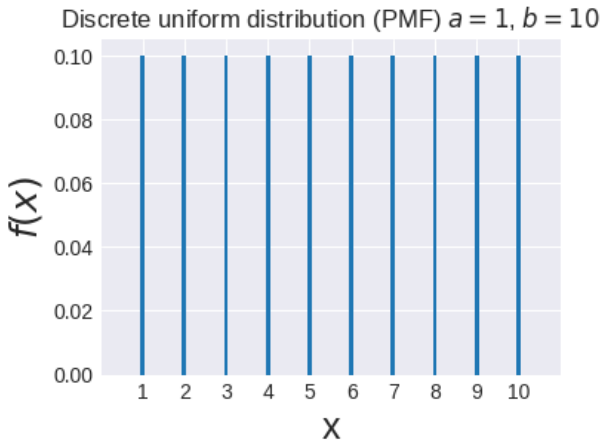
$$P(X = k) = f_X(k) \equiv f_X(k \mid a, b) = \frac{1}{b - a + 1}$$

Discrete uniform distribution

- Discrete distribution
- Applies to discrete numerical data
- PMF:
  - Equation

$$f_X(k \mid a, b) = \frac{1}{b - a + 1}, \quad a \leq b, \quad a, b \text{ integers}$$

  - Shape



- Parameters: integers a, b

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Gaussian distribution

Gaussian (normal) distribution

Still at your town, a scientist told you that the weights of your ducks follow a Gaussian distribution with mean $\mu = 7$ and standard deviation $\sigma = 2$.
What does that mean?

- Let X be a continuous random variable: $X = x$ if a duck weighs x kg
- Let $f_X(x)$ be the PDF, the probability distribution can be expressed as (10 secs):

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \text{ for all } a < b$$

- What is the PDF of a Gaussian distribution with $\mu = 7$ and $\sigma = 2$?

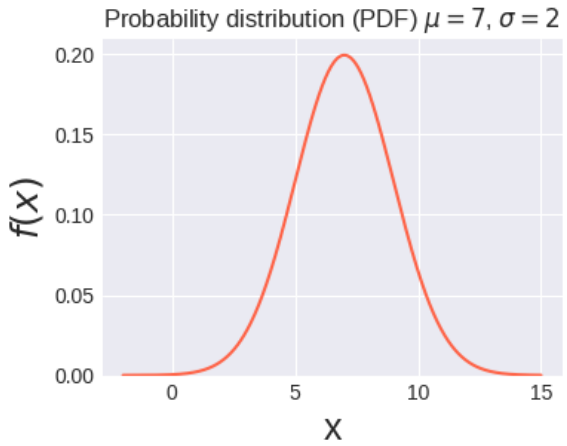
$$f_X(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}\left(\frac{x-7}{2}\right)^2}$$

Gaussian (normal) distribution

- Continuous distribution
- Applies to continuous numerical data
- PDF:
  - Equation

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \ \mu \in \mathbb{R}, \ \sigma \in \mathbb{R}^+$$

  - Shape



- Parameters: μ, σ

Probability distributions

Probability distribution	Continuous/discrete	Apply to data type
Bernoulli distribution	Discrete	Categorical (nominal)
Categorical distribution	Discrete	Categorical (nominal)
Discrete uniform	Discrete	Numerical (discrete)
Gaussian distribution	Continuous	Numerical (continuous)

Hooray!

An important note

These probability distributions DO NOT ONLY apply to duck related applications!

We are going to talk about more applications in the future (even though they won't be as important as ducks)

Today

Conditional probability

Conditional probability

Given events A and B,

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

The probability of event A given event B.

Conditional probability example

Example:

- **Experiment**: You ask your ducks to stand in a row again and look at their colors and sizes.
- **Sample space**: The color can be either red or blue; the size can be either slim or chonker.
- **Data**:

duck id	1	2	3	4	5	6
color	red	red	blue	blue	blue	red
size	chonker	slim	slim	chonker	chonker	slim

- **Event**:
  - A: a duck is blue
  - B: a duck is a chonker
- Estimate the conditional probability $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$ from this data
  - $P(A \text{ and } B)$: the probability that a duck is both blue and a chonker is (10 secs)

$$\frac{count(\text{blue and chonker})}{\text{total}} = \frac{2}{6}$$

  - $P(B)$: the probability that a duck is a chonker is (10 secs)

$$\frac{count(\text{chonker})}{\text{total}} = \frac{3}{6}$$

  - Conditional probability:

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)} = \frac{2/6}{3/6} = \frac{2}{3}$$

Conditional probability example (cont.)

An alternative way to estimate $P(A \mid B)$:

- Count the chonkers: 3 (event B)
- Count blue ducks within the chonker gang: 2 (event A)
- $P(A \mid B) = \frac{2}{3}$
  Is it the same as $P(B \mid A)$?

Side note: from this calculation, can you make a bold statement about the probability distribution of all the ducks in the world? No! It is only an estimation based on the data you got.

Conditional probability example (cont.)

As an exercise, let's define the random variables.

- **Experiment**: You ask your ducks to stand in a row again and look at their colors and sizes.
- **Sample space**: The color can be either red or blue; the size can be either slim or chonker.
- **Data**:

duck id	1	2	3	4	5	6
color	red	red	blue	blue	blue	red
size	chonker	slim	slim	chonker	chonker	slim

- **Event**:
  - A: a duck is blue
  - B: a duck is a chonker
- **Random variables**: X, Y

Are they continuous or discrete? (2 secs) Discrete

In practice, $X : \text{color} \to \mathbb{Z}$, $Y : \text{size} \to \mathbb{Z}$

$$X = \begin{cases} 0, & \text{duck is red} \\ 1, & \text{duck is blue} \end{cases} \quad \text{and} \quad Y = \begin{cases} 0, & \text{duck is slim} \\ 1, & \text{duck is a chonker} \end{cases}$$

Write the conditional probability in terms of the random variables X and Y (10 secs), i.e.

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = P(X = 1 \mid Y = 1) = \frac{P(X = 1 \text{ and } Y = 1)}{P(Y = 1)}$$

Independent events

Independent events

Two events A and B are independent if and only if

$$P(A \text{ and } B) := P(A \cap B) = P(A)P(B)$$

$$\iff P(A \mid B) = P(A),\ P(B \mid A) = P(B) \text{ (conditional probability)}$$

$$\iff log\left(P(A \text{ and } B) = \log\left(P(A \cap B)\right) = \log(P(A)) + \log(P(B))\right.$$

Bayes' rule

Bayes' rule

Given events A and B,

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

Just a heads-up - important concept in this course and machine learning in general

Cumulative distribution function (CDF)

Terminology alert

For a random variable X, the **cumulative distribution function (CDF)** F_X is defined as

$$F_X(x) = P(X \leq x)$$

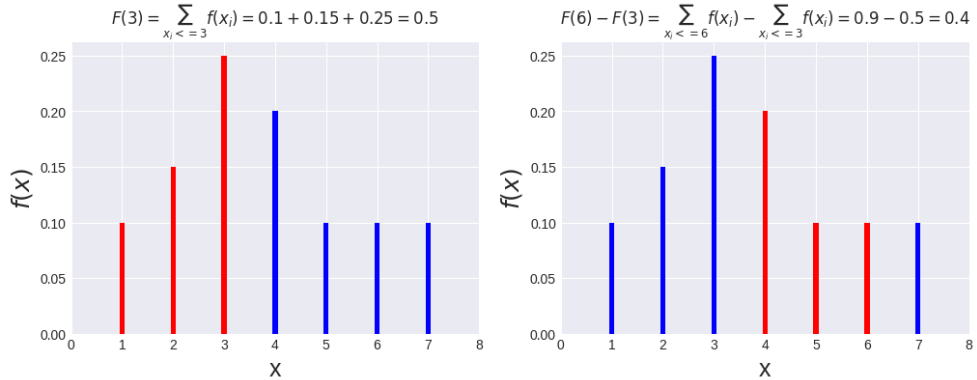
where X can be a discrete or a continuous random variable.

- X discrete random variable (categorical or discrete values):
  - **Definition**: given the PMF f_X,

$$F_X(\mathbf{x}) = P(X \leq \mathbf{x}) = \sum_{x_i \leq \mathbf{x}} P(X = x_i) = \sum_{x_i \leq \mathbf{x}} f_X(x_i)$$

    where x_i are all the values X can take.
  - Implication:

$$F_X(b) - F_X(a) = P(a < X \leq b) = \sum_{x_i \leq b} f_X(x_i) - \sum_{x_i \leq a} f_X(x_i)$$



Terminology alert

For a random variable X, the **cumulative distribution function (CDF)** F_X is defined as

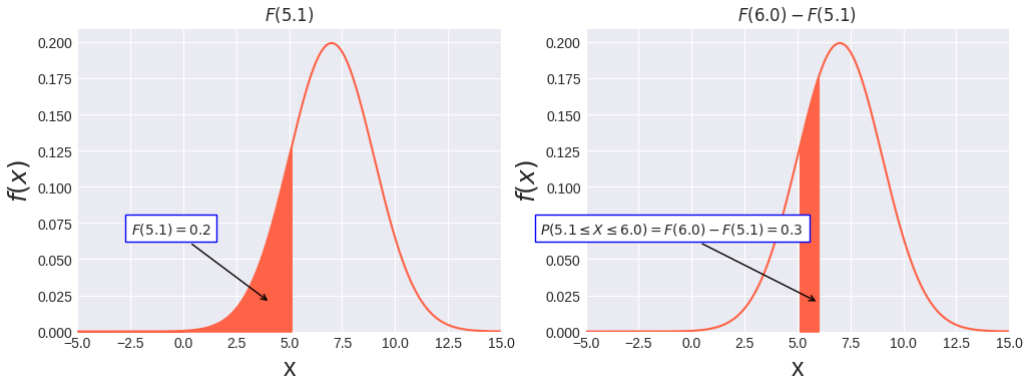
$$F_X(x) = P(X \leq x)$$

- X continuous random variable (continuous values):
  - **Definition**: given the PDF f_X,

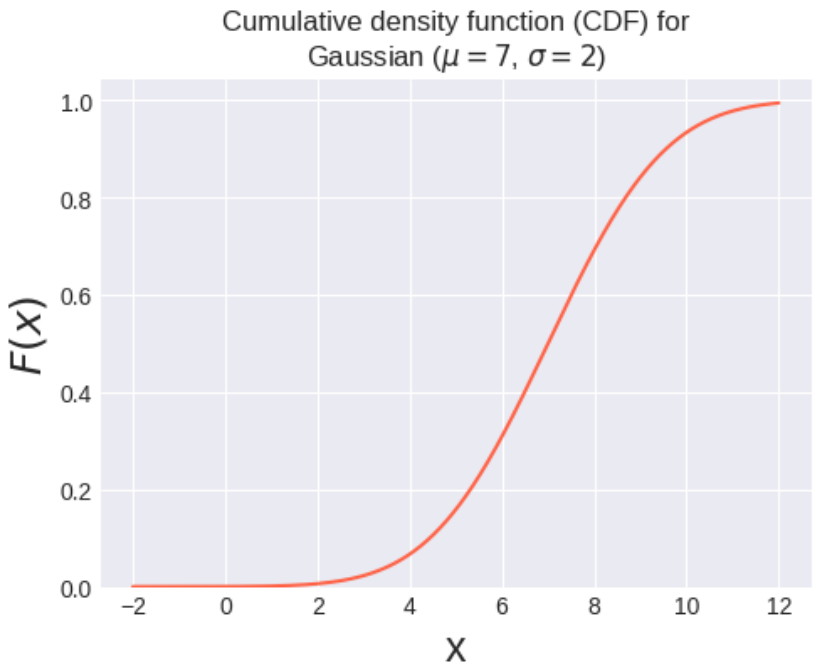
$$F_X(\mathbf{x}) = P(X \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_X(t)dt$$

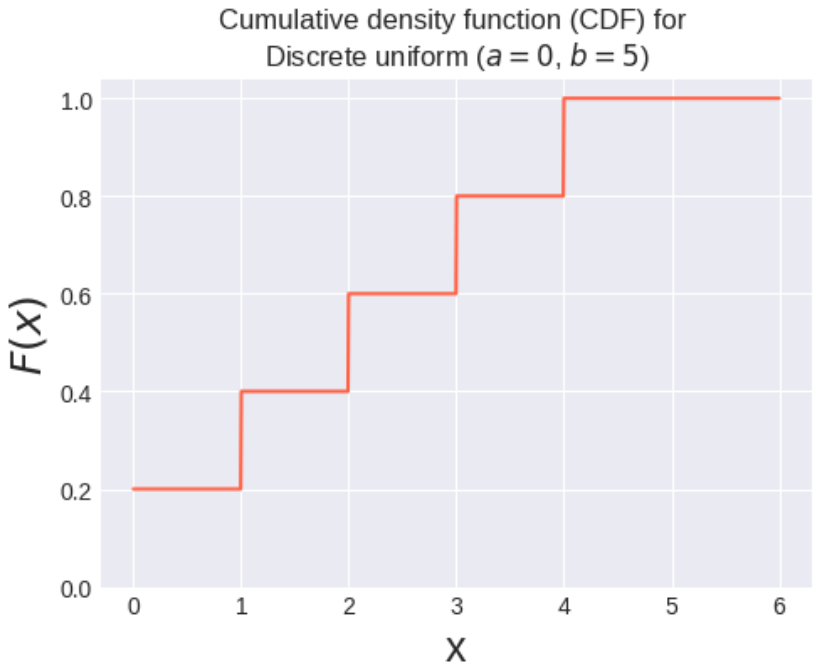
  - Implication:

$$F_X(b) - F_X(a) = P(a \leq X \leq b) = \int_{-\infty}^{b} f_X(t)dt - \int_{-\infty}^{a} f_X(t)dt$$



CDF example plot





Summary: Terminology

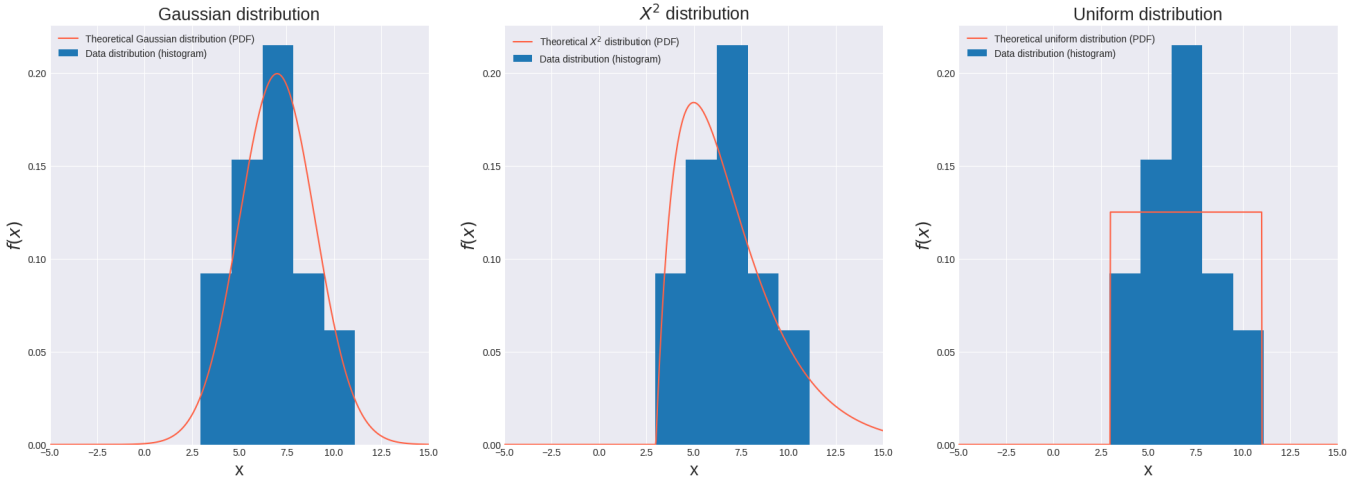
- Experiment
- Sample space
- Event
- Random variable:
  - Discrete random variable
  - Continuous random variable
- Data
- Probability distribution:
  - Discrete distribution: $P(\text{event})$ is described by the probability mass function (PMF)
  - Continuous distribution: $P(\text{event})$ is described by the probability density function (PDF)

  What are their differences?
- Conditional probability of events
- Independent events
- Bayes' rule
- Cumulative distribution function (CDF)

Today

What you will learn from this section

Given a data set, you will learn how to use the Q-Q plot to choose which probability distribution best fits the data.

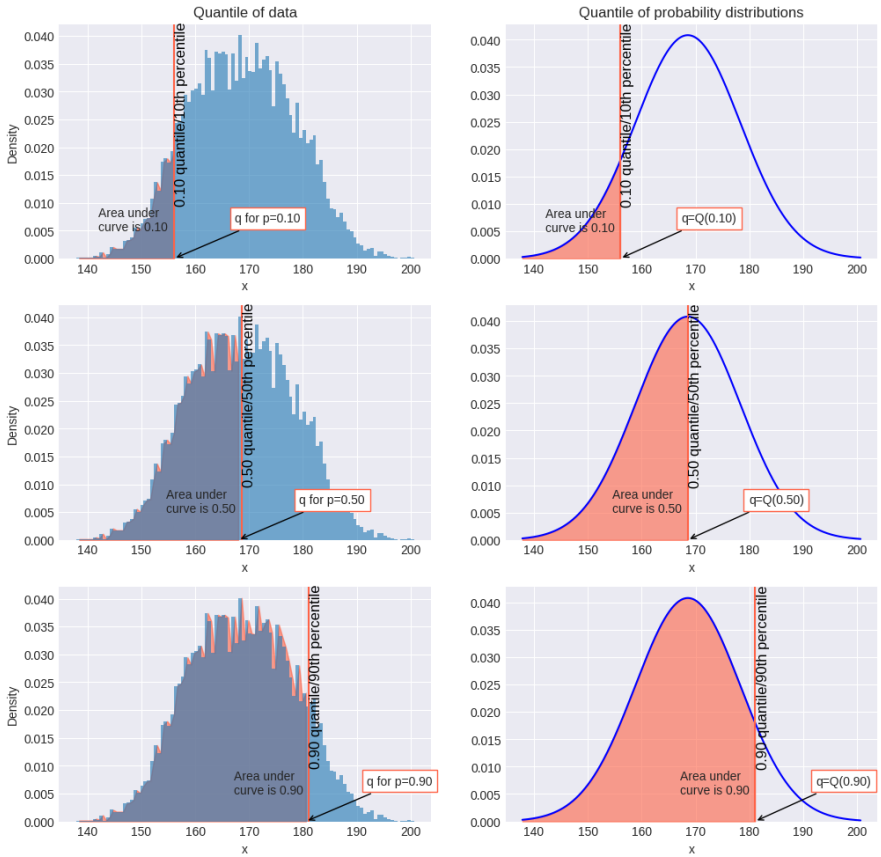


Which one of these three theoretical distributions seems to be the best fit?

Quantiles of a theoretical distribution

Data vs probability distribution

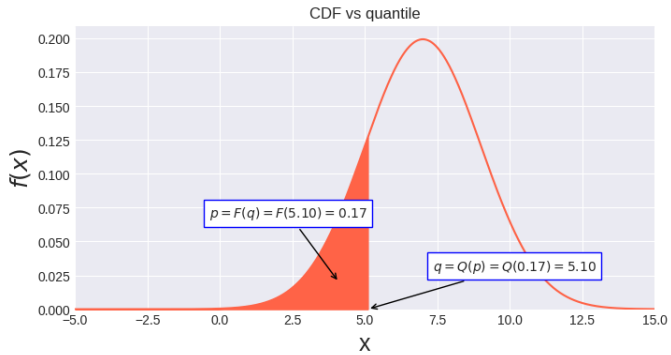
- Recall data quantile: given $p \in (0, 1)$, q is a p-quantile if $(p \times 100)\%$ of the data are below q
- "The area under the curve is the probability of data falling into that interval"



Quantile and CDF

- Quantile function Q is the (generalized) inverse CDF, i.e.

$$F_X(Q(p)) = p \text{ and } Q(F_X(q)) = q$$



- More precisely, given $p \in (0, 1)$, let $Q(p)$ be the quantile function. Then we have

$$Q(p) = \inf\{x : F_X(x) \geq p\}$$

  where inf is the infimum ("smallest") of the set
  Given a probability p (the area under the curve), we are looking for a threshold $x = q$, where $F_X(q)$ is at least p
- In Python (scipy.stats): ppf and cdf
  e.g. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html`

Q-Q plot (quantile-quantile plot)

Definition

- **Q-Q plot (quantile-quantile plot)**: a scatter plot of two sets of quantiles
- **Purpose**: to compare two distributions
- **Intuition**: similar distributions should have similar quantiles
- **Use cases**:
  - Compare a data distribution to a theoretical probability distribution (**one-sample tests**)
  - Compare two data sets to see if they are from the same distribution (**two-sample tests**)
  - Compare two theoretical probability distributions (less common)

How to make the Q-Q plot

Steps: given two distributions

- Choose a set of m probabilities $p_1, p_2, \cdots, p_m \in [0, 1]$ (make sure they spread evenly between 0 and 1)
- For $i = 1, 2, \cdots, m$:
  - Compute the quantile q_i^1 of the first distribution at p_i
  - Compute the quantile q_i^2 of the second distribution at p_i
  - Make a scatter plot of the pair (q_i^1, q_i^2)

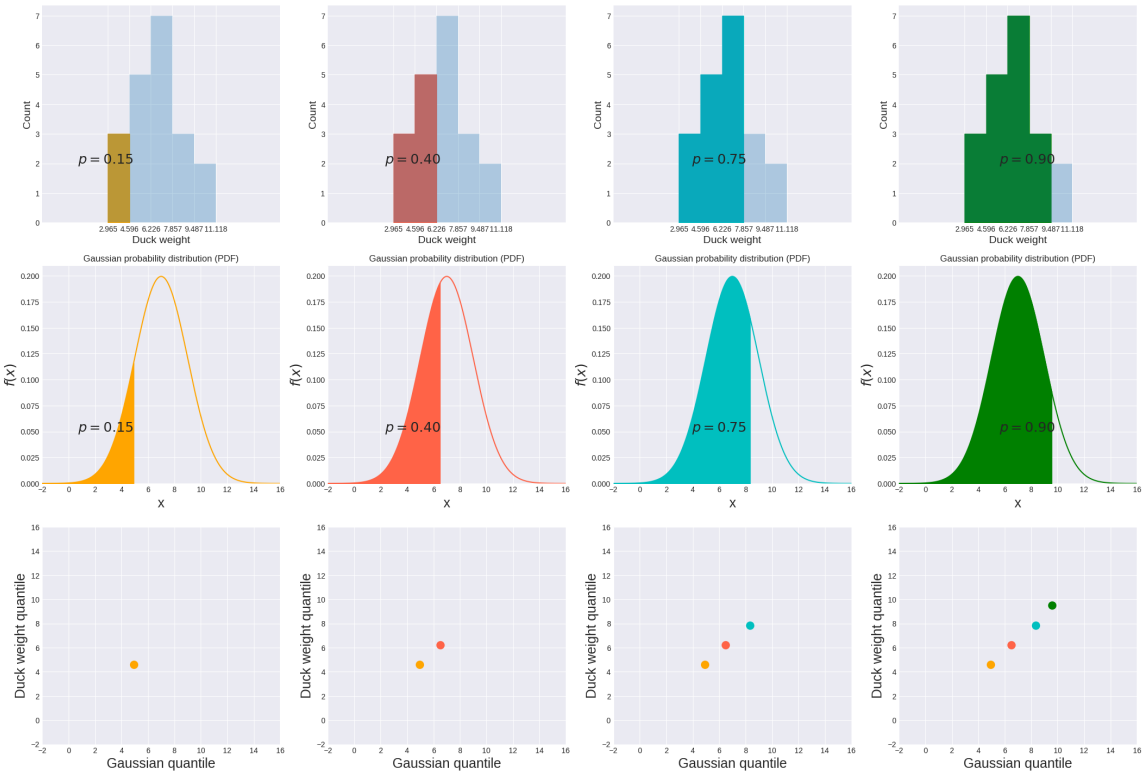
Compare two distributions

Example

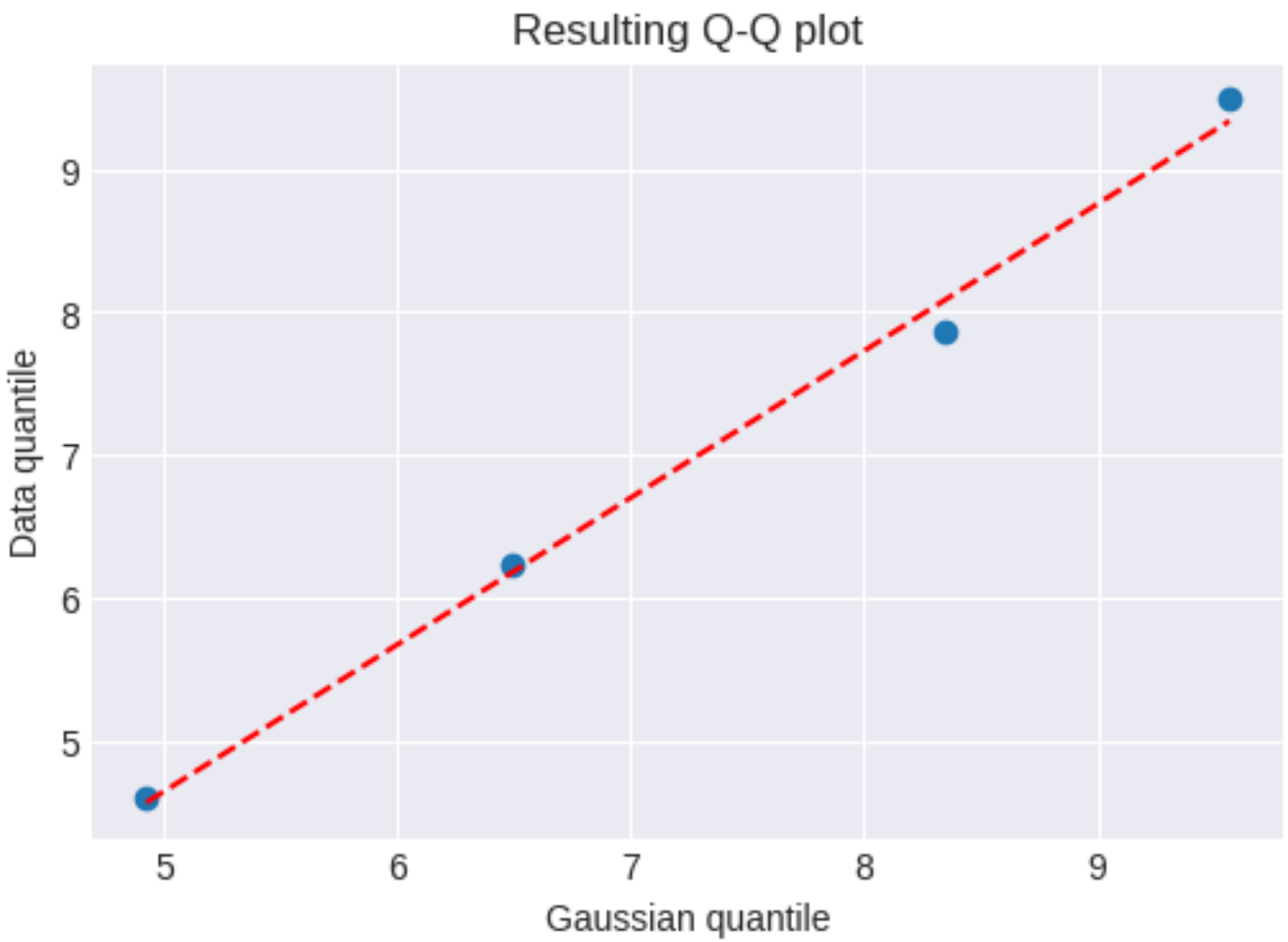
To answer the question "how do you know if my data follows a Gaussian distribution?" Let us look at your ducks

duck id	1	2	3	4	···	19	20
weight	6.98	5.43	2.97	7.07	···	4.63	7.27

and make the Q-Q plots by calculating the quantiles from your data distribution and a Gaussian distribution with given $\mu = 7$ and $\sigma = 2$. **Three steps (cf. 47)**: choose $p = [0.15, 0.40, 0.75, 0.90]$

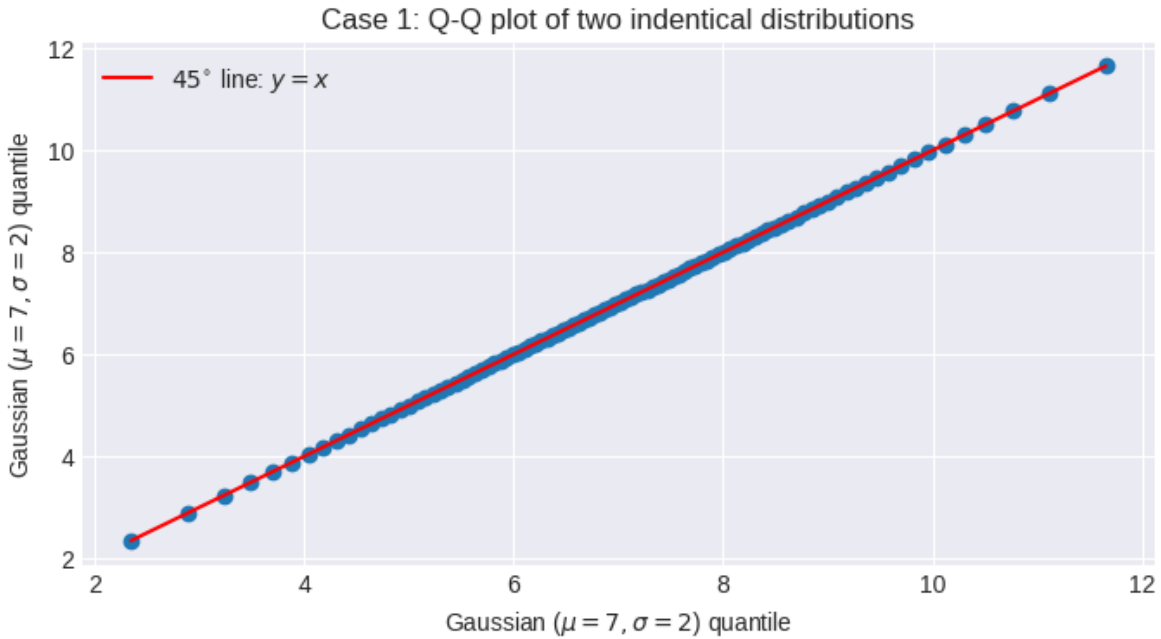


Fit a line to the Q-Q plot



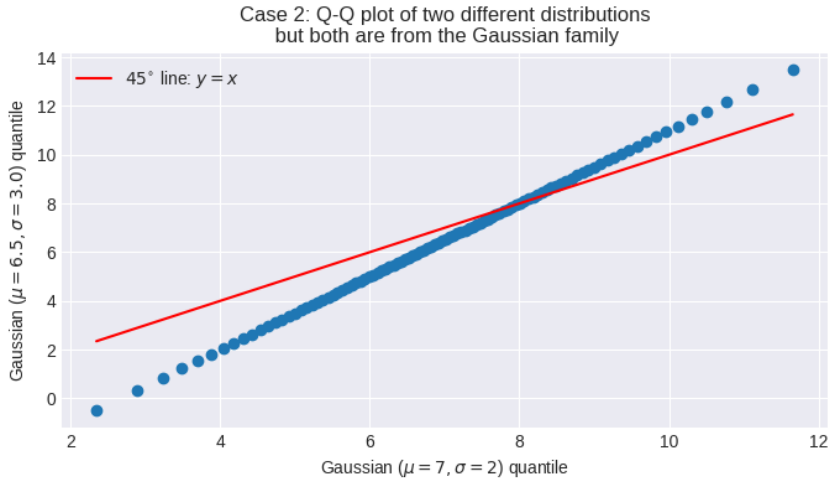
Q-Q plot interpretation: case 1

- Case 1: if the two distributions are identical, the points in the Q-Q plot should follow a 45° straight line $y = x$



Q-Q plot interpretation: case 2

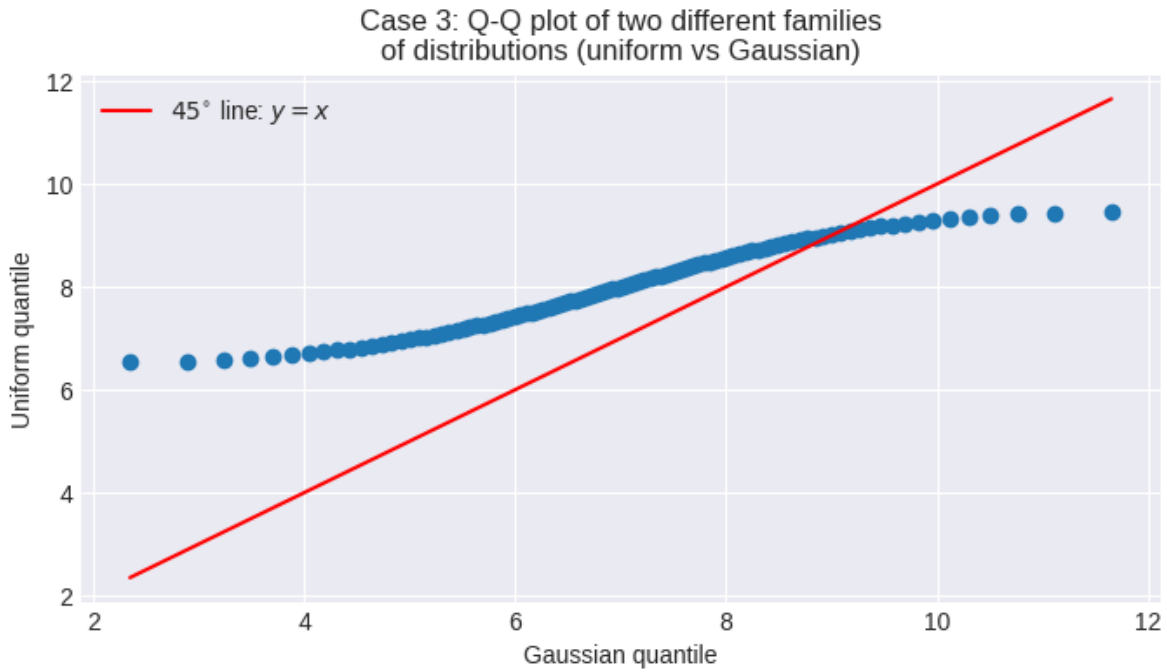
- Case 2: if the two distributions are linearly related, the points in the Q-Q plot follow a straight line that is not necessarily $y = x$



  - Note: if one of the two distributions is a theoretical distribution from a **location-scale family** (e.g. Gaussian distributions), it is very likely that the other distribution is from the same family of distributions.
  - Example: if the two distributions are 1) a theoretical Gaussian distribution with parameters (μ_1, σ_1) and 2) a data distribution; if the points in the Q-Q plot follow a straight line that is not $y = x$, it is very likely that the data follows a Gaussian distribution with a different set of parameters (μ_2, σ_2).

Q-Q plot interpretation: case 3

- Case 3: if the two distributions are from different families of distributions, the points in the Q-Q plot are not lying on a straight line.



Use the Q-Q plot to find a theoretical probability distribution

Steps:

- Given a data set $\mathcal{X} = \{x_1, \cdots, x_N\}$
- Choose several candidate theoretical distributions $D_1, D_2, \cdots$
- Make the Q-Q plot for $\mathcal{X}$ vs D_i for all D_i
- Investigate the resulting Q-Q plots (case 1-3)

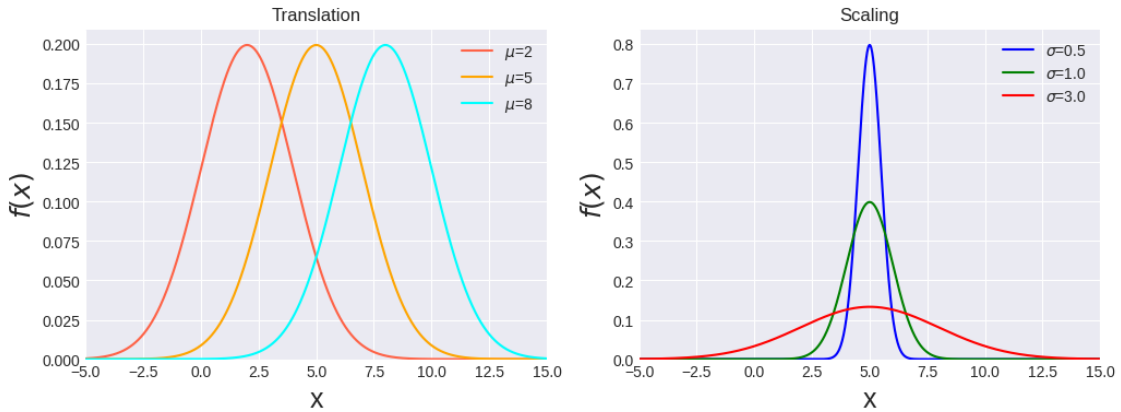
Q-Q plot: additional notes

for interested readers

- The location-scale family of distributions:
  - You will recognize this when you use the **scipy.stats** library!
  - **A family of distributions**: a set of probability distributions, whose PDF/PMF have the **same functional form** with **different parameters**
  - **Definition**: a location-scale family is a family of distributions formed by translation and scaling of a standard family member, where the CDF G can be written as

$$G(x \mid location, scale) = F\left(\frac{x - location}{scale}\right)$$

    where $location \in (-\infty, \infty)$, $scale > 0$, F is the CDF of a standard family member
  - If a distribution family is a location-scale family, we know that they have nice properties we can use; for instance, the family members are linearly related
  - Gaussian distribution (PDF: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$) is a location-scale family



Q-Q plot: additional notes  for those who are interested

- Transformation to a Gaussian distribution:
  - Gaussian is great, because 1) we know everything about it; 2) it's linear - we love linearity - we know how to handle linearity; 3) many things in the world are naturally Gaussian (spoiler alert: central limit theorem).
  - What if data is not Gaussian distributed? - one option is to **transform** data into a Gaussian-like distribution as a preprocessing step.
  - **Example transformations**: power transformation (e.g. Box-Cox transformation, Yeo-Johnson transformation), square root transformation, reciprocal transformation, etc.

    You can try it out in your project if you want! Does it work as expected? If not, what seems to be the problem?

A note on statistical tests for interested readers

- The Q-Q plot is essentially a visualization technique to check similarities between distributions
- There are more analytical testing techniques for the same purpose, for instance, z-test, t-test, Kolmogorov-Smirnov test, Wilcoxon's signed-rank test, Mann-Whiteney U test, $\mathcal{X}^2$-test, etc
- How do you know which test to choose? One can ask the following questions to find an appropriate statistical test to use
  - What are the data types? Categorical? Numerical? Discrete? Continuous?
  - How many variables you have? One? Two? Many?
  - Parametric test or nonparametric test?
  - Are variables independent?
  - Do you want to compare two data distributions or a data distribution against a theoretical probability distribution?
  - If you want to compare two data distributions, are they paired?
  - ...
- We will revisit this topic soon

Summary

- We used a Q-Q plot to visually verify the hypothesis that the data follows a Gaussian distribution by showing that points in the Q-Q plot follow a straight line
- We learned how to use a Q-Q plot to compare different probability distribution candidates for describing a data set
- Some useful concepts: cumulative distribution function (CDF), quantiles of a theoretical distribution, location-scale family of distributions
- Statistical tests as analytical alternatives to the Q-Q plot

Today

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
- Q-Q plot, CDF

Not yet:

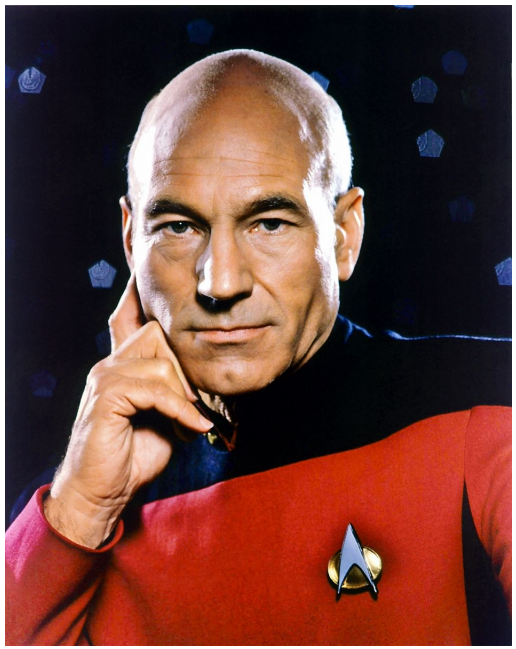
- How to estimate parameters, such as μ and σ in a Gaussian distribution?

Next:

- mathematical modeling, parameter estimation

Before next lecture:

- PMF and PDF
- Independent events
- Bayes' rule



Pretty confident