Lecture 3: Q-Q plot and mathematical modeling

Statistical Methods for Data Science

Yinan Yu

Department of Computer Science and Engineering

November 7, 2022

Today

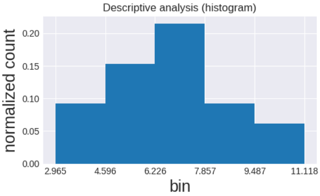
1. Compare two distributions using a Q-Q plot
   - Cumulative distribution function (CDF)
   - Quantiles of a theoretical distribution
   - Q-Q plot (quantile-quantile plot)
   - Compare two distributions
2. Mathematical modeling
3. Summary

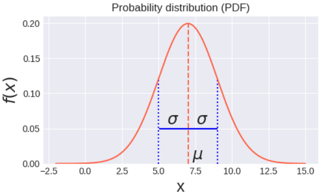
Learning outcome

- 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
- Be able to explain different components in a mathematical model $y = g(x; \theta \mid h)$

Recap: three questions from lecture 2

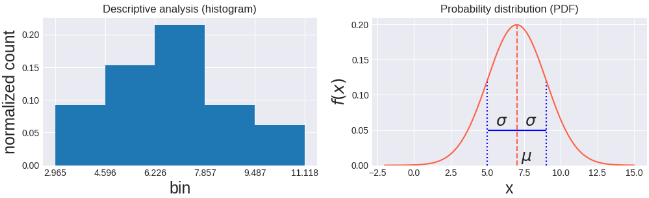
Jack suggested to use a Gaussian distribution to model your data.





Recap: three questions from lecture 2

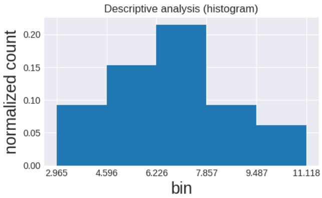
Jack suggested to use a Gaussian distribution to model your data.

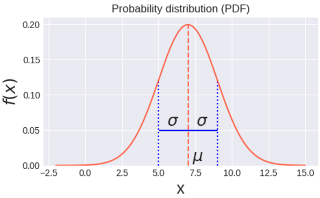


✅ Question 1: Why should I use probability distributions instead of histograms?

Recap: three questions from lecture 2

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

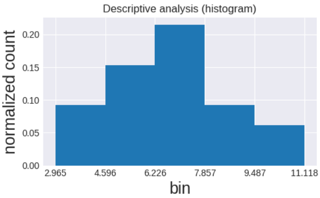


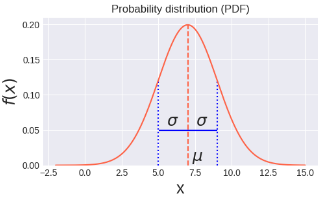


- ✅ Question 1: Why should I use probability distributions instead of histograms?
- ? Question 2: How do you know if my data follows a Gaussian distribution?

Recap: three questions from lecture 2

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

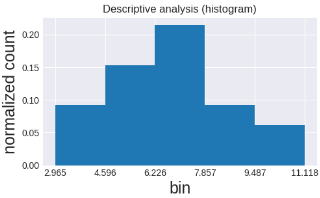


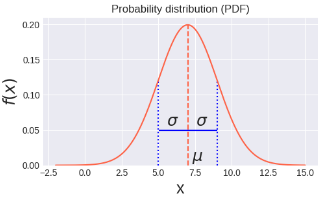


- ✅ Question 1: Why should I use probability distributions instead of histograms?
- ? Question 2: How do you know if my data follows a Gaussian distribution?
- ? Question 3: How do I find the unknown parameters?

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?

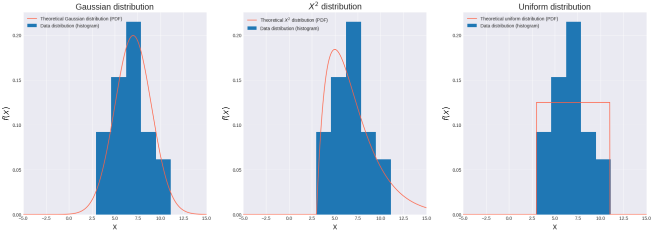
In today's lecture, we are going to address question 2.

Today

1. Compare two distributions using a Q-Q plot
   - Cumulative distribution function (CDF)
   - Quantiles of a theoretical distribution
   - Q-Q plot (quantile-quantile plot)
   - Compare two distributions
2. Mathematical modeling
3. Summary

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?

Cumulative distribution function (CDF)

🚨 Terminology alert 🚨

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

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

🚨 Terminology alert 🚨

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

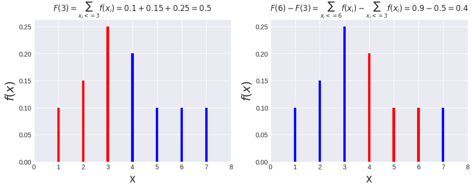
$$F_X(x) = P(X \leq x)$$

- X discrete random variable:
  - **Definition**: given the PMF f_X,

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

    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

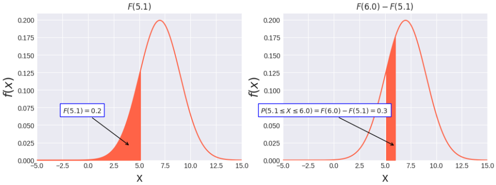
$$F_X(x) = P(X \leq x)$$

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

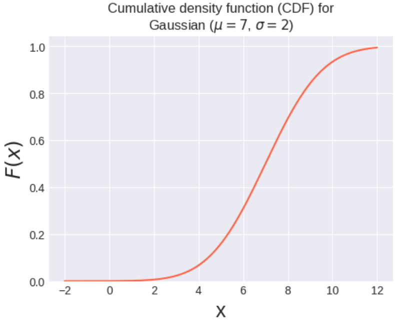
$$F_X(\mathbf{x}) = P(X \leq \mathbf{x}) = \int_{-\infty}^{\mathbf{x}} f_X(t)dt$$

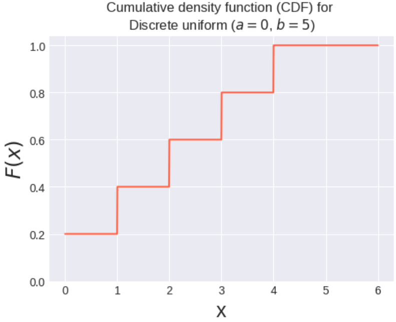
  - Implication:

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



CDF example plot

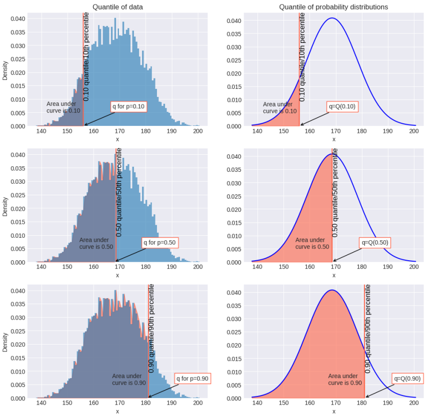




Quantiles of a theoretical distribution

Data vs probability distribution

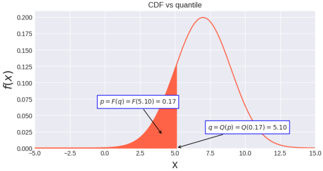
- Recall data quantile: given $p \in (0, 1)$, q is a p-quantile if $p \times 100\%$ of the data are below q
- "The area under the curve is the probability of data falling into that interval"



Quantile and CDF

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

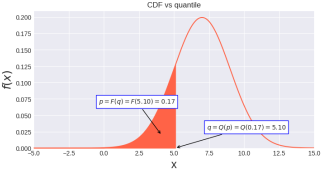
$$F_X(Q(p)) = p \text{ and } Q(F_X(q)) = q$$



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

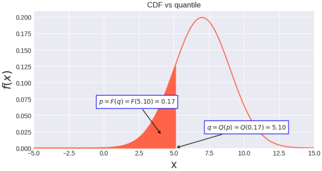
$$Q(p) = \inf\{x : F_X(x) \geq p\}$$

where inf is the infimum ("smallest") of the set

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

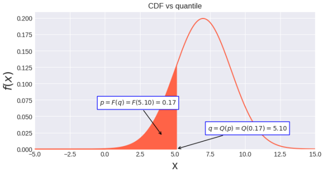
$$Q(p) = \inf\{x : F_X(x) \geq p\}$$

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

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

Definition

- **Q-Q plot (quantile-quantile plot)**: a scatter plot of two sets of quantiles
- **Purpose**: to compare two distributions

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

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**:

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**)

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**)

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

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)

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 $i = 1, 2, \cdots, m$:

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

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

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 $\sigma = 2$.

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

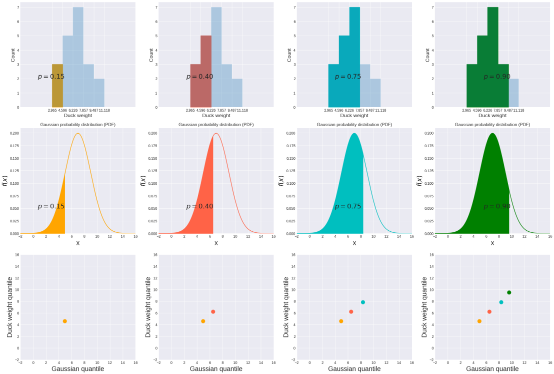
and make the Q-Q plots by calculating the quantiles from your data distribution and a Gaussian distribution with given $\mu = 7$ and $\sigma = 2$. **Three steps (cf. 16)**: choose $p = [0.15, 0.40, 0.75, 0.90]$

Example

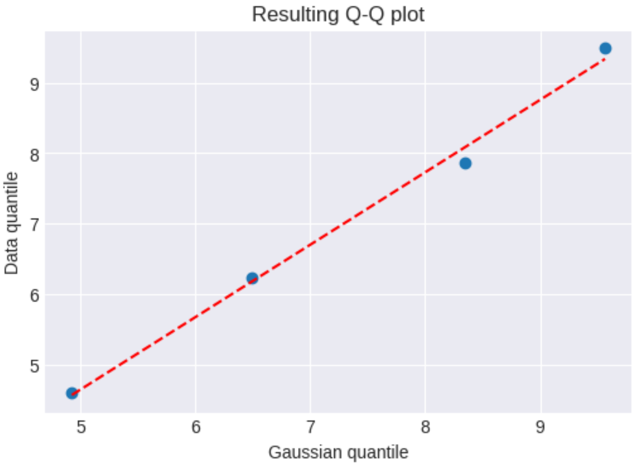
To answer the question "how do you know if my data follows a Gaussian distribution?" Let us look at your ducks

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. 16)**: choose $p = [0.15, 0.40, 0.75, 0.90]$

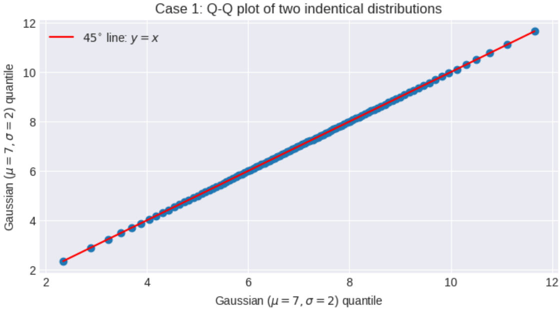


Fit a line to the Q-Q plot



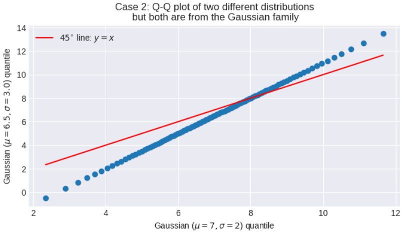
Q-Q plot interpretation: case 1

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



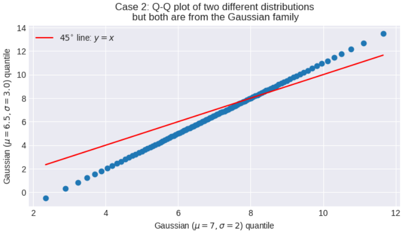
Q-Q plot interpretation: case 2

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



Q-Q plot interpretation: case 2

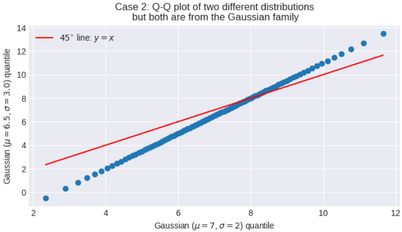
- Case 2: if the two distributions are linearly related, the points in the Q-Q plot follow a straight line that is not necessarily $y = x$



  - 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.

Q-Q plot interpretation: case 2

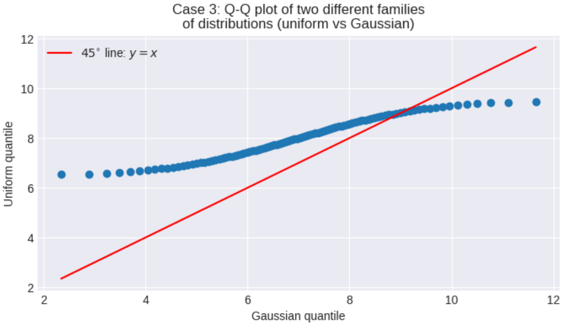
- Case 2: if the two distributions are linearly related, the points in the Q-Q plot follow a straight line that is not necessarily $y = x$



- 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:

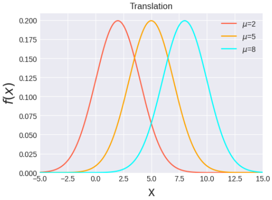
- Given a data set $\mathcal{X} = \{x_1, \cdots, x_N\}$
- Choose several candidate theoretical distributions $D_1, D_2, \cdots$
- Make the Q-Q plot for $\mathcal{X}$ vs D_i for all D_i
- Investigate the resulting Q-Q plots (case 1-3)

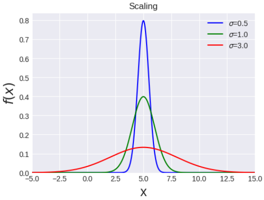
Q-Q plot: additional notes ♫ ♫ for interested readers

- 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:

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).

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?

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.

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.

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

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?
  - ...

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

What you will learn from this section

In the previous section, we have touched upon the topic of choosing a probabilistic model to describe a given data set. This is also known as mathematical modeling.

Generally speaking, given a data set and a problem to be solved, you need to formulate the solution mathematically so that you can write a computer program to solve the problem. This is the main task for a data scientist.

This section aims to help you get started by providing explicit components and steps for formulating mathematical models.

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

    - Left hand side:

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict;

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems;

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption;

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
    - h: hyperparameters - part of the model g (given or derived from your assumption);

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
    - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g;

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
- Symbols:

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
- Symbols:
  - Semicolon (";") is used to emphasize that θ is not known for free - it needs to be estimated

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

- Left hand side:
  - y: target or label - what you want to predict; **a result** that answers the question at hand
- Right hand side:
  - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
  - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
  - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
  - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
- Symbols:
  - Semicolon (";") is used to emphasize that θ is not known for free - it needs to be estimated
  - Bar ("|" pronounced "*given*") is used to indicate that h is known to you

Terminology

- What is mathematical modeling? - Mathematical modeling is to *describe* a system using the language of mathematics in order to solve **a range of problems**.
- What the *description* looks like in data science:

$$y = g(x; \theta \mid h)$$

  - Left hand side:
    - y: target or label - what you want to predict; **a result** that answers the question at hand
  - Right hand side:
    - x: variables or features - placeholder for data in order to solve *a range of* problems; **the input**
    - g: model - a mathematical function that can be used to solve a given range of problems - a model is given by domain experts or derived from your assumption; can be selected from established models; **known** except for some parameters
    - h: hyperparameters - part of the model g (given or derived from your assumption); **known** (but you might need to "guess" them first)
    - θ: parameters - part of the model g; in a data-driven paradigm θ is **unknown**; need to be estimated from data
  - Symbols:
    - Semicolon (";") is used to emphasize that θ is not known for free - it needs to be estimated
    - Bar ("|" pronounced "*given*") is used to indicate that h is known to you
  - Note: x, y, θ and h are not necessarily scalars; they can be multiple scalars, vectors or more complex data structures; g can be complex functions, for instance, a machine learning model or a deep neural network.

Five questions

Overwhelmed? Take it easy! Here is something that helps you get started! Answer these five questions in the language of mathematics step by step:

- 1) What do we want to predict, i.e. what is the target y?
- 2) What are the variables x?
- 3) What is the mathematical function g that relates variables x to the target y?
- 4) Are there any hyperparameters h in the function g? How do we choose them?
- 5) What are the unknown parameters θ in g? How do we estimate them from data?

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it?

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$;

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs):

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

  - Expression (15 secs):

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you predict the probability of this new duck weighing between 5 kg and 7 kg before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can use established mathematical models to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem using probabilistic models
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) assumption - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

  - Expression (15 secs): given the assumption in Eq. (1), we can write g as

$$g(x; \theta) = g(x_1, x_2; \theta) = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2}$$

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

  - Expression (15 secs): given the assumption in Eq. (1), we can write g as

$$g(x; \theta) = g(x_1, x_2; \theta) = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2}$$

- 4) Are there any hyperparameters h in the function g? How do we choose them? (5 secs)

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \quad (1)$$

  - Expression (15 secs): given the assumption in Eq. (1), we can write g as

$$g(x; \theta) = g(x_1, x_2; \theta) = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \quad (2)$$

- 4) Are there any hyperparameters h in the function g? How do we choose them? (5 secs)
  Answer: By looking at Eq. (2), we don't seem to have any hyperparameter here

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

    $$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

  - Expression (15 secs): given the assumption in Eq. (1), we can write g as

    $$g(x;\theta) = g(x_1, x_2; \theta) = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2}$$

- 4) Are there any hyperparameters h in the function g? How do we choose them? (5 secs)
  Answer: By looking at Eq. (2), we don't seem to have any hyperparameter here
- 5) What are the unknown parameters θ in g? (10 secs)

Example - modeling walkthrough it's like a video game walkthrough but twice the fun!

You will get a new duck tomorrow and you will measure its weight when it arrives (exciting!). Can you **predict the probability** of this new duck **weighing between 5 kg and 7 kg** before measuring it? Let's answer the five questions!

- 1) What do we want to predict, i.e. what is the target y? (15 secs)
  Answer: $y = P(5 \leq \text{weight} \leq 7)$
  Now we want to generalize the problem so that we can **use established mathematical models** to make the prediction $y = g(x; \theta \mid h) = P(5 \leq \text{weight} \leq 7)$; let's model the problem with a probability distribution
- 2) What are the variables x? (15 secs)
  Answer: define two variables $x = (x_1, x_2)$ so that we can pose the question $P(5 \leq \text{weight} \leq 7)$ in a more generic form $P(x_1 \leq \text{weight} \leq x_2)$ and solve the problem **using probabilistic models**
- 3) What is the mathematical function g that relates variables x to the target y? This comes from your (hopefully reasonable) **assumption** - hint: write down your assumption first and then try to construct g
  Answer:
  - Assumption (15 secs): the weight is generated from a theoretical probability distribution - here we assume a Gaussian distribution

$$\text{weight} \sim \mathcal{N}(\mu, \sigma^2) \text{ , with PDF } f_{\text{weight}}(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{1}$$

  - Expression (15 secs): given the assumption in Eq. (1), we can write g as

$$g(x; \theta) = g(x_1, x_2; \theta) = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2}$$

- 4) Are there any hyperparameters h in the function g? How do we choose them? (5 secs)
  Answer: By looking at Eq. (2), we don't seem to have any hyperparameter here
- 5) What are the unknown parameters θ in g? (10 secs)
  Answer: From Eq. (2), we see two **unknown** parameters $\theta = (\mu, \sigma)$

Example - modeling walkthrough

- Put everything together, we get our model:

$$y = P(x_1 \leq weight \leq x_2) = g(x_1, x_2; \mu, \sigma) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{3}$$

- As soon as we find the values for μ and σ, we can answer the question by plugging $x_1 = 5$ and $x_2 = 7$ into Eq. (3):

$$y = \int_{5}^{7} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$$

Example - Python implementation

- How do we implement this model in Python?

Example - Python implementation

- How do we implement this model in Python?
- Recall the cumulative distribution function (CDF) function F on page 9

$$y = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = F_{\text{weight}}(x_2) - F_{\text{weight}}(x_1)$$

Example - Python implementation

- How do we implement this model in Python?
- Recall the cumulative distribution function (CDF) function F on page 9

$$y = P(x_1 \leq \text{weight} \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = F_{\text{weight}}(x_2) - F_{\text{weight}}(x_1)$$

```
from scipy.stats import norm # Gaussian (normal) distribution
mean = ... # \mu: unknown for now
std = ...  # \sigma: unknown for now
F_x1 = norm.cdf(x=5, loc=mean, scale=std) # CDF at 5
F_x2 = norm.cdf(x=7, loc=mean, scale=std) # CDF at 7
y = F_x2 - F_x1
```

There are many available probability distributions in the scipy.stats library:
https://docs.scipy.org/doc/scipy/reference/stats.html

A nonrigorous note on functions and variables

- Let g be a function that relates input variables x to a target y:

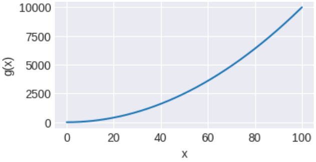
$$y = g(x)$$

- Typically, we care about the behavior of y for **all possible values for x**. This is called **generalization** in machine learning.
- Even if we add parameters θ and hyperparameters h to g, $g(x; \theta \mid h)$ is still a function of x.
- In a plot, we typically place the variable on the x-axis!
- If we are interested in the behavior of y in terms of θ, we can construct a different function L that takes θ as the variables $y = L(\theta)$ to relate θ to y.

A nonrigorous note on functions and variables (cont.)

- Example: $y = g(x) = x^2$
- In Python, **all possible values for x** means something like this:

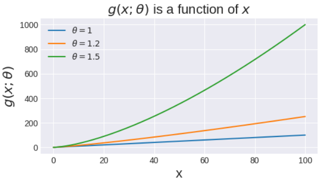
```
# Assume x can take any value between 0 and 100
xmin, xmax = 0, 100
N = 10000 # ideally, N should be infinity. But sadly, computers are discrete
          # so N has to be finite.
x = np.linspace(xmin, xmax, num=N) # all possible values for x
# Plot a function
def g(t):
    return np.power(t, 2)
y = g(x)
plt.plot(x, y)
```



A nonrigorous note on functions and variables (cont.)

- Now we add a parameter θ to g: $y = g(x; \theta) = x^{\theta}$

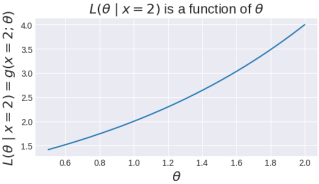
```
def g_theta(t, theta):
    return np.power(t, theta)
xmin, xmax = 0, 100 # assume x can take any value between 0 and 100
N = 10000
x = np.linspace(xmin, xmax, num=N) # all possible values for x
y = g_theta(x, 1)
plt.plot(x, y)
y = g_theta(x, 1.2)
plt.plot(x, y)
y = g_theta(x, 1.5)
plt.plot(x, y) # x is still on the x-axis
```



A nonrigorous note on functions and variables (cont.)

- Now we define a new function: $y = L(\theta \mid x = 2) = g(x = 2; \theta) = 2^\theta$

```
def L(t):
    return g_theta(2, t)
# Now theta is the variable! So we need to get all possible values for theta
# Assume theta can take any value between 0.5 and 2
theta_min, theta_max = 0.5, 2
N = 10000
thetas = np.linspace(theta_min, theta_max, num=N) # all possible values for theta
y = L(thetas)
plt.plot(thetas, y) # theta is on the x-axis now
```



A nonrigorous note on functions and variables (cont.)

- Make sure you are comfortable with this
- This is important for understanding the (¡spoiler alert!) **likelihood function**

Summary

- Mathematical modeling is to describe a system with a mathematical expression $y = g(x; \theta \mid h)$ in order to solve a range of problems.
- Five questions to help you get started:
  - 1) What do we want to predict, i.e. what is the target y?
  - 2) What are the variables x?
  - 3) What is the mathematical function g that relates variables x to the target y?
  - 4) Are there any hyperparameters h in the function g? How do we choose them?
  - 5) What are the unknown parameters θ in g?

Summary

- Mathematical modeling is to describe a system with a mathematical expression $y = g(x; \theta \mid h)$ in order to solve a range of problems.
- Five questions to help you get started:
  - 1) What do we want to predict, i.e. what is the target y?
  - 2) What are the variables x?
  - 3) What is the mathematical function g that relates variables x to the target y?
  - 4) Are there any hyperparameters h in the function g? How do we choose them?
  - 5) What are the unknown parameters θ in g? **How do we estimate them from data?**

Summary

- Mathematical modeling is to describe a system with a mathematical expression $y = g(x; \theta \mid h)$ in order to solve a range of problems.
- Five questions to help you get started:
  - 1) What do we want to predict, i.e. what is the target y?
  - 2) What are the variables x?
  - 3) What is the mathematical function g that relates variables x to the target y?
  - 4) Are there any hyperparameters h in the function g? How do we choose them?
  - 5) What are the unknown parameters θ in g? **How do we estimate them from data?**

Practice makes perfect! Try to formulate a problem at hand using these steps to see if you understand them completely! If you have any questions, do not hesitate to ask me!

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, mathematical modeling

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, mathematical modeling

Not yet:

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

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, mathematical modeling

Not yet:

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

Next:

- parameter estimation

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, mathematical modeling

Not yet:

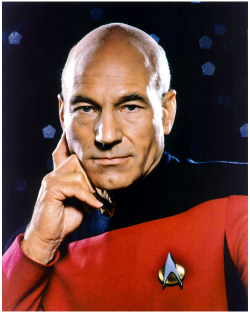
- How to estimate parameters, such as μ and σ in a Gaussian distribution?

Next:

- parameter estimation

Before next lecture:

- PMF and PDF
- Independent events
- Bayes' rule



Pretty confident