ECE 20875
Python for Data Science

Chris Brinton and David Inouye

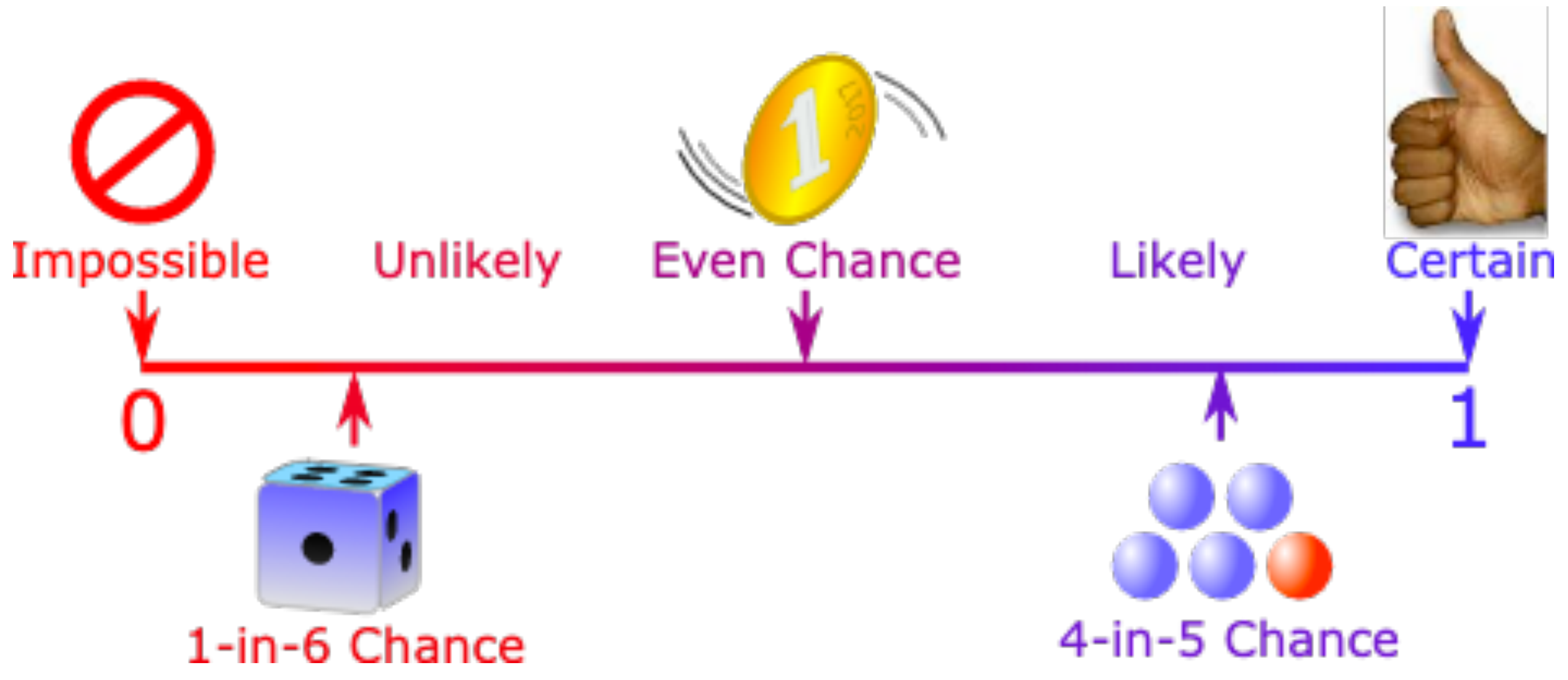
(Adapted from material developed by Prof. Milind Kulkarni and Prof. Chris Brinton)

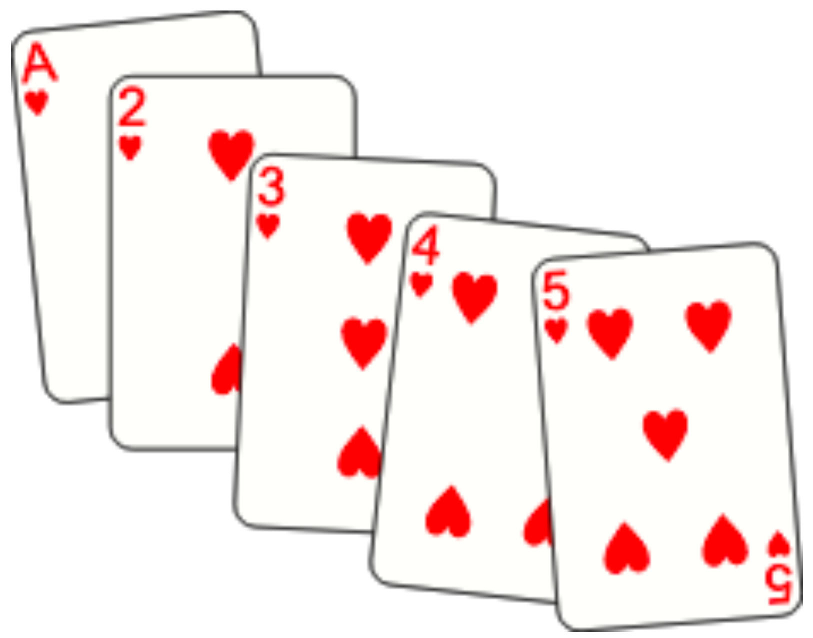
Probability and Random Variables

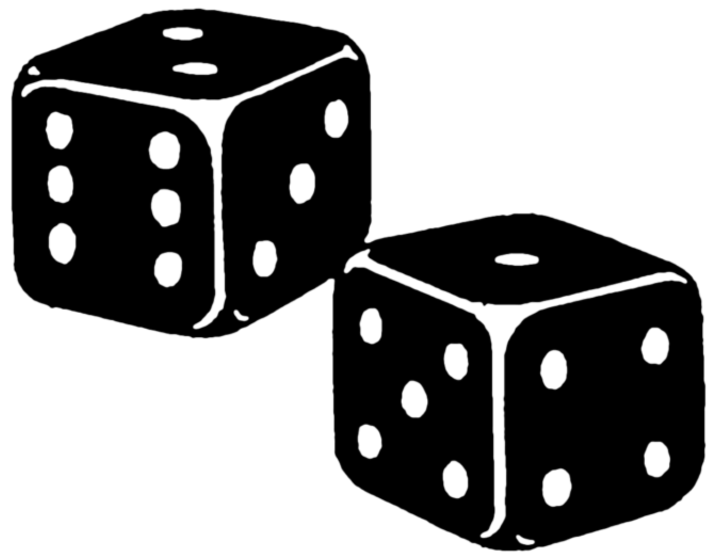
what is a probability?

- Measure of likelihood that an event occurs
- A number between 0 and 1
- The higher the number, the more likely the event occurs
    - A probability of 0 means the event never occurs, and a probability of 1 means the event always occurs
- Example: What is the probability of the event "heads" when flipping a coin?

$$P(H) =$$

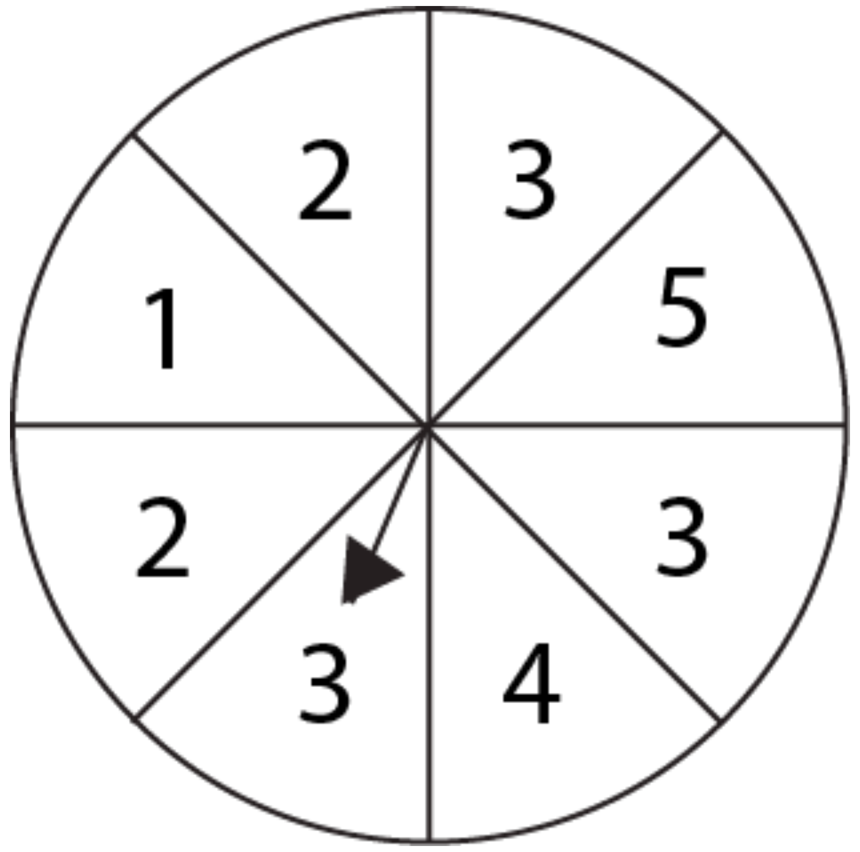






elements of a probability model

- Conduct an **experiment**, which results in an **outcome**
  - Each outcome has a probability between 0 and 1
  - Set of all possible outcomes is the **sample space** Ω
  - Sum of probability of all outcomes is 1
- An **event** is a set of possible outcomes
  - Probability of event is the sum of the probabilities of individual outcomes



$$\Omega = \{1,\ldots,5\}$$

$$P(3) = 3/8$$

$$P(\{1,3,5\}) = 5/8$$

visualization

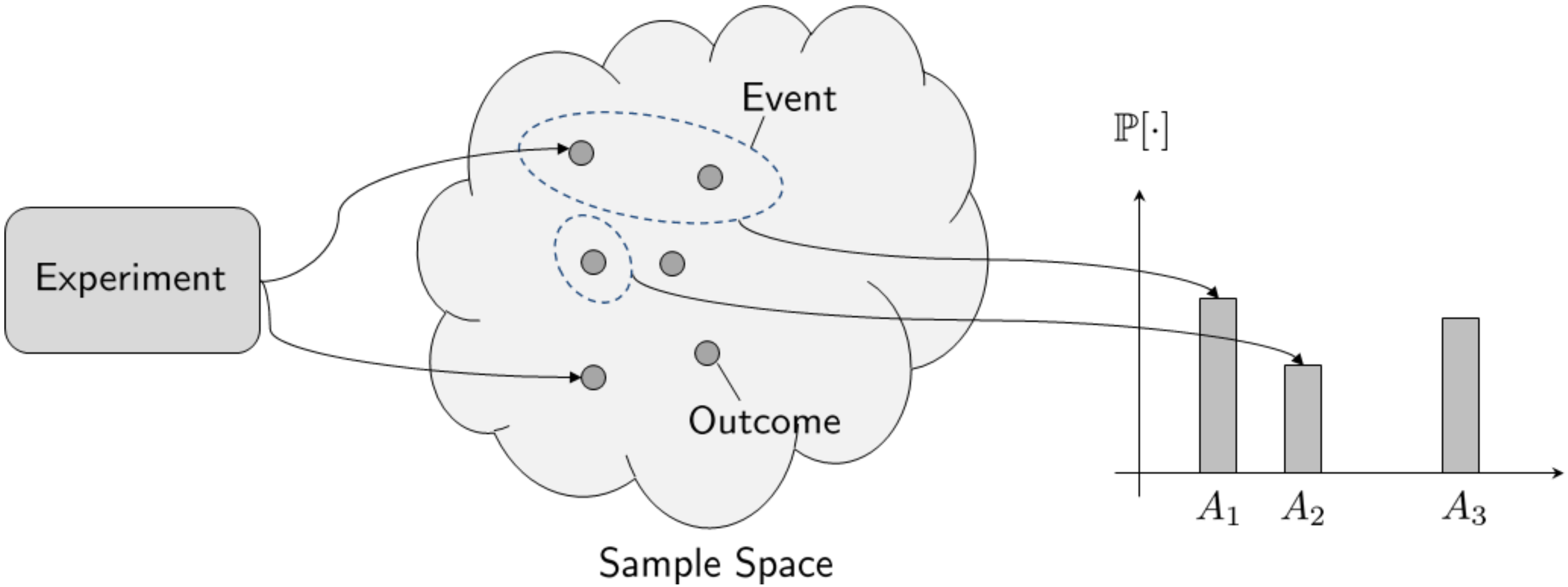
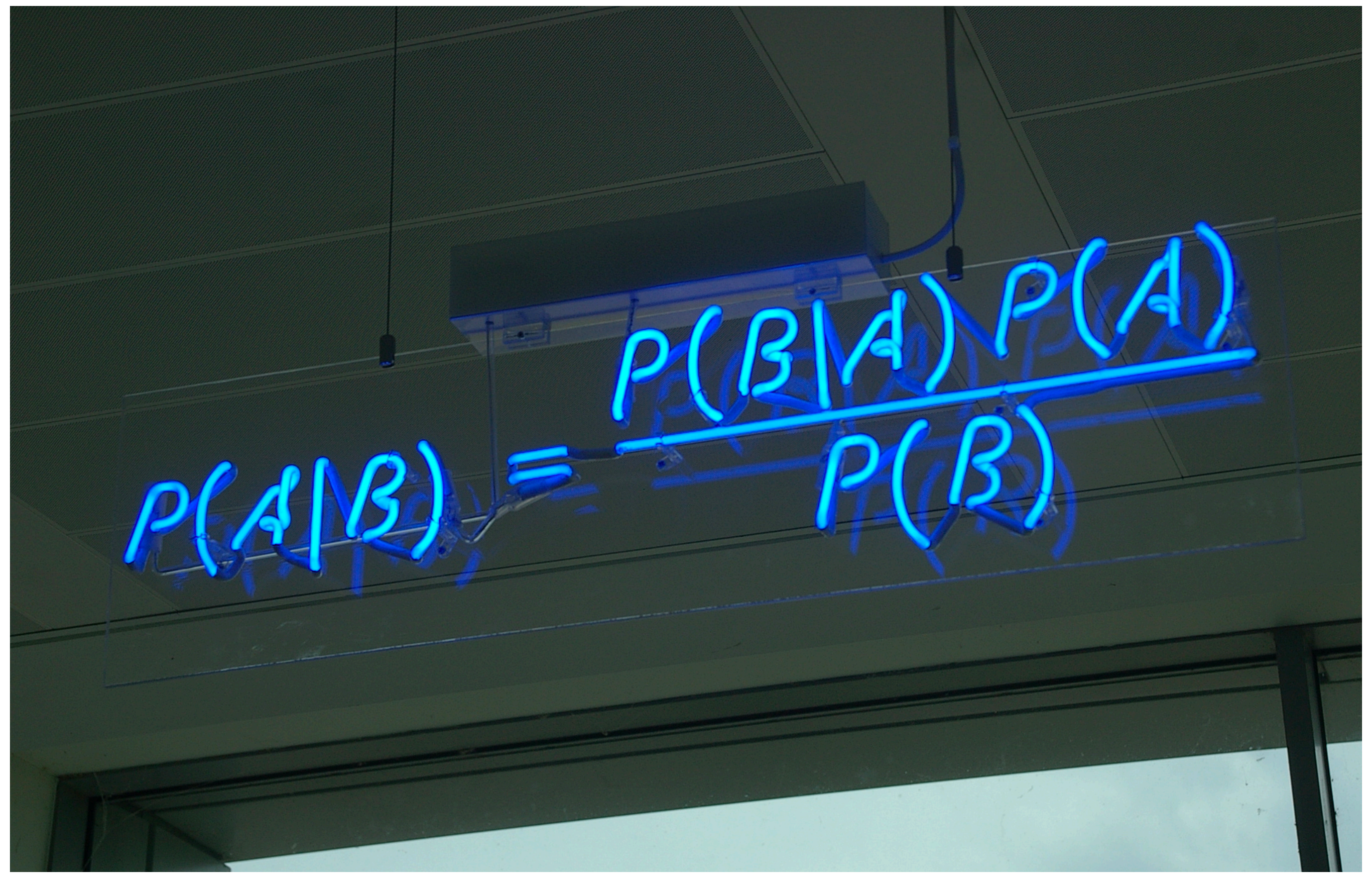


Figure courtesy Stanley Chan

what does probability mean?

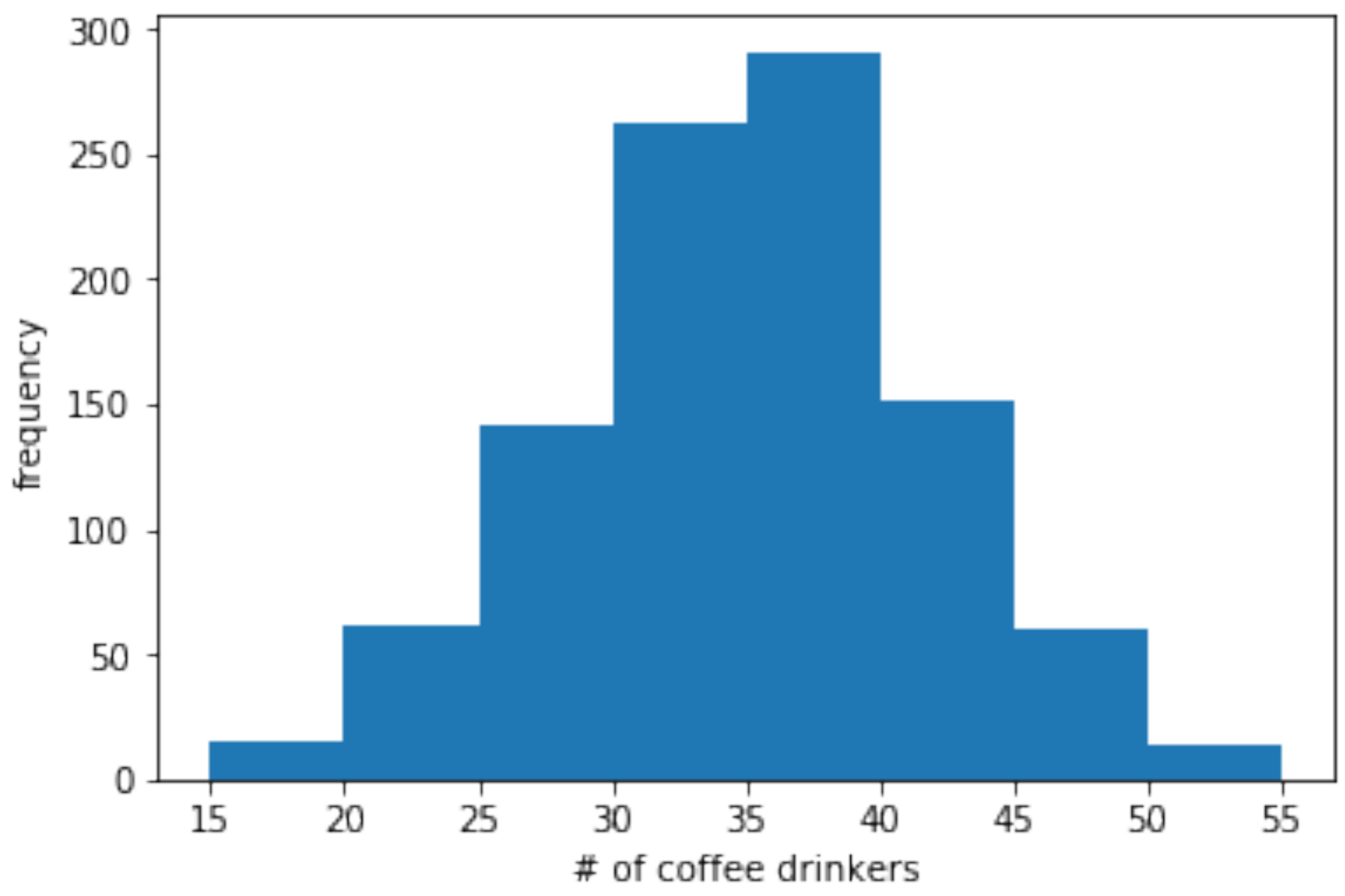
- Lots of different interpretations
  - All outcomes x are equally probable (e.g., roll a die, each number has the same chance). Probability of an event is number of outcomes in event divided by total number of outcomes.
  - **Frequentist**: Repeat an experiment over and over again, probability of an event is fraction of the time the event happens during the experiment.
  - **Bayesian**: Probability is a reflection of your belief about the likelihood of something happening (e.g., based on prior knowledge).

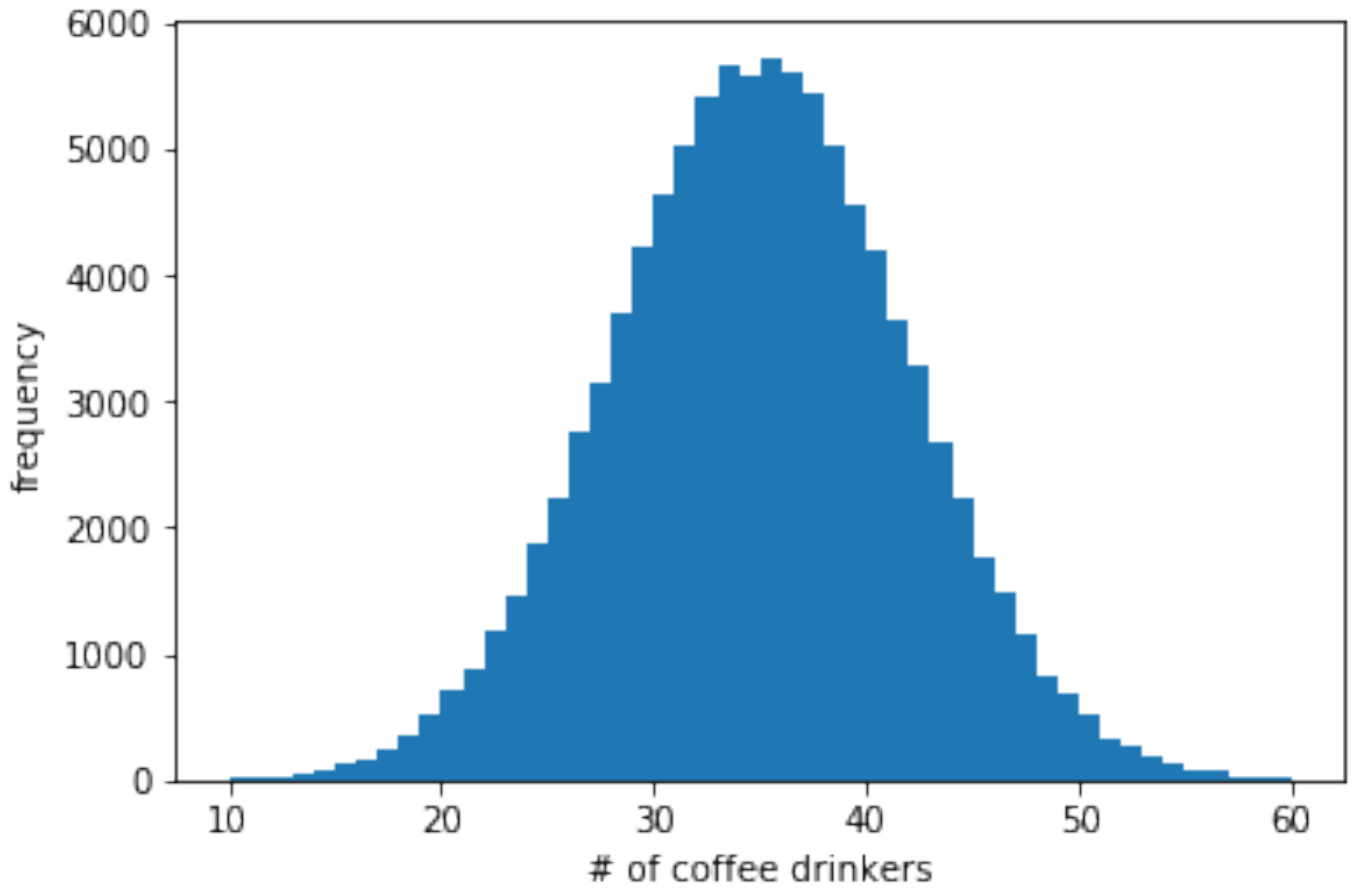


random variables

- A **random variable** X is a function that assigns an outcome to a number
  - A way of letting us treat outcomes, which may not be numbers, in a mathematical way
  - E.g., in flipping a coin, X could map Heads to 0 and Tails to 1
- A random variable has a probability distribution which tells us the probability of its values
  - E.g., in flipping a coin, $P[X = 0] = 0.5, \;\; P[X = 1] = 0.5$
- Informal intuition: The random variable is the horizontal value on the histogram, with the height being the probability
- Random variables can be **continuous** or **discrete**

from histogram to probability



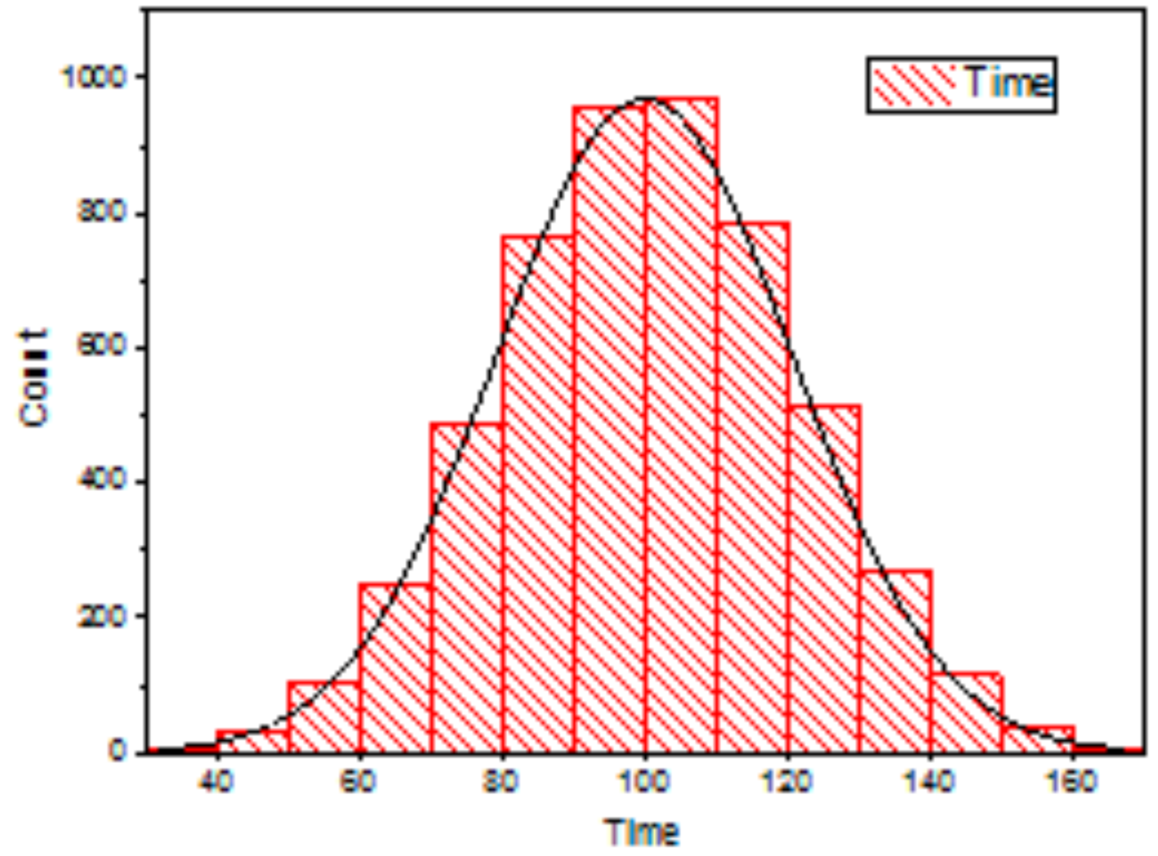


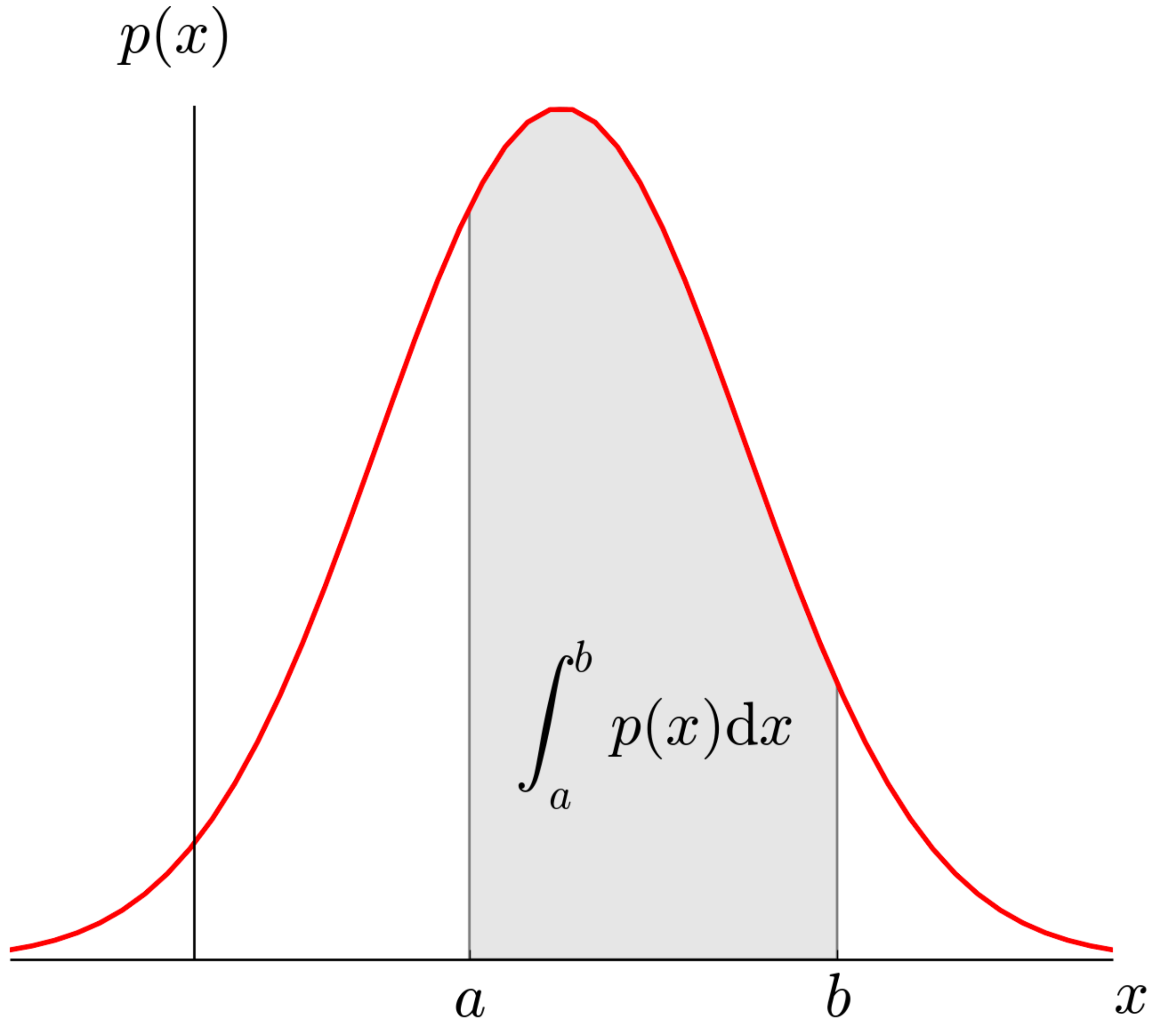
probability density function

- One loose definition: A histogram when …
  - (i) the number of samples goes to infinity
  - (ii) the bin width approaches zero
  - When this happens, the estimate $\hat{p}_k$ approaches p_k of the population
- More formal definition: $f_X(x)$ is the **probability density function** (PDF) for X if

$$P[a \le X \le b] = \int_a^b f_X(x)\, dx$$

  - X is a random variable

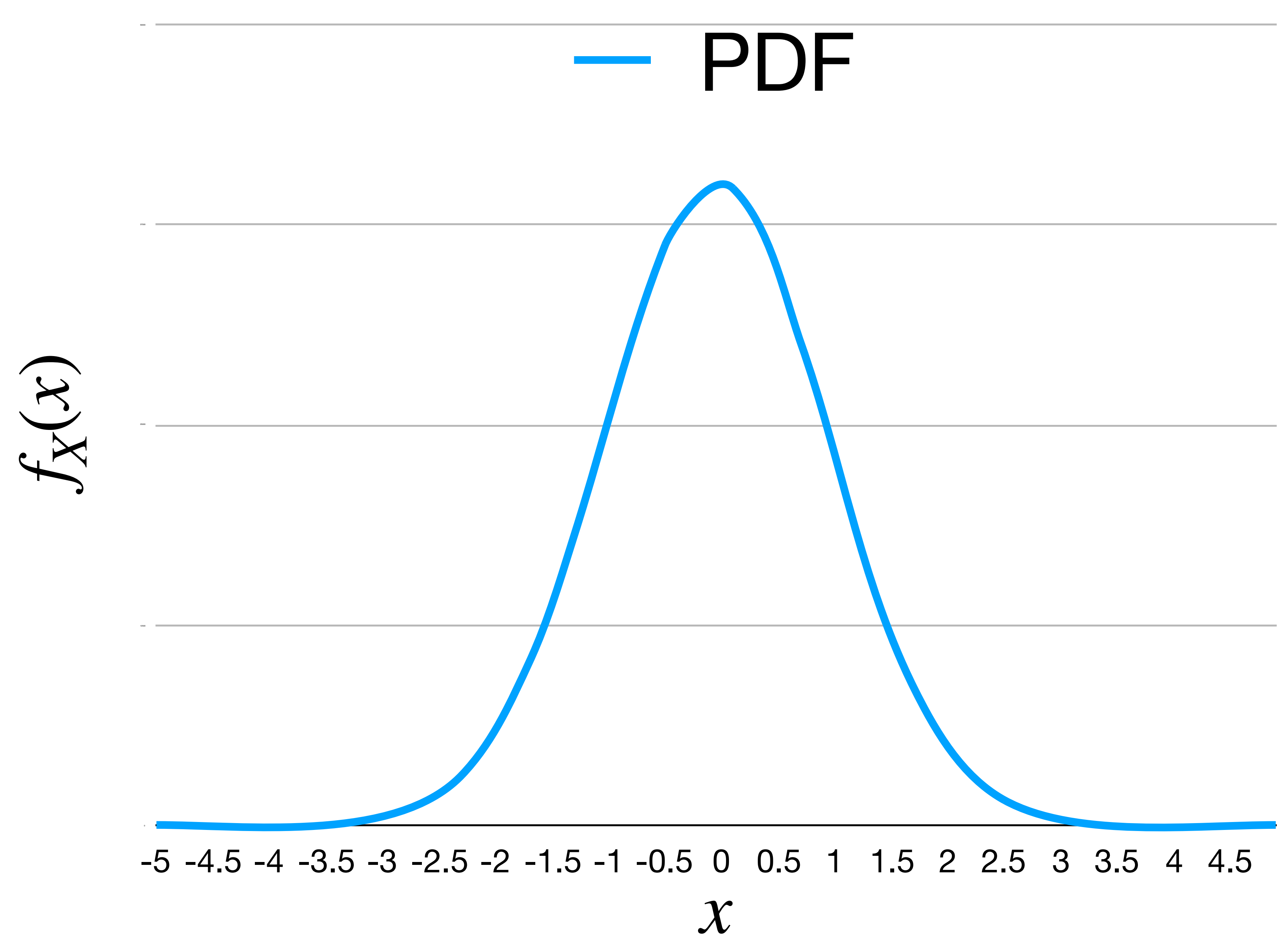


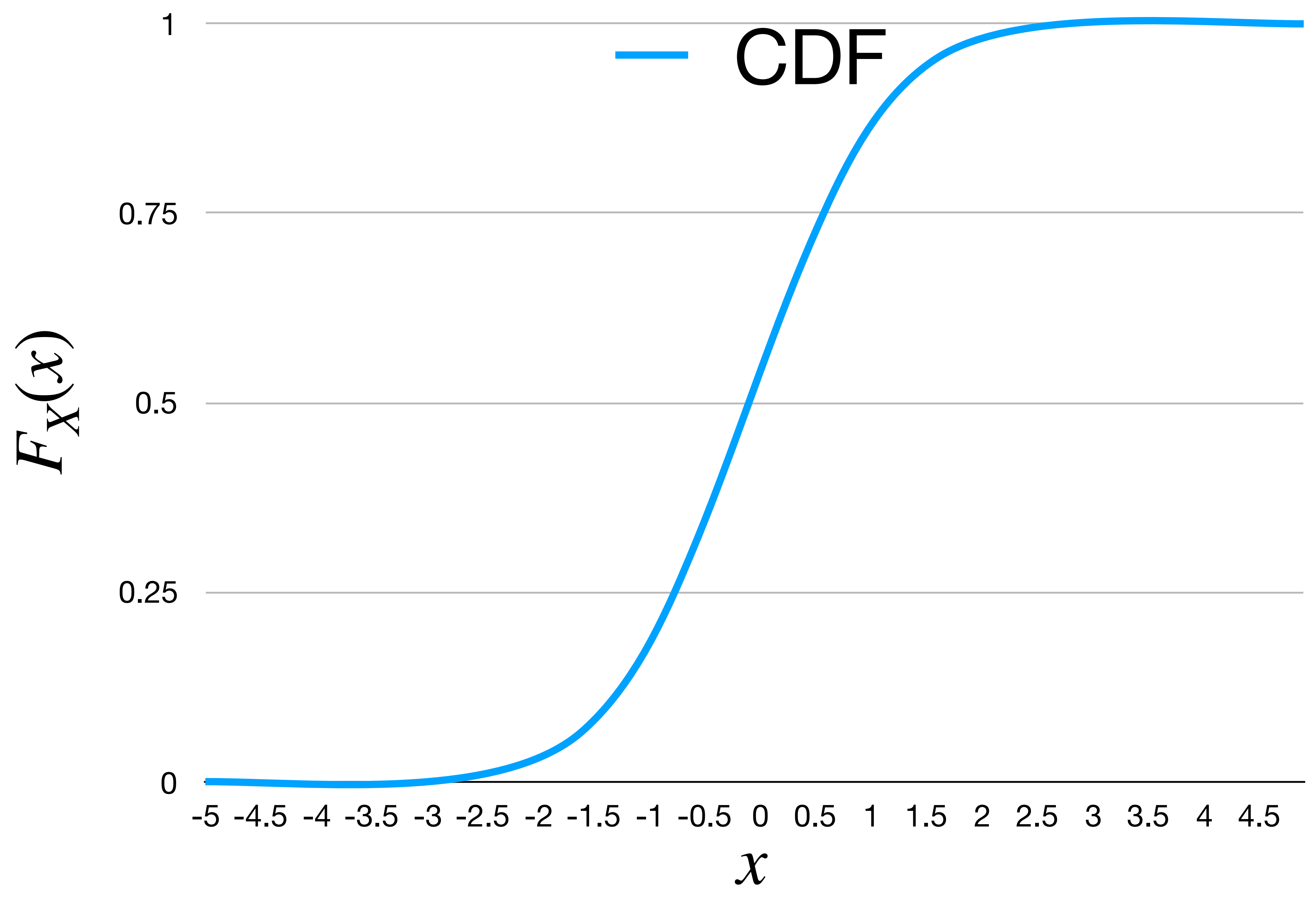


cumulative distribution function

- The **cumulative distribution function** (CDF) of a random variable X is

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

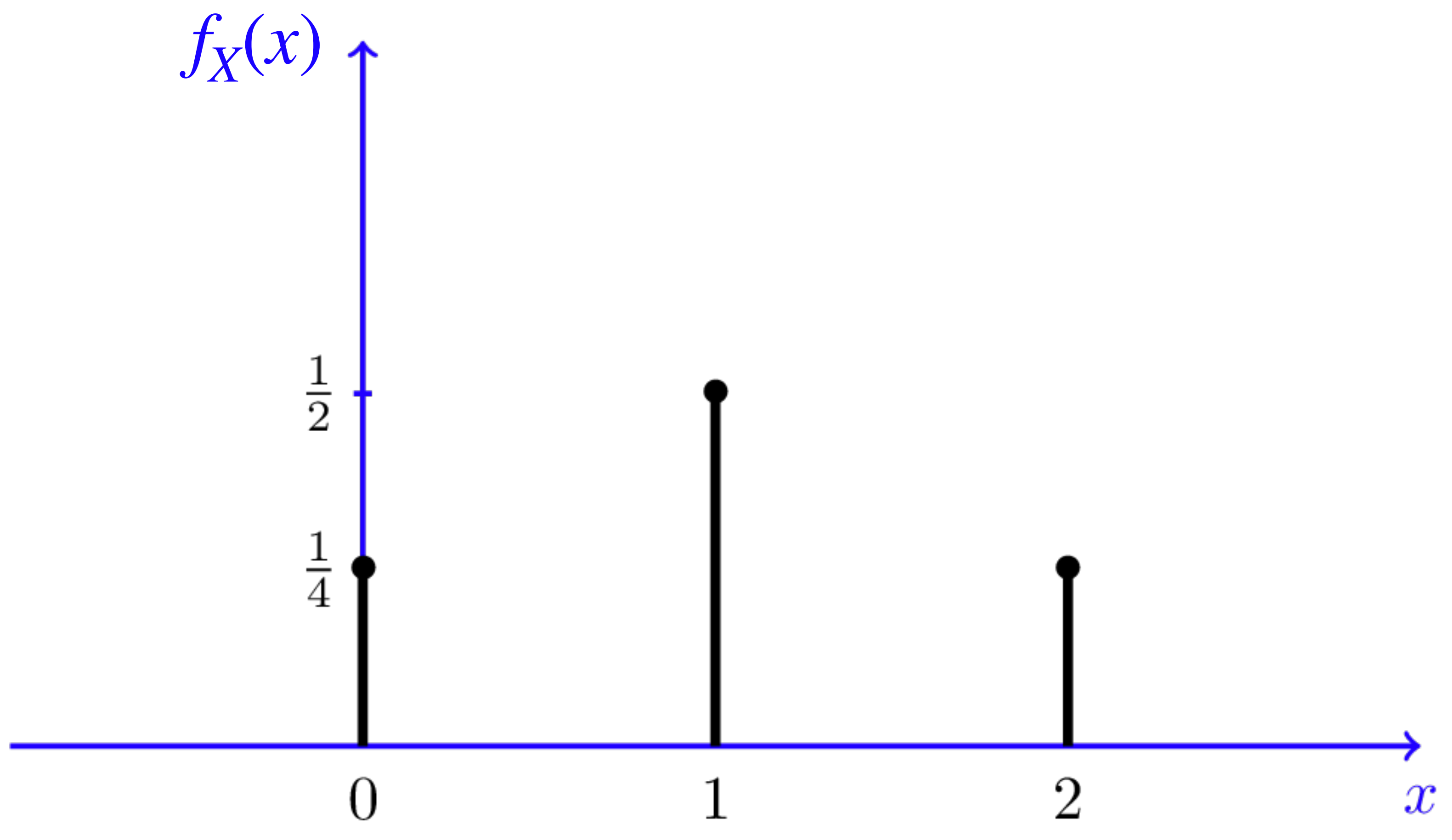




probability mass/density function

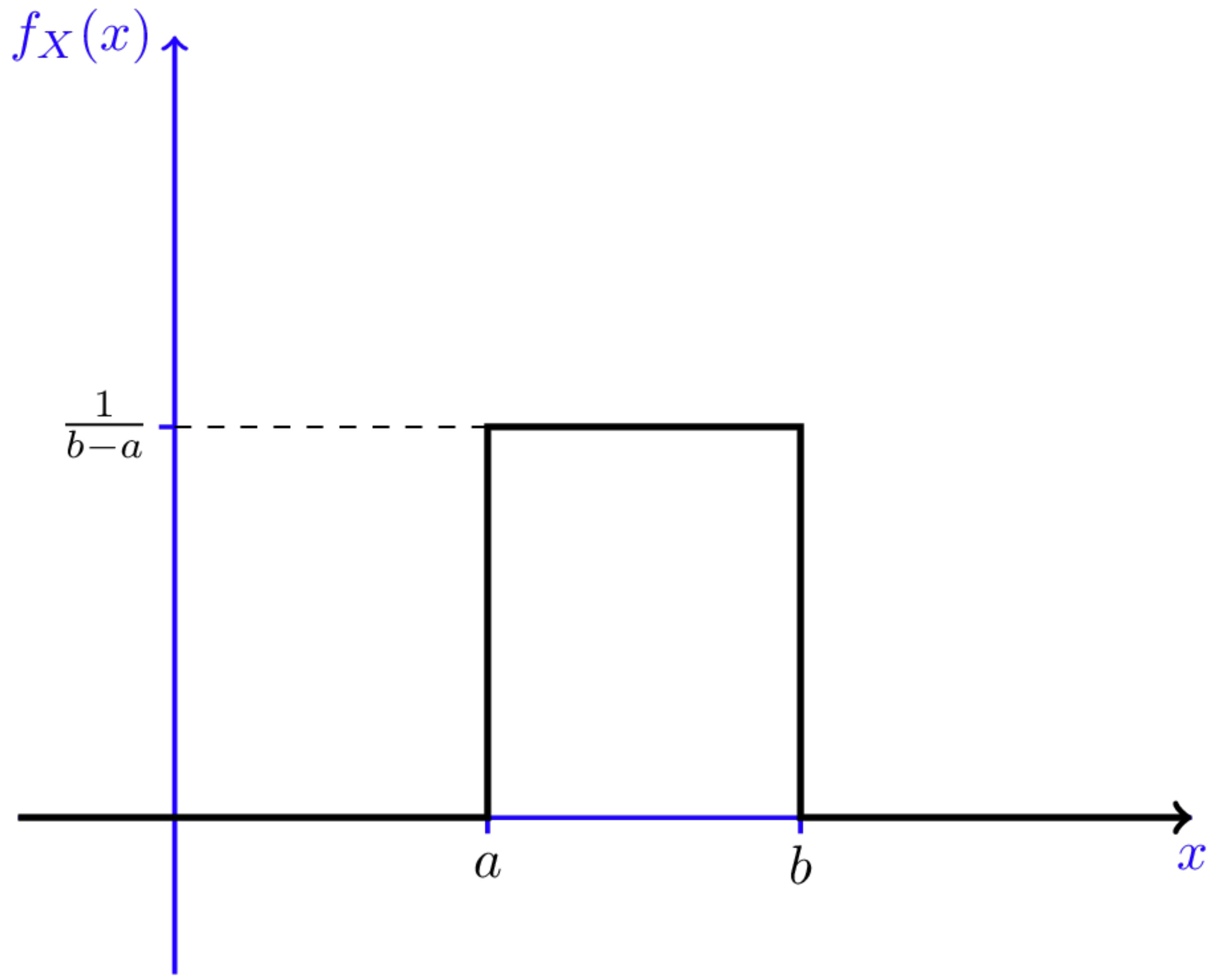
- If X is a discrete random variable, it has a **probability mass function** (**PMF**). The PMF is defined directly from the probabilities of events (essentially a histogram with bars interpreted as frequencies):

$$f_X(x) = P[X = x]$$



- If X is a continuous random variable, it has a PDF, which is a little tricker to define since the probability of any single number is actually 0. As a result, we typically define the PDF in terms of the CDF:

$$f_X(x) = \frac{dF_X(x)}{dx}$$



CDF from PDF/data

- The continuous CDF $F_X(x)$ in terms of the PDF $f_X(x)$:

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

- The discrete CDF $F_X(x)$ in terms of the PMF $f_X(x) = P[X = x]$:

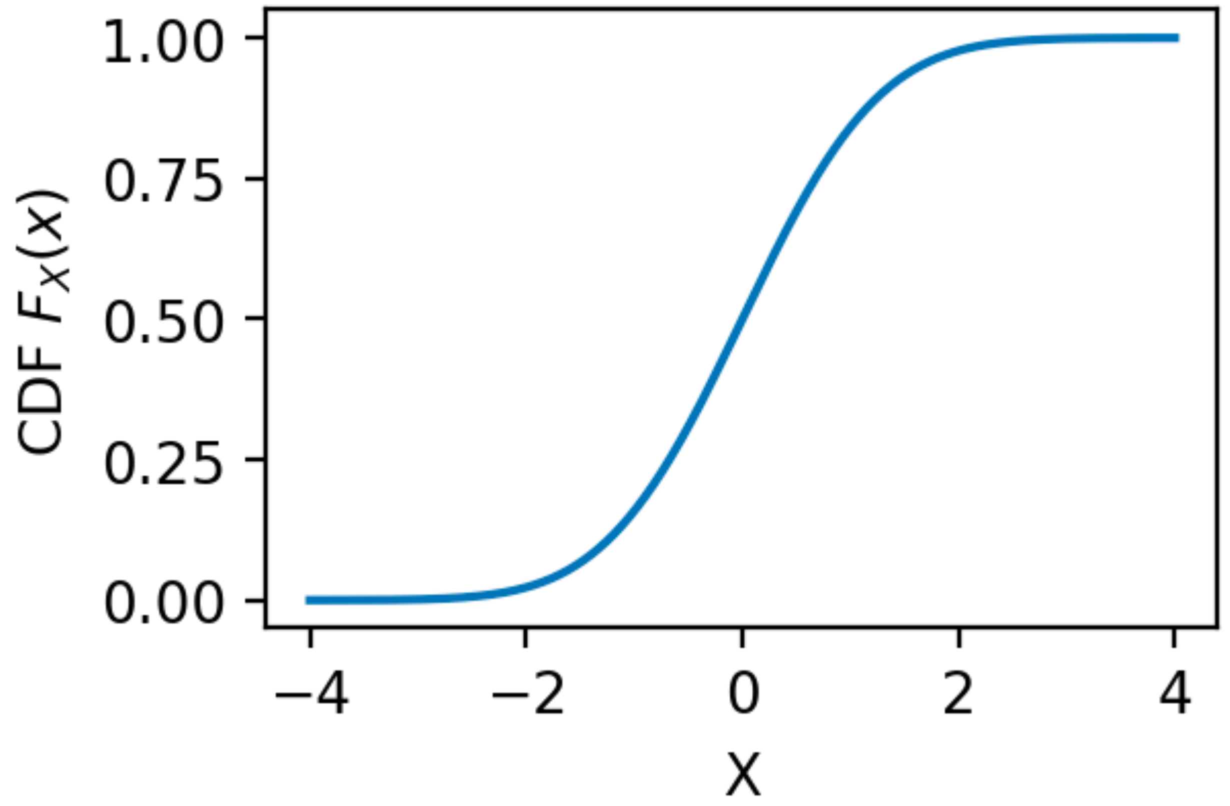
$$F_X(x) = P[X \leq x] = P[-\infty \leq X \leq x] = \sum_{x_i \leq x} f_X(x_i) = \sum_{x_i \leq x} P[X = x_i]$$

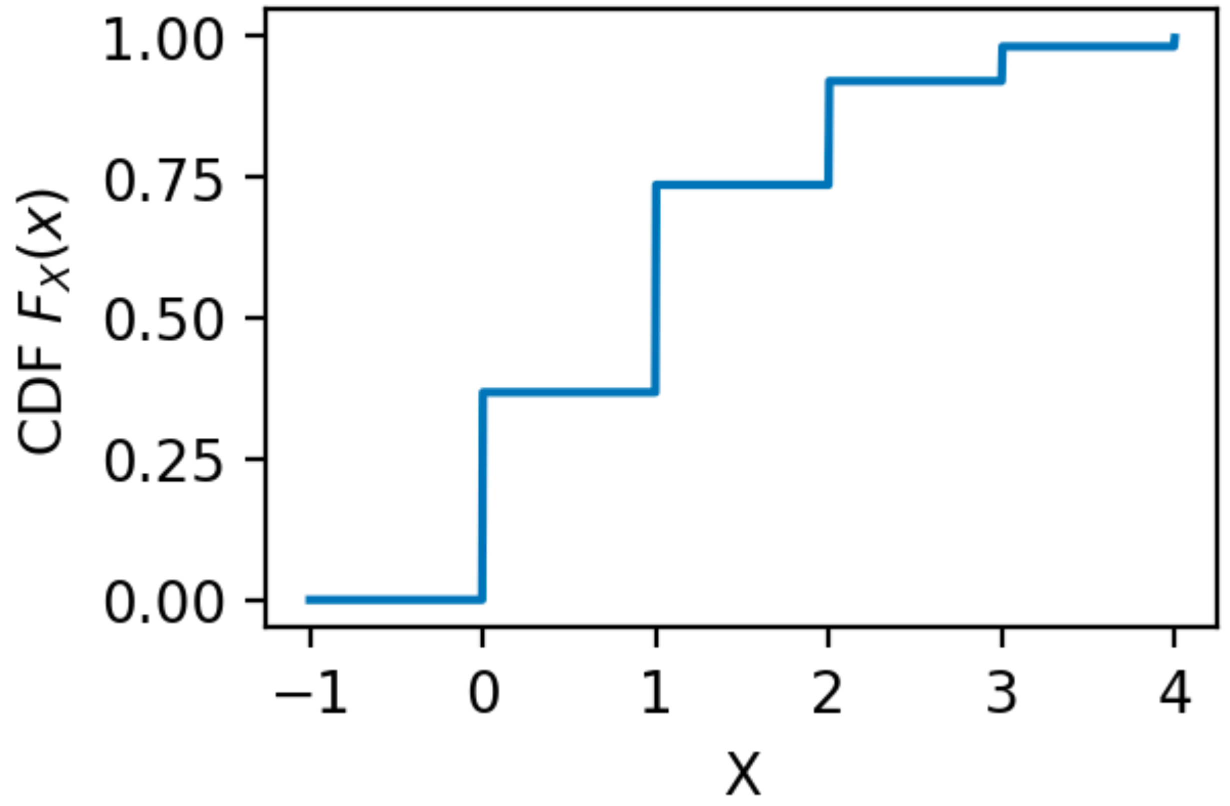
where x_i are possible discrete values (e.g., 0, 1, 2, …)

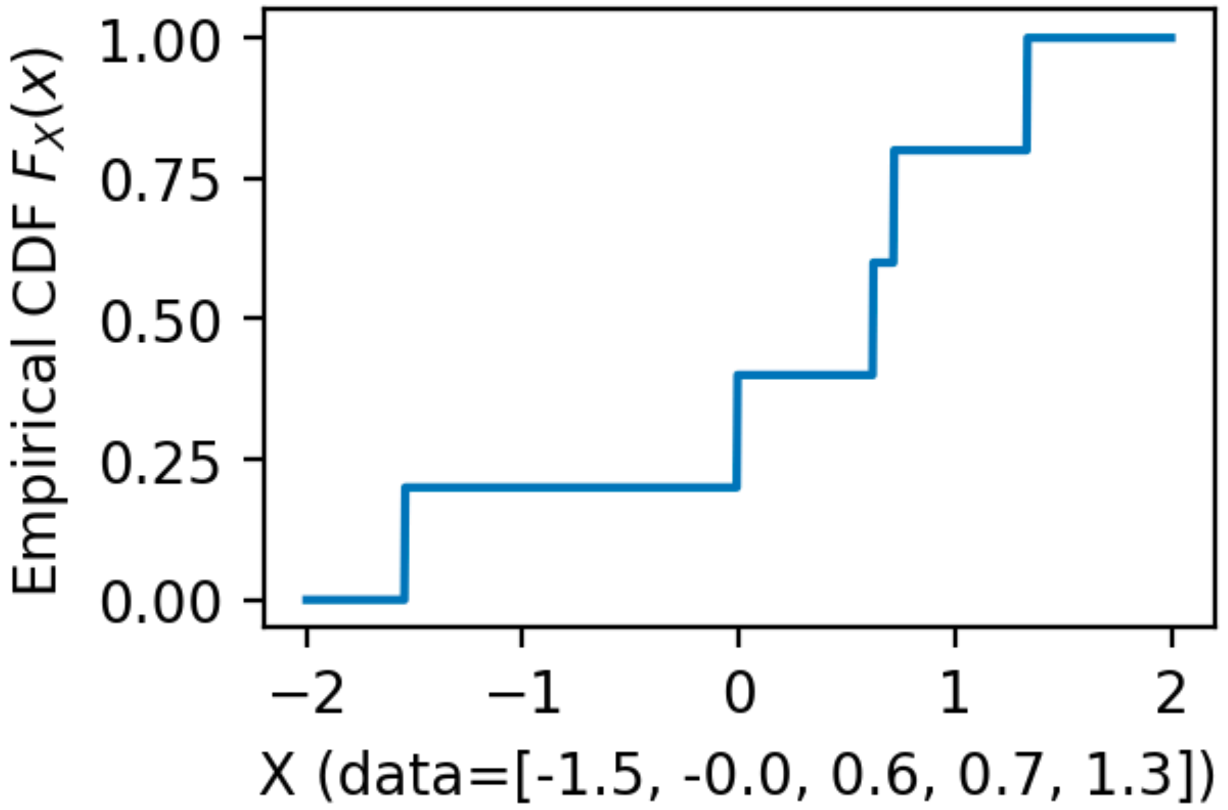
- For a dataset of n points, we can define a **discrete empirical CDF**:

$$F_X(x) = P[X \leq x] = P[-\infty \leq X \leq x] = \sum_{x_i \leq x} f_X(x_i) = \sum_{x_i \leq x} \frac{1}{n}$$

where x_i are the samples (e.g., height in feet 5.8, 6.1, 5.1, …)

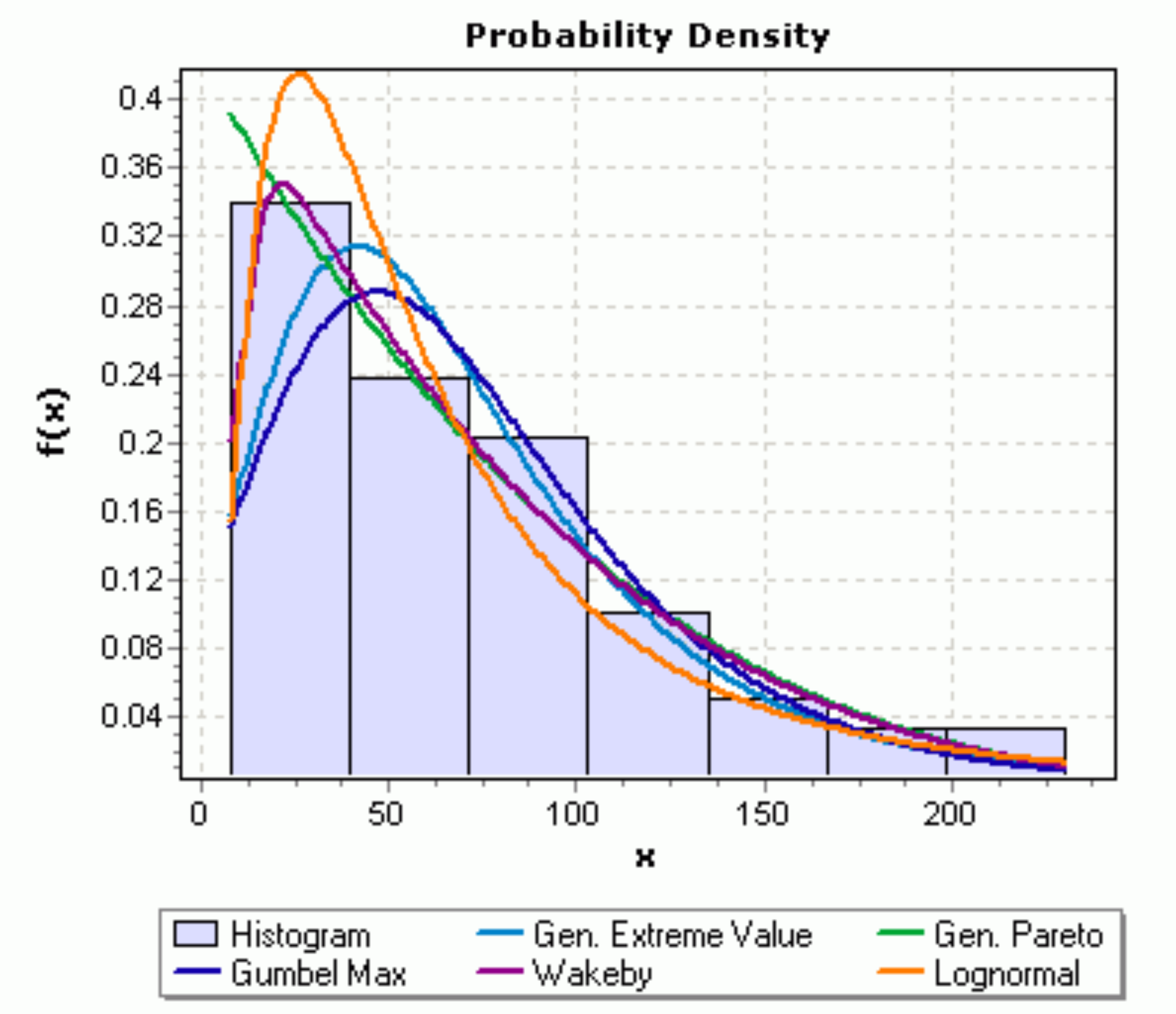






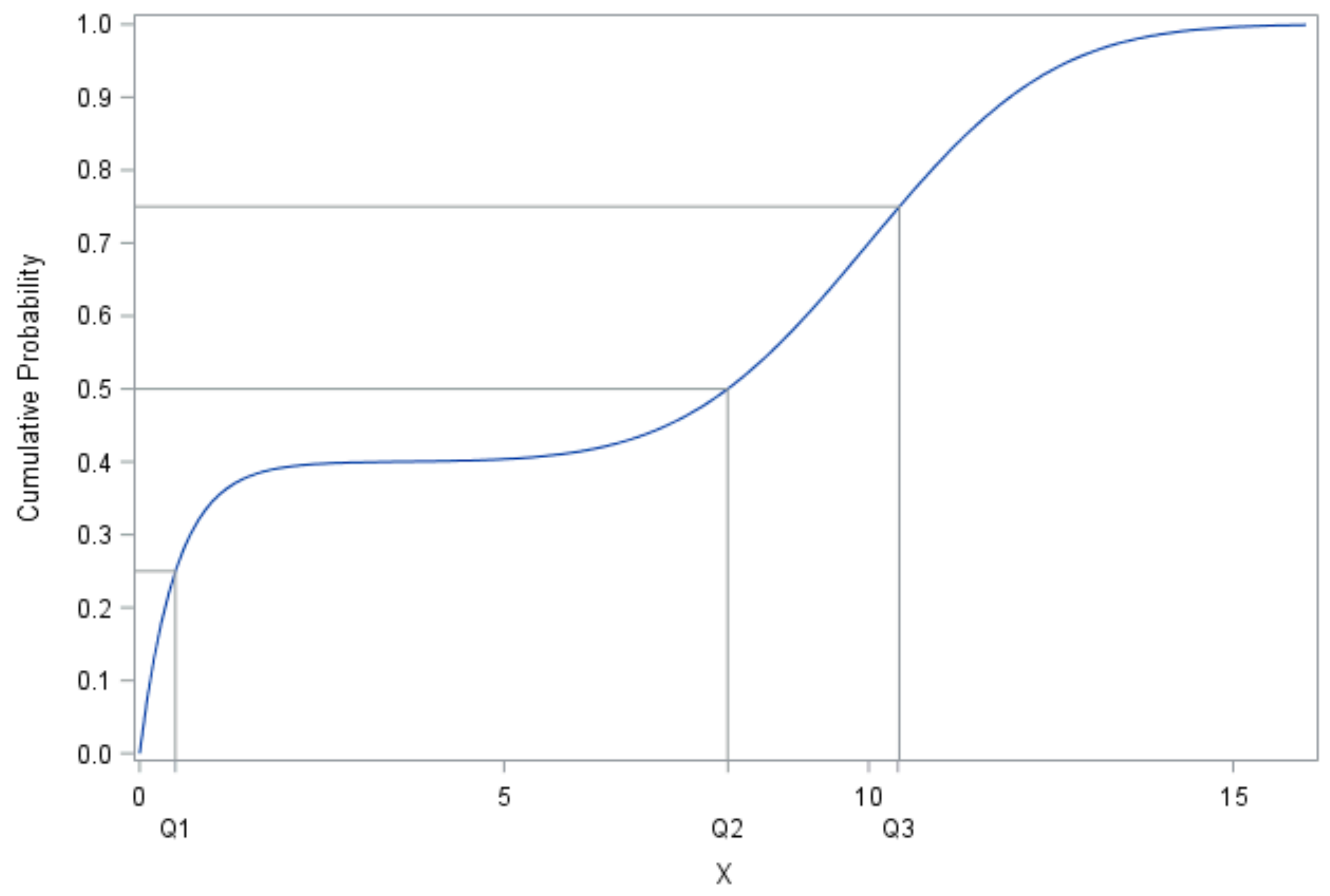
picking a distribution

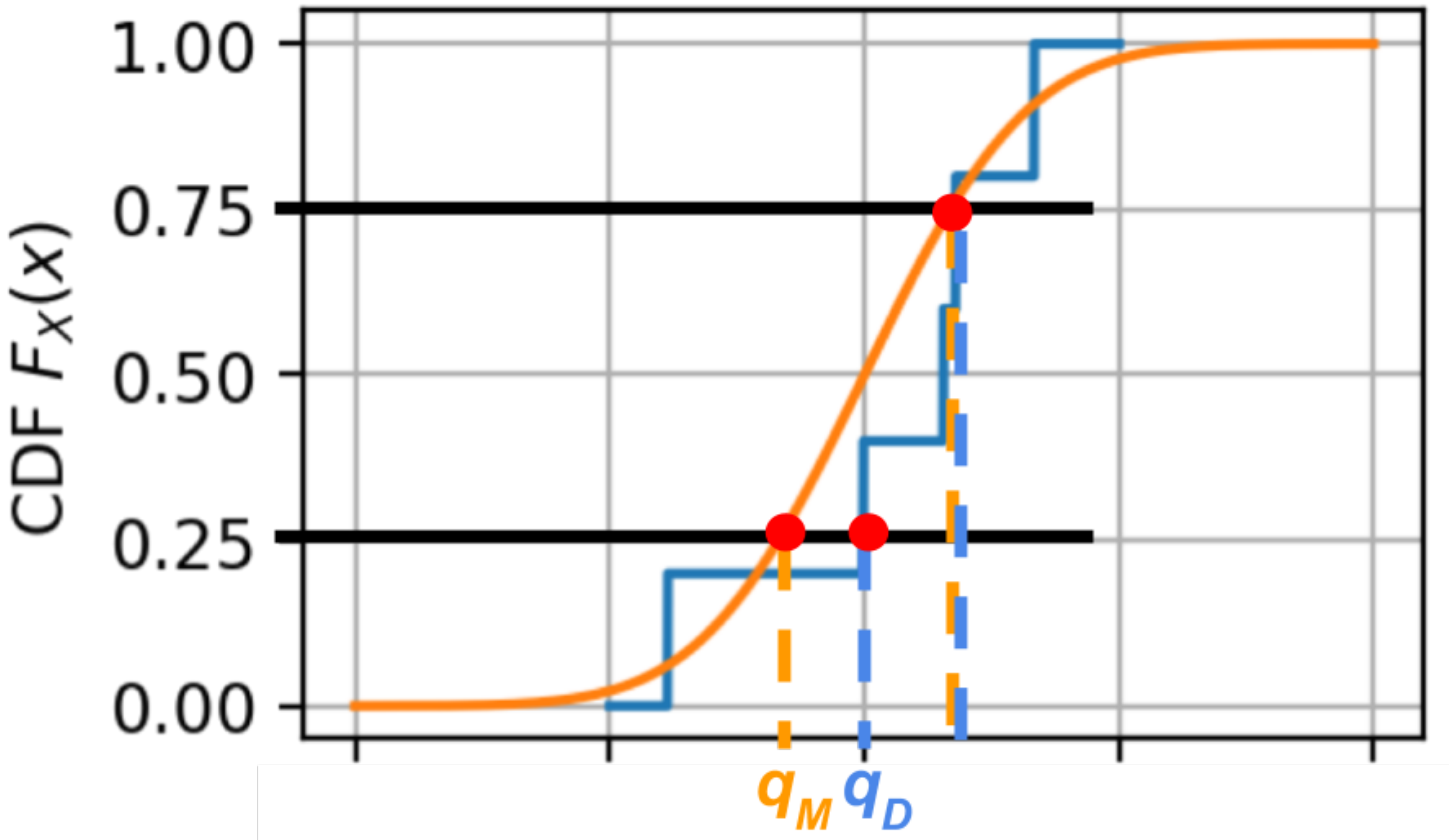
- Common problem in data science
- You have (empirical) data, and you need to choose how to (analytically) model it
  - What distribution is your data coming from?
  - What distribution is most likely to predict future samples?
- Important choice because distribution often determines how your model works



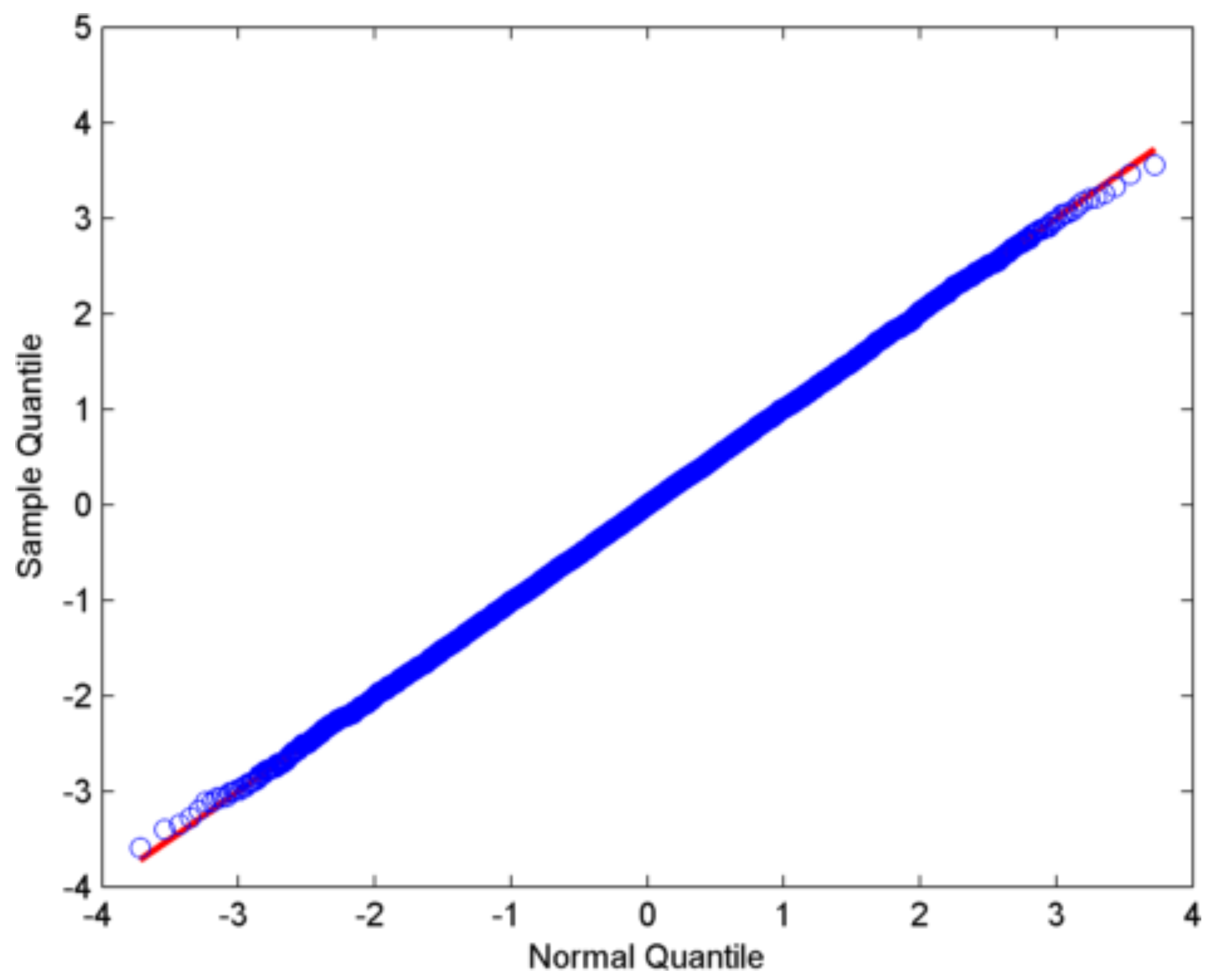
qq plots

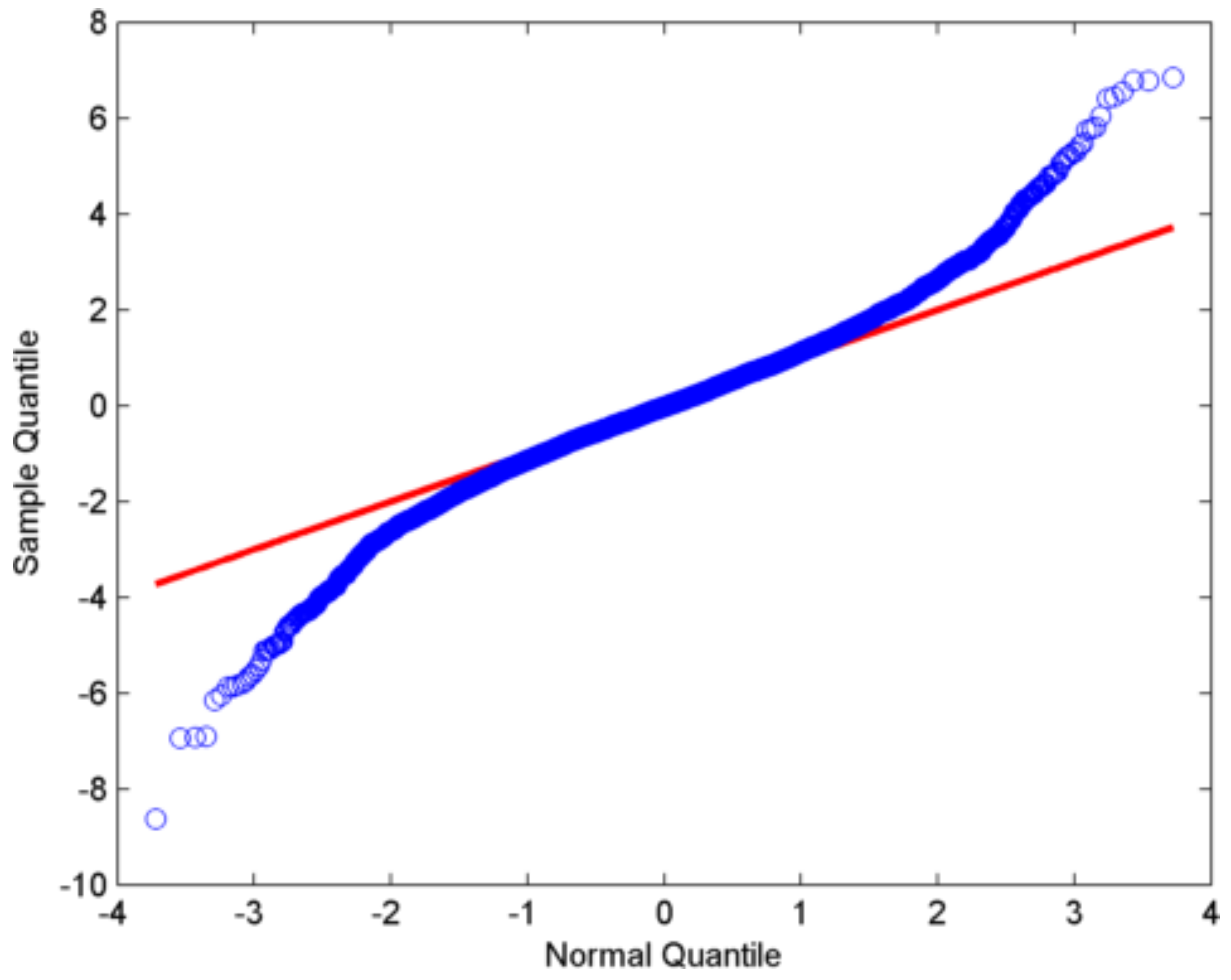
- Basic idea: Compare the CDF of your data to the CDF of a proposed model
- Use **quantiles** to do this
  - Quantile q is the value of x such that $P[X \leq x] = q$
  - Sometimes expressed in terms of **percentiles**, e.g., scoring in the 95th percentile on a test
- For each datapoint in your sample, find:
  - The quantile with respect to the dataset, q_D
  - The quantile with respect to the model, q_M
- Add each point (q_M, q_D) to a scatter plot
  - If the distributions are similar, the quartiles will appear to form the line $y = x$





qq plots





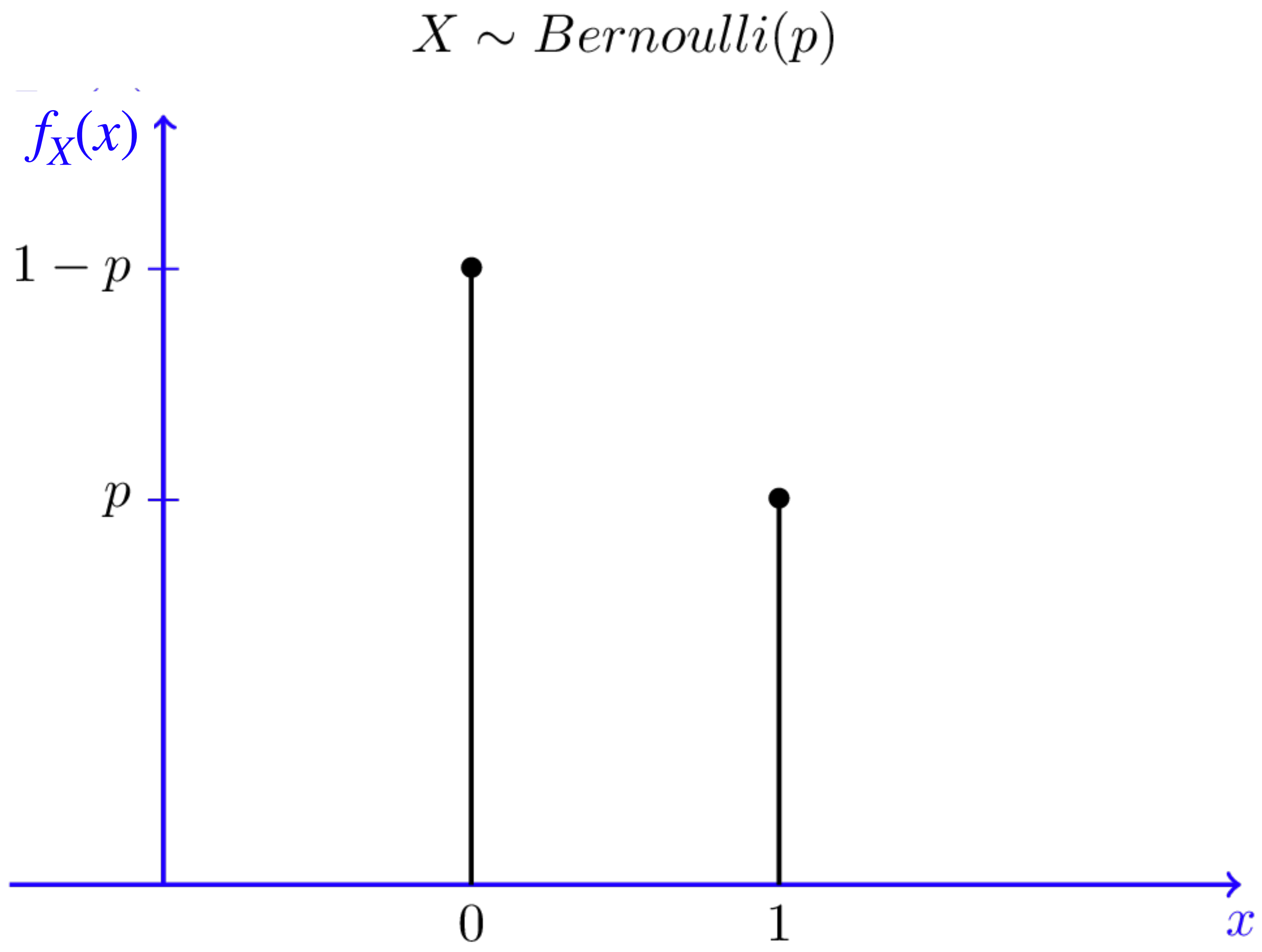
- See `scipy.stats.probplot`

bernoulli distribution

- Two states: $X = 0$ or $X = 1$
  - Think flipping a coin, or a single "bit" of information
  - But it doesn't have to be a fair coin!
- PMF:

$$P[X = x] = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \end{cases}$$

  - Here, $p \in [0,1]$ is the probability of "success" (i.e., $X = 1$)

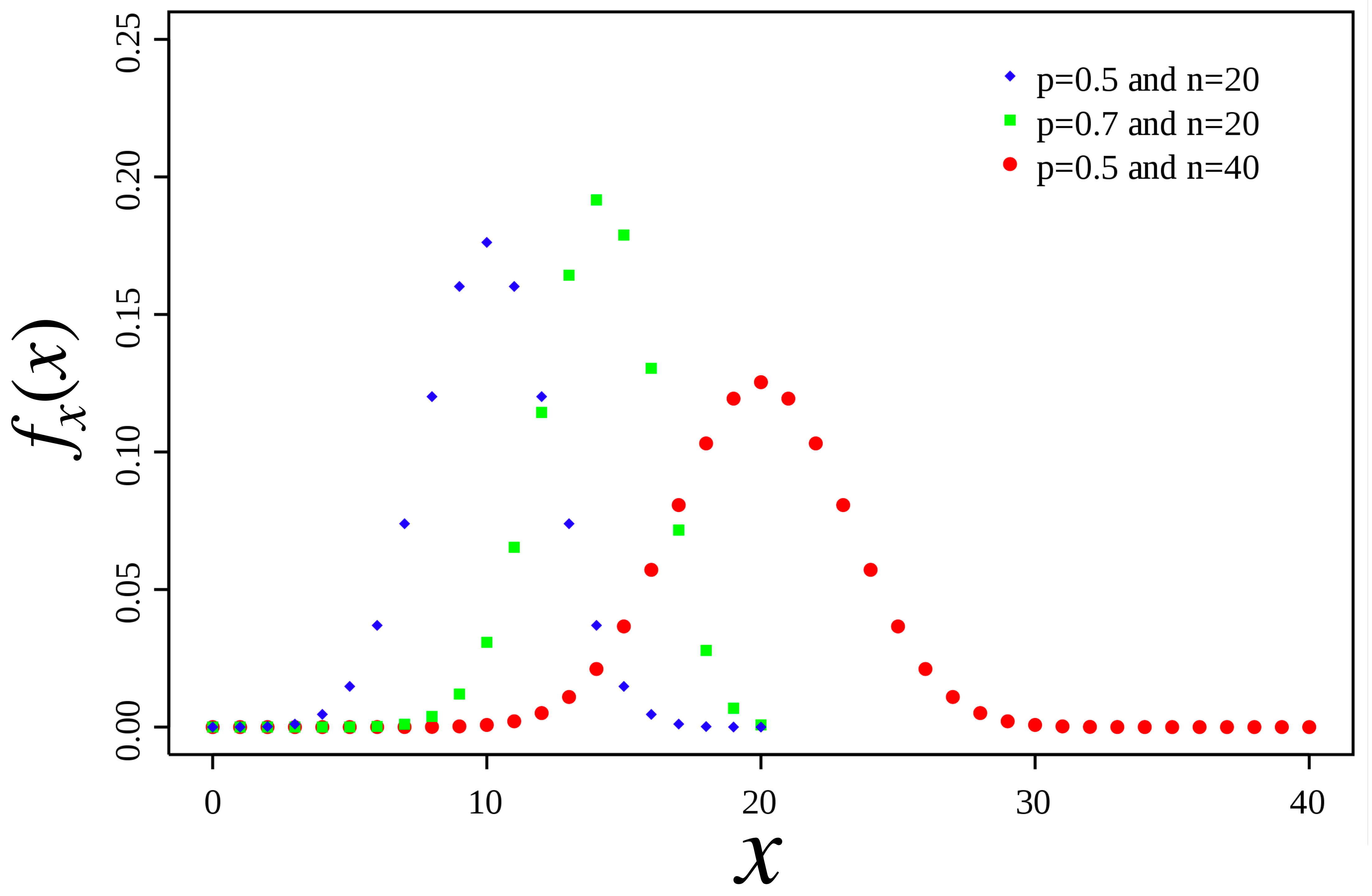


binomial distribution

- Bernoulli trials repeated n times
  - Think flipping a coin n times and counting the number of heads, or transmitting n bits and counting the number of 1's
- PMF:

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

- Here, $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ is the binomial coefficient



example

We are interested in modeling whether a machine produces outputs in spec or not. We collect 200 samples and find 20 are out of spec. Model the next output as a random variable. What is its density function?

example

Let $X = 0$ denote "out of spec" and $X = 1$ denote "in spec".

X is a Bernoulli random variable, and from the data, we can estimate $p = 180/200 = 0.9$ as the probability of success.

Hence,

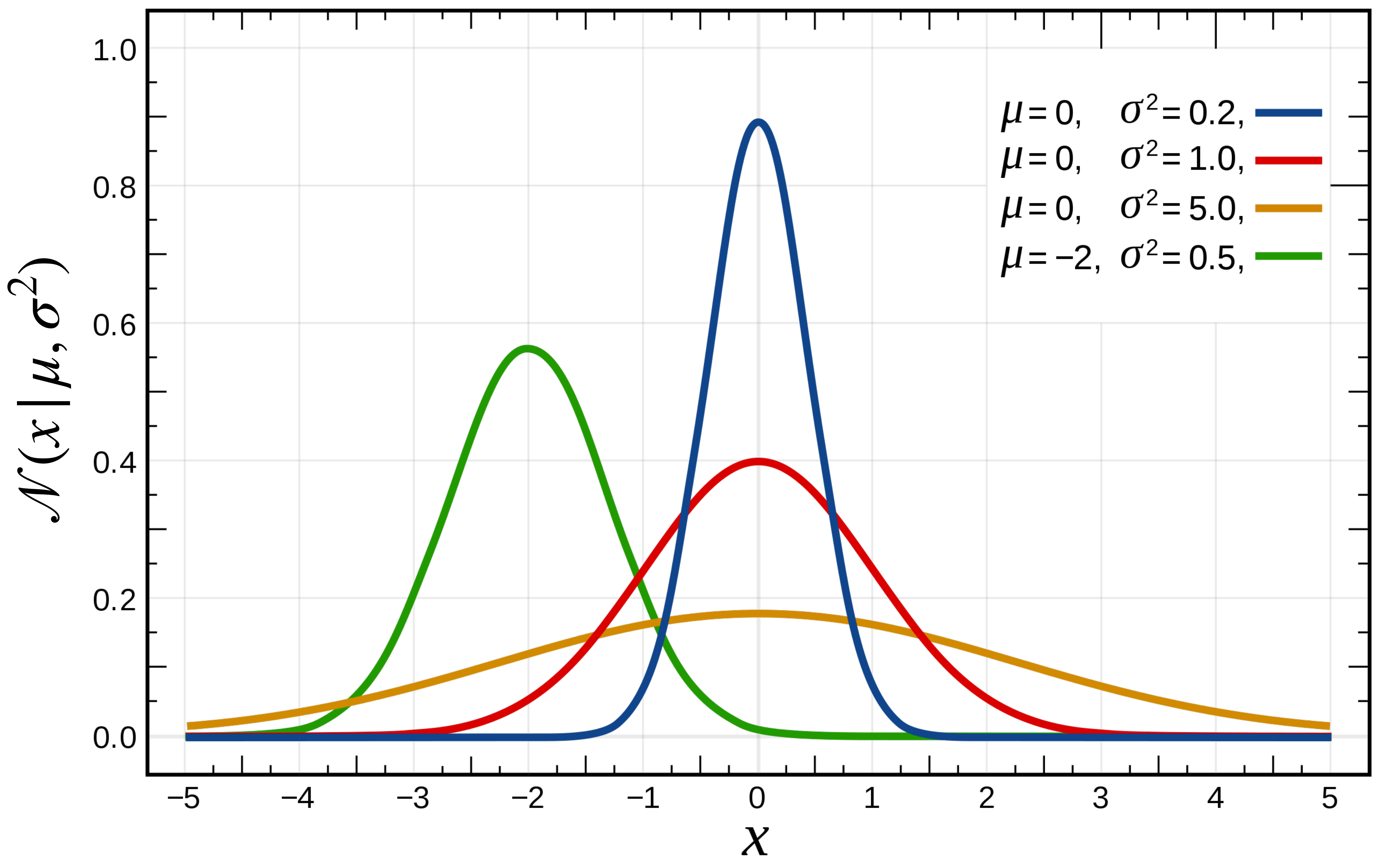
$$f_X(x) = \begin{cases} 0.1, & x = 0 \\ 0.9, & x = 1 \end{cases} \qquad F_X(x) = \begin{cases} 0, & x < 0 \\ 0.1, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

gaussian distribution

- Also called the **normal** distribution, or the bell curve
  - Very common distribution in natural processes
  - The sum of many independent processes is often normal (more on this later)
- PDF:

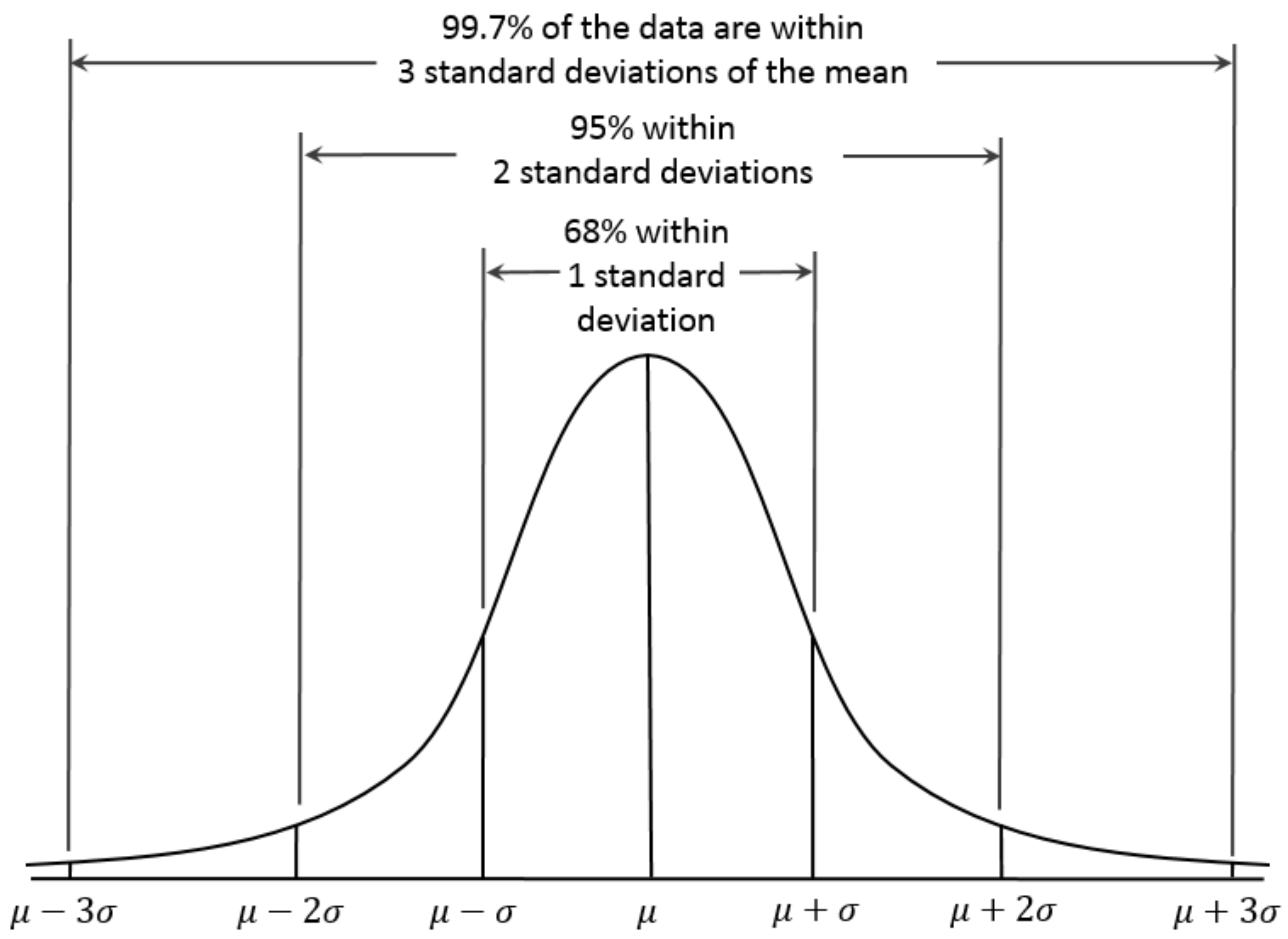
$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - Its parameters are the **mean** μ and the **variance** σ^2



gaussian distribution

- The PDF of the normal distribution has several useful properties
- The **3-sigma rule**
  - ~68% of points within $\pm\sigma$ of μ
  - ~95% of points within $\pm 2\sigma$ of μ
  - ~99.7% of points within $\pm 3\sigma$ of μ
- Useful in constructing confidence intervals and hypothesis testing (more on this later)

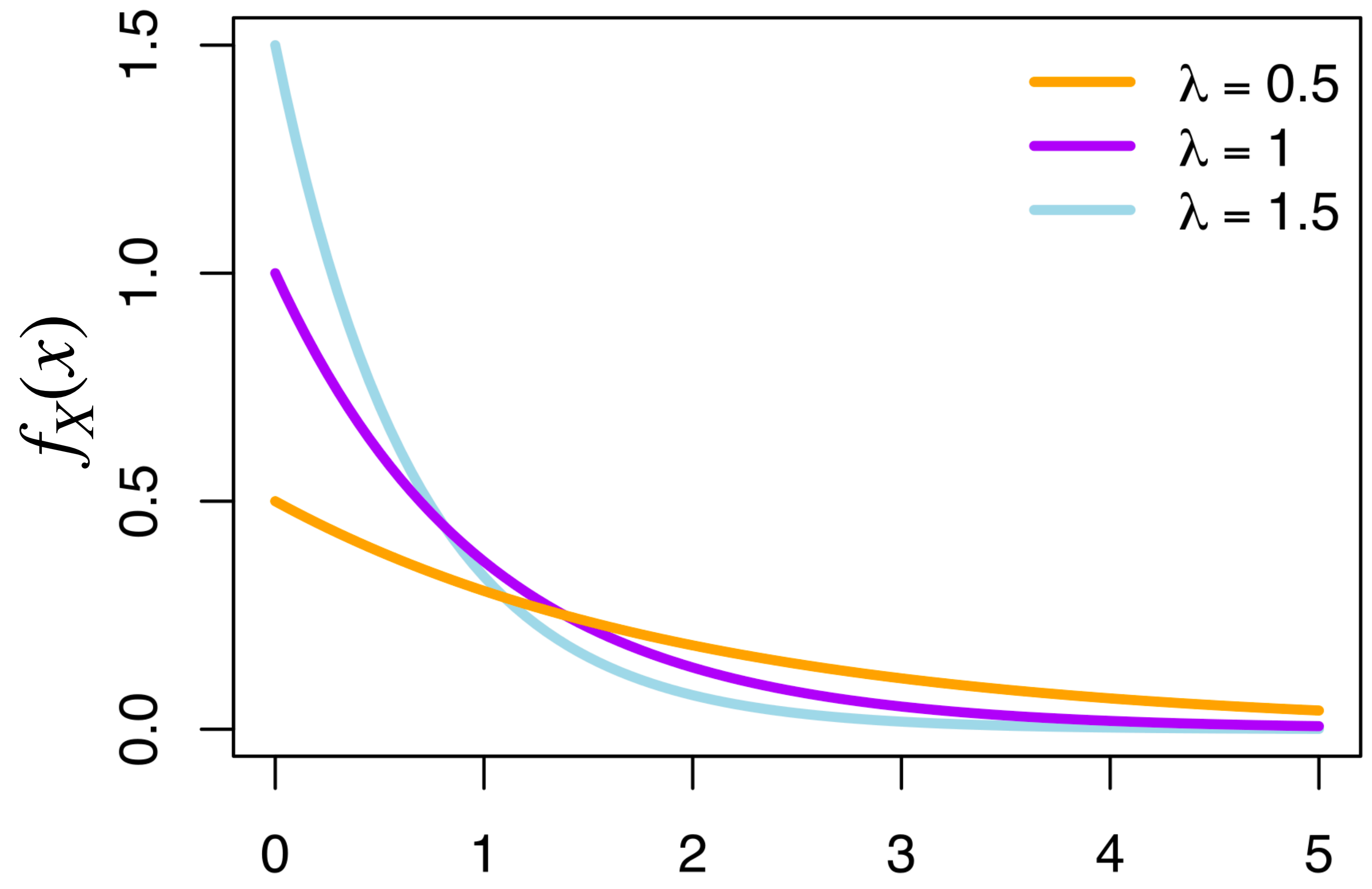


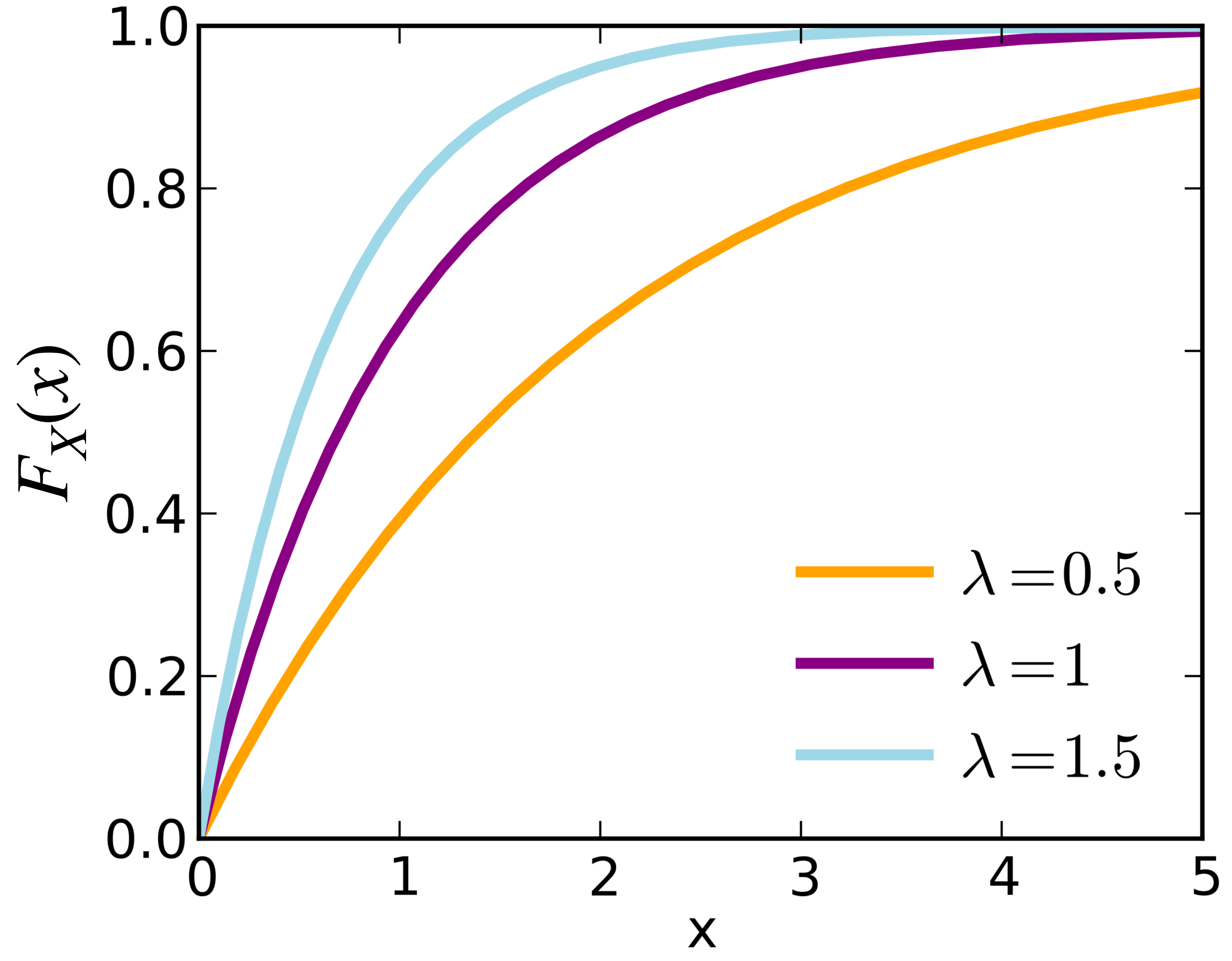
exponential distribution

- Useful for modeling decay processes, inter-arrival times, and occurrences of events
  - Probability of a radioactive item decaying
  - Time between arrival of visitors to a website, or customers to a store
- PDF:

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

  - $\lambda > 0$ is the **rate parameter**





example

We are told that the time between visits to a website, measured in minutes, is exponentially distributed with a rate parameter $\lambda = 2$. Find the CDF of this random variable. What is the probability that that there is more than 0.5 minutes between visits?

example

The random variable X has the following PDF:

$$f_X(x) = \begin{cases} 0, & x < 0 \\ 2e^{-2x}, & x \geq 0 \end{cases}$$

We can find the CDF as:

$$F_X(x) = \int_{-\infty}^{x} f_x(t)dt = \int_{0}^{x} 2e^{-2t}dt = e^{-2t}\Big|_0^x = \begin{cases} 0, & x < 0 \\ 1 - e^{-2x}, & x \geq 0 \end{cases}$$

The probability of $X > 0.5$ is:

$$P[X > 0.5] = 1 - F_X(0.5) = e^{-1} = 0.368$$

many more!

- Geometric: "How many times do I need to flip a coin to get heads?"
- Uniform: Every event in an interval is equally likely
- Student's t: Behavior of normal distribution with fewer samples
- Poisson: Discrete version of the exponential distribution
- …
- See more here: https://docs.scipy.org/doc/numpy-1.14.1/reference/routines.random.html