Probability and Distribution

what is a probability?

- The measure of the likelihood that an event occurs
- A number between 0 and 1
- The higher the number, the more likely the event (informally: 0 means the event never occurs, 1 means the event always occurs)
- Example: probability of getting a head when tossing a coin:

  $P(H) = ?$

elements of a probability model

- Conduct an experiment, and the experiment has a set of possible *outcomes*
    - Sample space, or Ω
    - Each outcome has some *probability* between 0 and 1
    - Sum of probabilities of all outcomes = 1
- An *event* is a set of possible outcomes
    - Probability of an event is the sum of the probabilities of the individual outcomes

what does a probability mean?

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

what does a probability mean?

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

a picture of a probability model

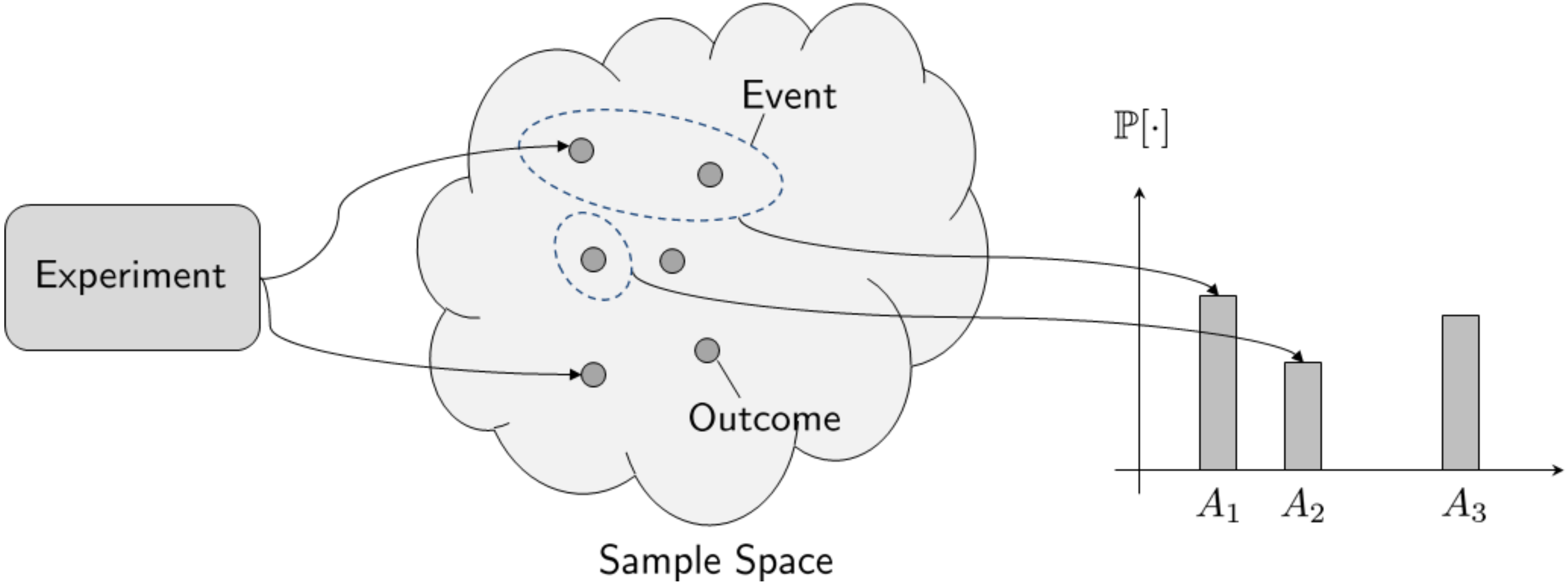
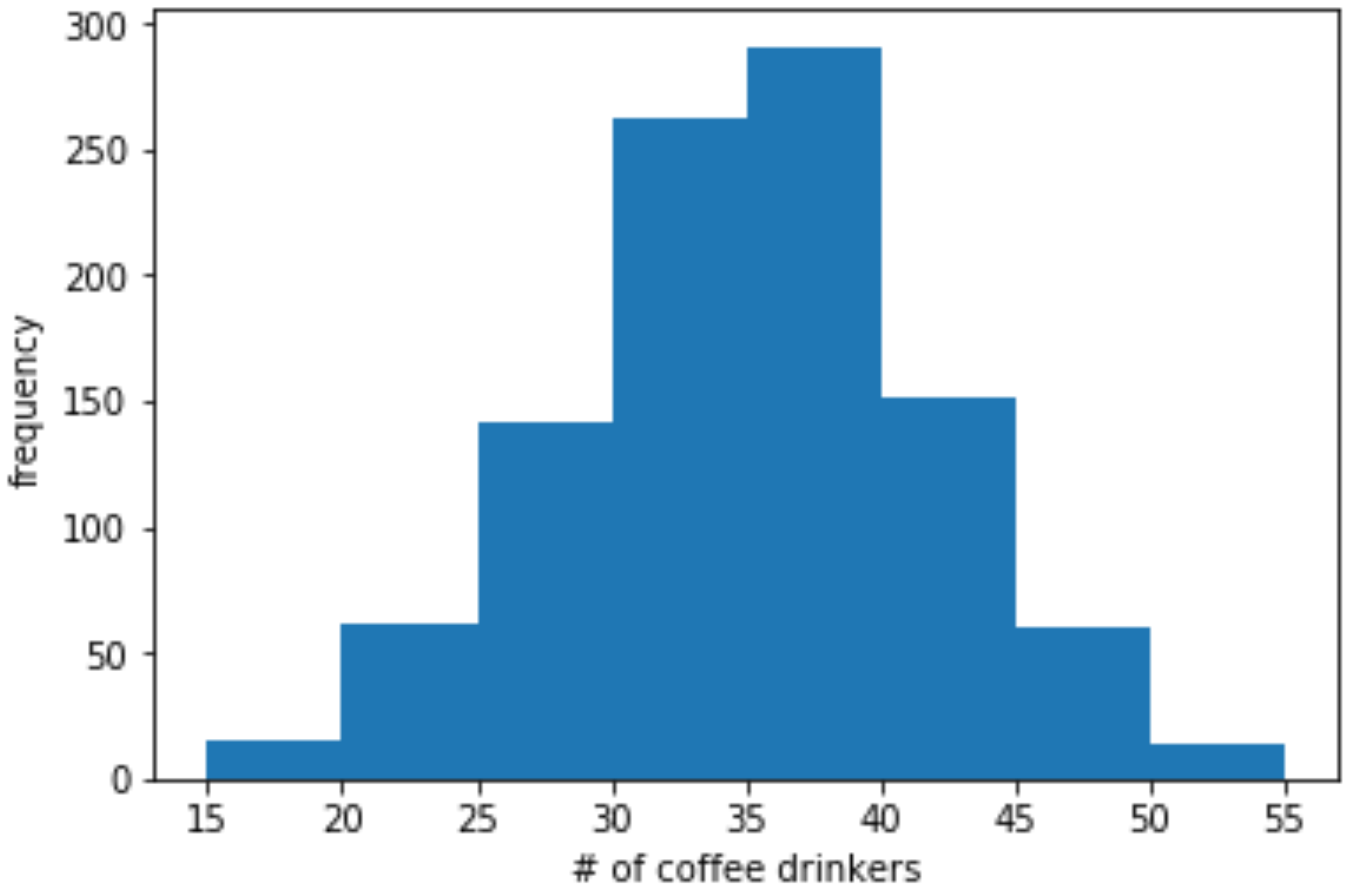
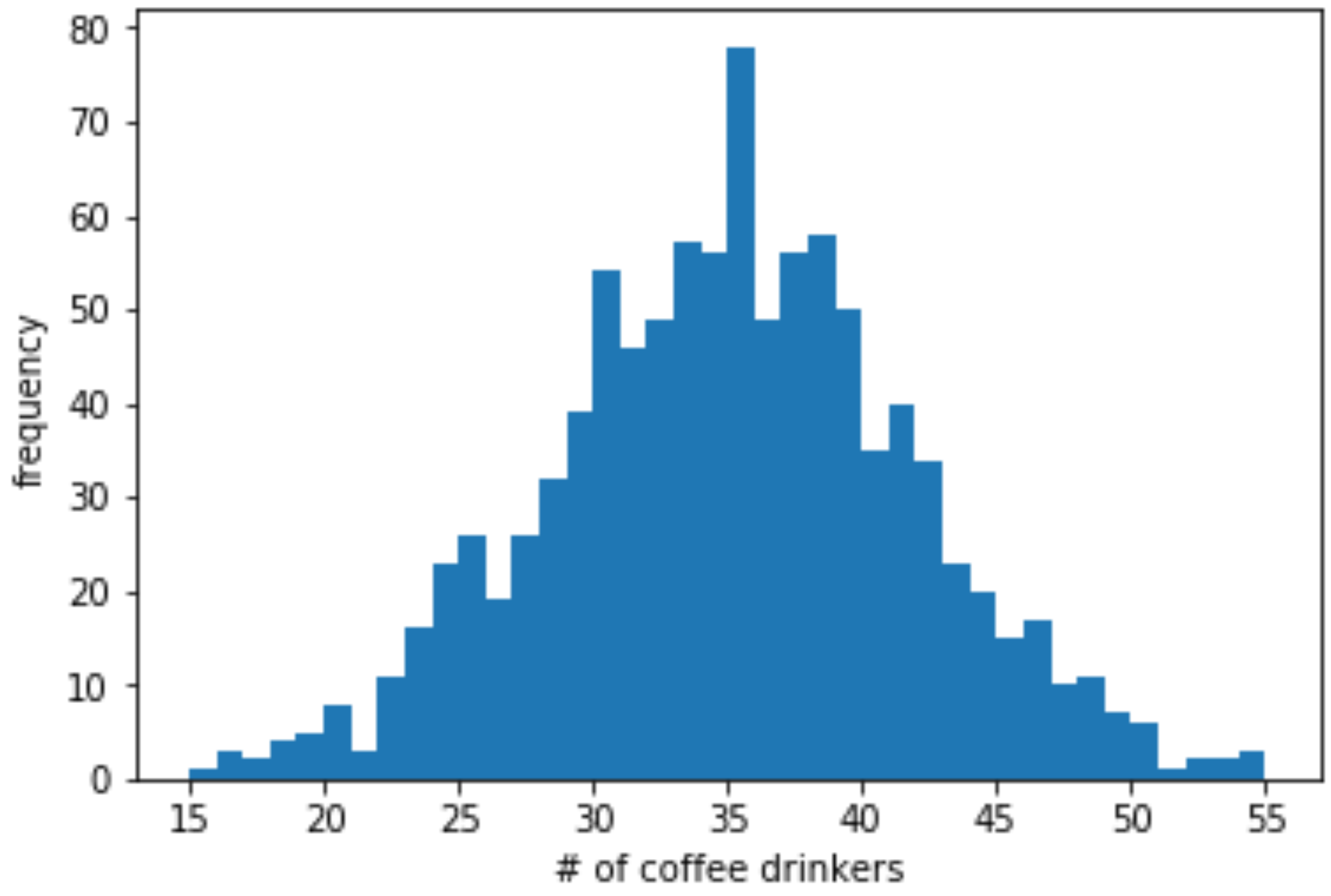


Figure courtesy Stanley Chan

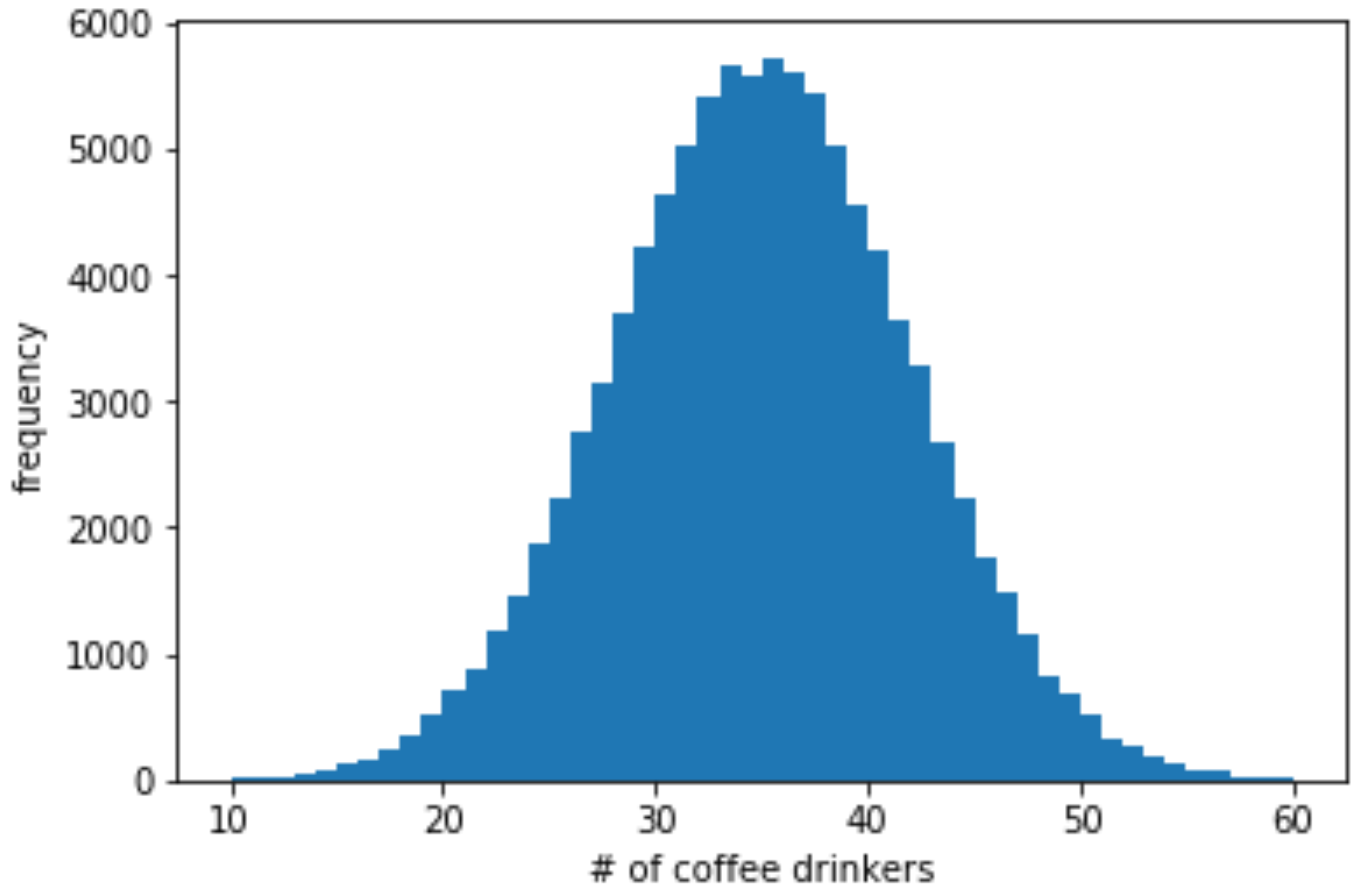
probability density function



probability density function



probability density function



probability density function

- One loose definition:
  - A histogram when the number of data points goes to infinity
- When this happens:

$$\textbf{height of bar} = \frac{\textbf{number of times } x_j \textbf{ happens}}{\textbf{number of trials}} = P[X = x_j]$$

- X is a *random variable*

random variables

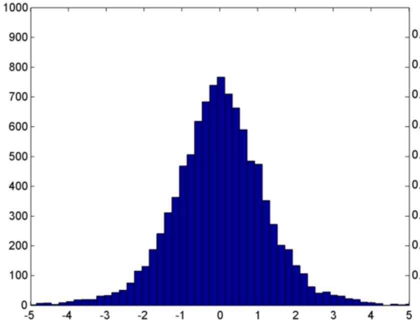
- A **random variable** *X* is a function that maps an outcome to a number
  - Just a way of letting us treat outcomes, that may not be numbers, in a mathematical way
  - e.g., *X* maps Heads to 0 and Tails to 1
- A random variable has a probability distribution which tells us the probability of its values
  - $Pr(X = 0) = Pr(X = 1) = 0.5$
- Informal intuition: the random variable *is* the histogram: the x-value of each bar is the number that outcome is mapped to, the probability of the bar is the probability of that value

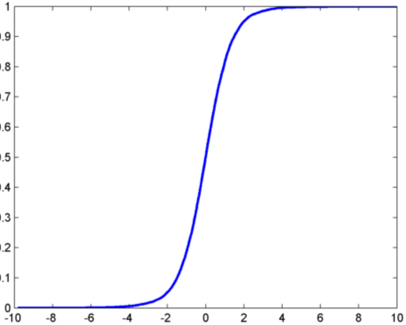
Cumulative Distribution Function

Definition

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

$$F_X(x) \stackrel{\text{def}}{=} \mathbb{P}[X \leq x]$$





Probability Density Function

Theorem

The **probability density function** *(PDF) is the derivative of the cumulative distribution function (CDF):*

$$p_X(x) = \frac{dF_X(x)}{dx}, \tag{1}$$

if F_X is differentiable at x.

If F_X is not differentiable at x, then $p_X(x)$ is defined as

$$p_X(x) = F_X(x) - \lim_{h \to 0} F_X(x - h). \tag{2}$$

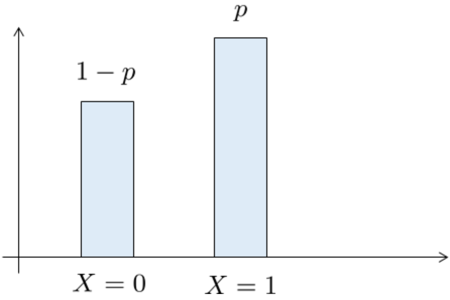
The resulting $p_X(x)$ is called the **probability mass function** (PMF).

Examples of Probability Distributions

Examples

Bernoulli Distribution:

- Two states: $X = 1$ or $X = 0$.
- Flip a coin.
- Probability: $p_X(0)$ or $p_X(1)$.
- We call X a **Bernoulli** random variable.



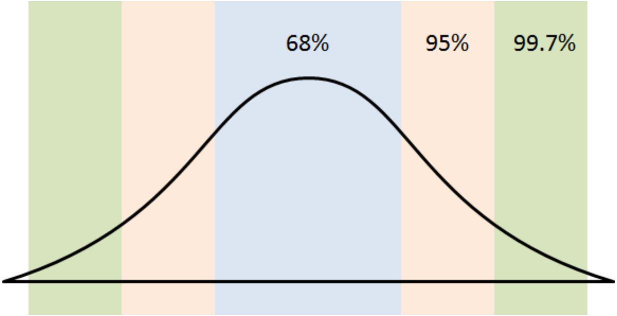
Gaussian Distribution

Also called the **Normal** distribution.

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = \mathcal{N}(x \,|\, \mu, \sigma)$$

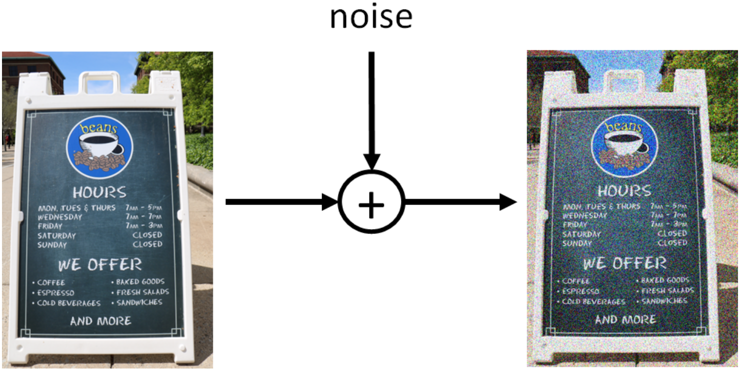
Two parameters:

- μ: mean of the Gaussian
- σ: standard deviation of the Gaussian



Example of Gaussian Distribution

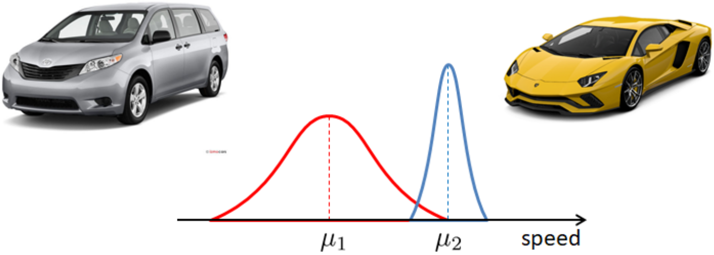
- Noise in cameras can be modeled as Gaussian
- Also Gaussian for communication systems
- A lot of natural phenomena are Gaussian



Example of Gaussian Distribution

Practical questions we can ask:

- You don't see a car, but you measure its speed
- There are only two types of cars: Mini-van and Sports car
- Given the speed, which one would you guess?



Exponential Distribution

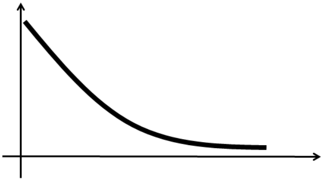
Exponential Distribution:

$$p_X(x) = \lambda \exp\{-\lambda x\}$$

- λ is the rate
- Large λ, decay faster

Usage:

- Use to model inter-arrival time
- Use to model traffic
- Use to model decay processes



Model Selection

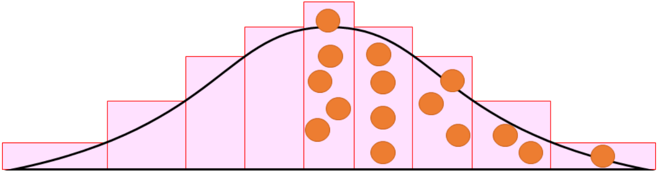
Model Selection

How to Select a Distribution?

- You have data
- A few candidate distributions
- How to choose?

Main Idea

Trick: Divide the candidate distribution into equal **area** bins

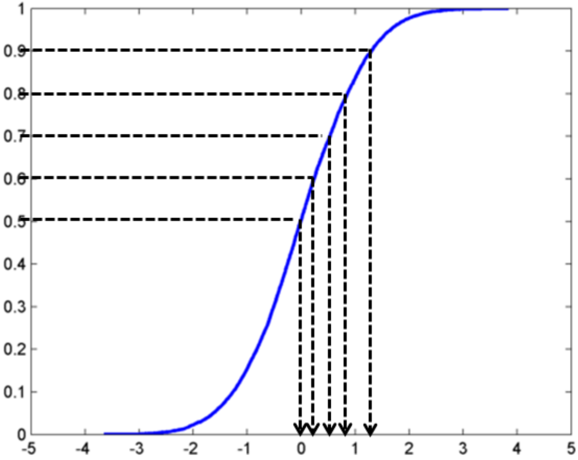


Two sets of numbers:

- Ideal area for each bin
- Actual number of samples fall into each bin

Practice

In practice, you can do this in the CDF domain.



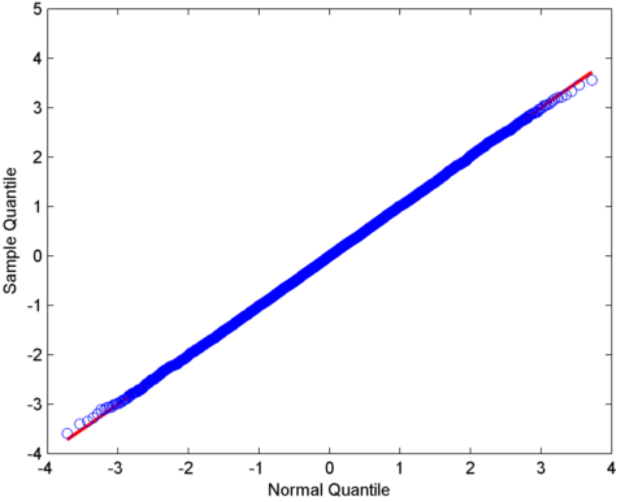
- Equally space cut the probability (y-axis)
- Find correspondingly the value (x-axis)

Algorithm

- Given a set of N data points: $x_1, \ldots, x_N$.
- Sort the numbers as $x_{[1]}, \ldots, x_{[N]}$.
- Create a dummy set $v_1, \ldots, v_N$.
- These v_i's are equally spaced in the range $[0, 1]$.
- Look at your candidate CDF, say $F_Z(z)$.
- Compute $z_i = F_Z^{-1}(v_i)$.
- Plot $x_{[i]}$ VS z_i.

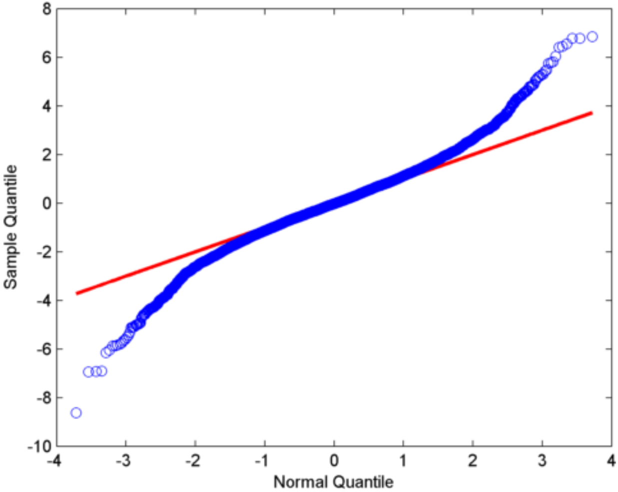
QQ-Plot

- If straight line, then actual fits ideal
- That means your candidate model is good



QQ-Plot

Bad fit:



This type of plot is called the **QQ-plot**.