CSCI 567 Discussion: Probability Review

Jan 30, 2026

What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?

What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?
- Effect of “conditioning” on an event?

What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?
- Effect of “conditioning” on an event?
- Sequences of events?

What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?
- Effect of “conditioning” on an event?
- Sequences of events?
- Dependence / Independence?

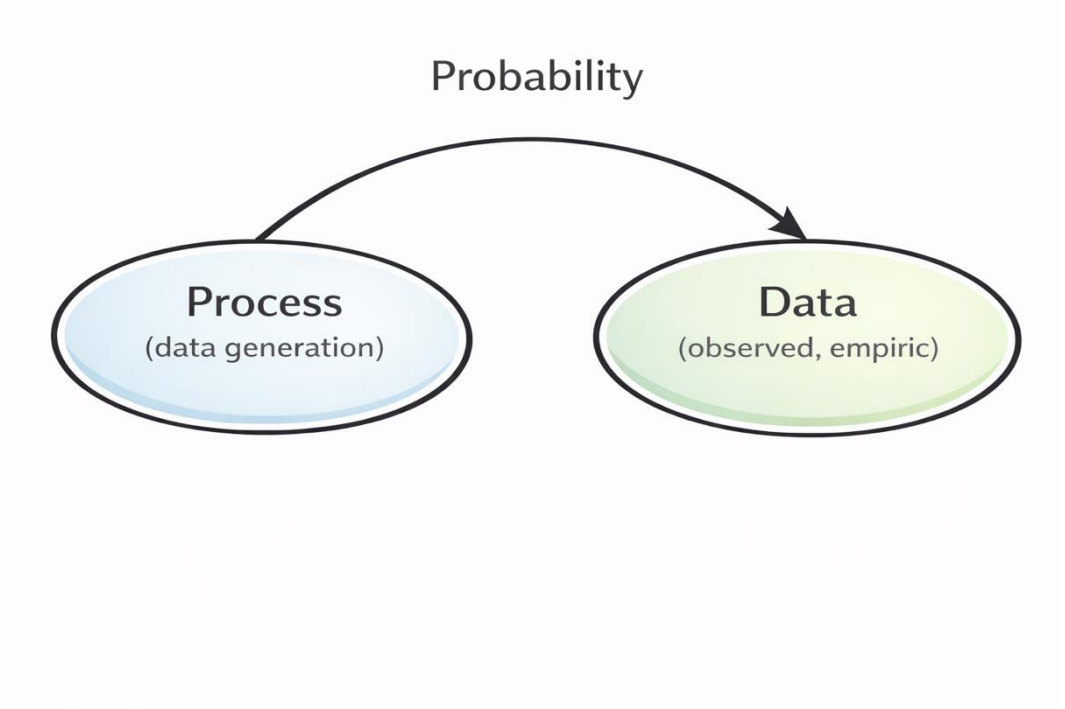
What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?
- Effect of “conditioning” on an event?
- Sequences of events?
- Dependence / Independence?



Drgoňa, Ján. (2017). Model Predictive Control with Applications in Building Thermal Comfort Control.

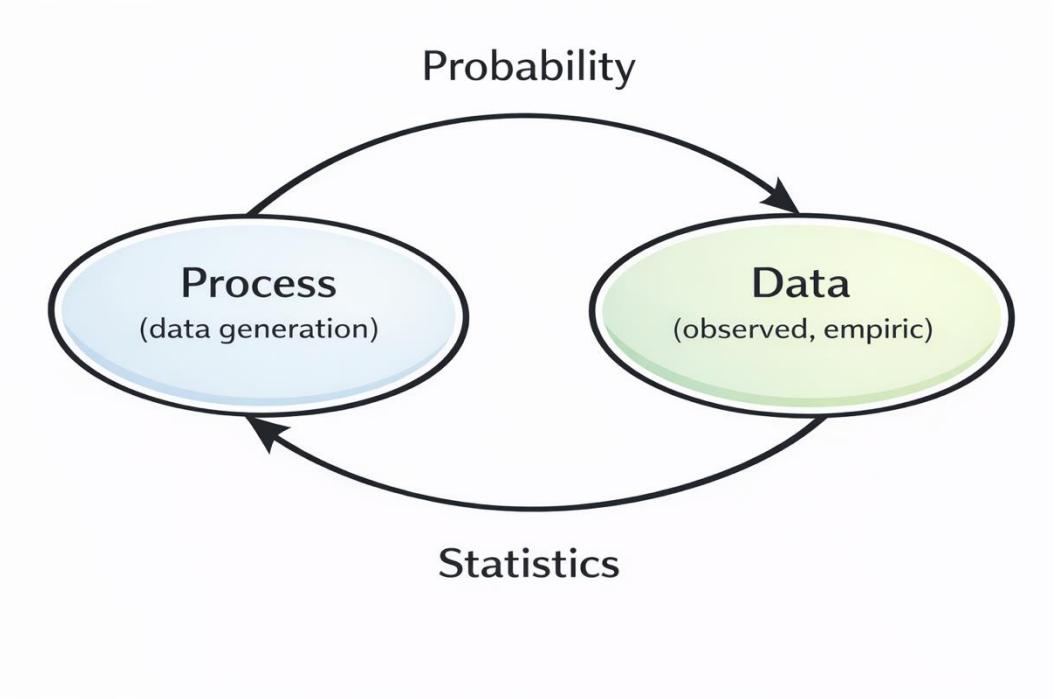
What is Probability? (versus Statistics?)

Probability studies the behavior of data-generating processes.

First, assume we have complete knowledge of how data is generated.

Then, what can we say about:

- Probabilities of interesting events?
- Effect of “conditioning” on an event?
- Sequences of events?
- Dependence / Independence?



Drgoňa, Ján. (2017). Model Predictive Control with Applications in Building Thermal Comfort Control.

What is Probability? (versus Statistics?)

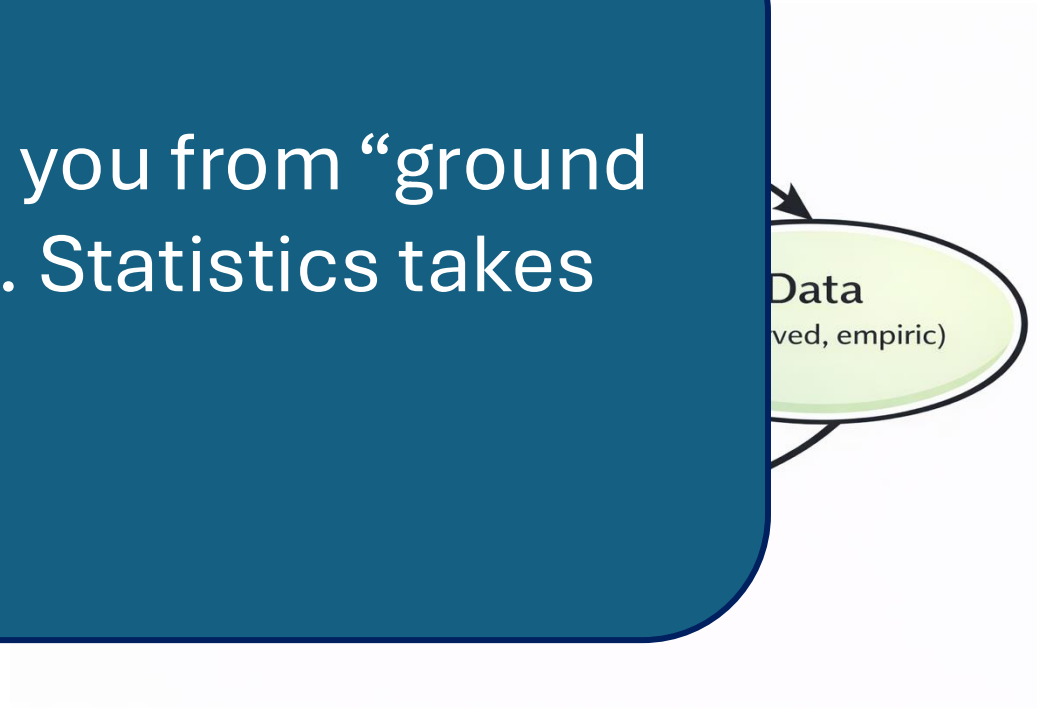
Probability studies the behavior of
data-gen

First, ass
knowled

Then, wh

- Probab
- Effect o
  event?
- Sequences of events?
- Dependence / Independence?

TLDR: Probability takes you from "ground truth" to observed data. Statistics takes you the other way!



Drgoňa, Ján. (2017). Model Predictive Control with Applications in Building Thermal Comfort Control.

Probability Spaces & Random Variables

Two key definitions underly all of probability. Don’t confuse them!

Probability space:

Random variable:

Probability Spaces & Random Variables

Two key definitions underly all of probability. Don't confuse them!

Probability space: Set Ω of events that can take place, and probabilities for each event $A \subseteq \Omega$.

- Intuition: Random events you are observing "in the wild" (e.g., flipping of coin), and nature's rules for the chances of each event.

Random variable:

Probability Spaces & Random Variables

Two key definitions underly all of probability. Don't confuse them!

Probability space: Set Ω of events that can take place, and probabilities for each event $A \subseteq \Omega$.

- Intuition: Random events you are observing "in the wild" (e.g., flipping of coin), and nature's rules for the chances of each event.

Random variable: Function $\Omega \to \mathbb{R}$, assigning number to each $\omega \in \Omega$.

- Intuition: Our choice of assigning "value" to each outcome (e.g., "Heads" → 1). Lets us use math!

Probability Space

A **probability space** is a tuple (Ω, P) where:

- Ω is a non-empty set
- P is a function $2^{\Omega} \to \mathbb{R}$

such that:

Probability Space

A **probability space** is a tuple (Ω, P) where:

- Ω is a non-empty set
- P is a function $2^{\Omega} \to \mathbb{R}$

such that:

1. $P(\Omega) = 1$
2. If $\{A\_i\}_{i \in \mathbb{N}}$ is a sequence of disjoint sets, then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Exercise

Q For events A, B and C, which of the following identities are correct?

(a) $P(A) - P(A \cap B) = P(A \cup B) - P(B)$

(b) $P(A \cup B) \leq P(A) + P(B) - P(A)P(B)$

(c) $P(A) = P(A \cap C) + P(A \cap \overline{C})$, where $\overline{C}$ denotes the complement of event C.

Random Variables & Distributions

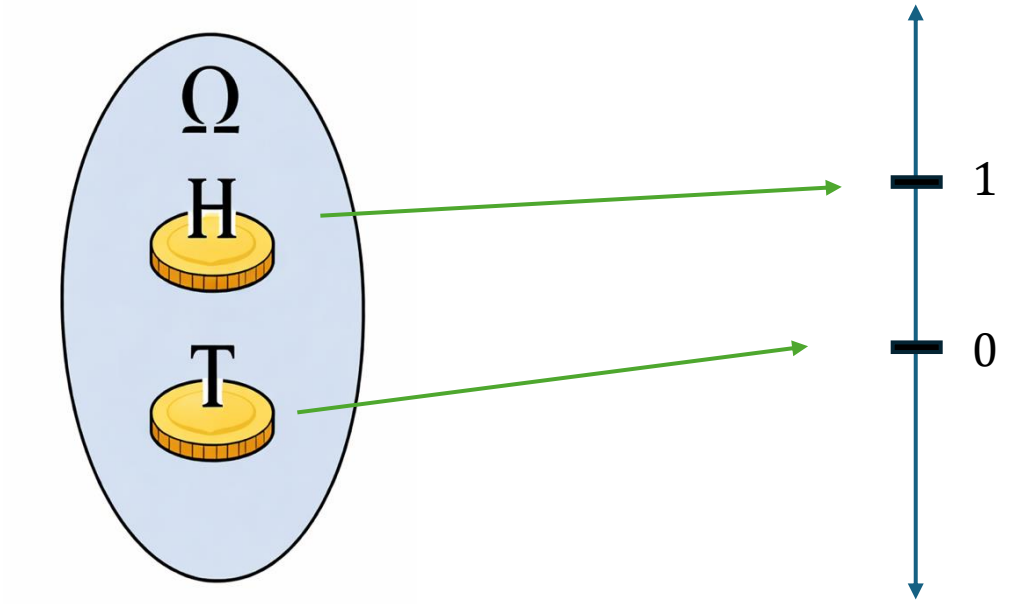
A **random variable (RV)** on a probability space (Ω, P) is a function

$$X: \Omega \to \mathbb{R}.$$

Random Variables & Distributions

A **random variable (RV)** on a probability space (Ω, P) is a function

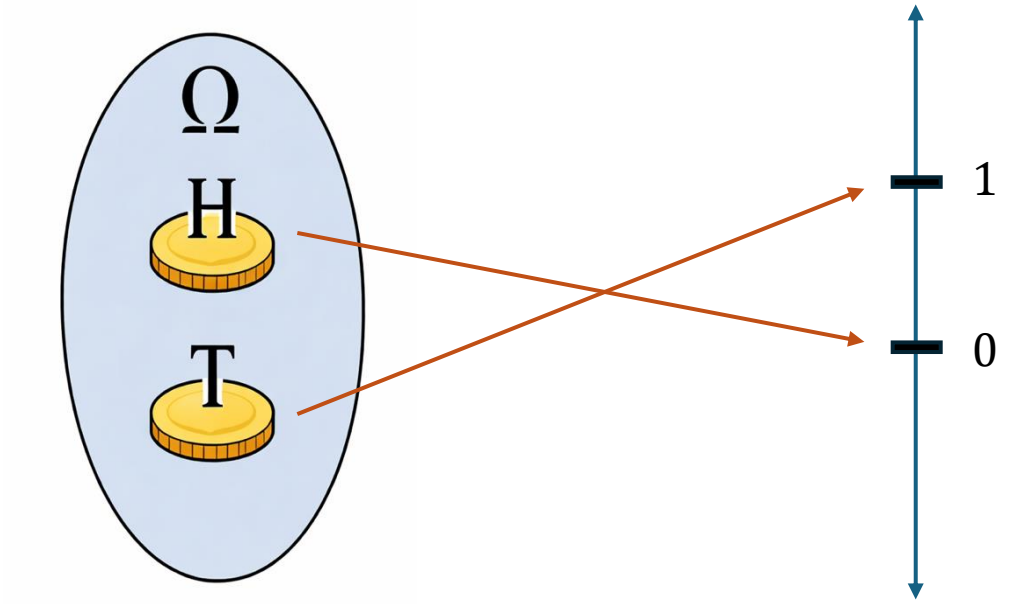
$$X : \Omega \to \mathbb{R}.$$



Random Variables & Distributions

A **random variable (RV)** on a probability space (Ω, P) is a function

$$Y : \Omega \to \mathbb{R}.$$



Random Variables & Distributions

A **random variable (RV)** on a probability space (Ω, P) is a function

$$X: \Omega \to \mathbb{R}.$$

This naturally produces probabilities over $\mathbb{R}$ via

$$P_X(A) = P(X \in A) = P(X^{-1}(A)).$$

This is called the **distribution** of X.

Random Variables & Distributions

A **random variable (RV)** on a probability space (Ω, P) is a function

$$X: \Omega \to \mathbb{R}.$$

This naturally produces probabilities over $\mathbb{R}$ via

$$P_X(A) = P(X \in A) = P(X^{-1}(A)).$$

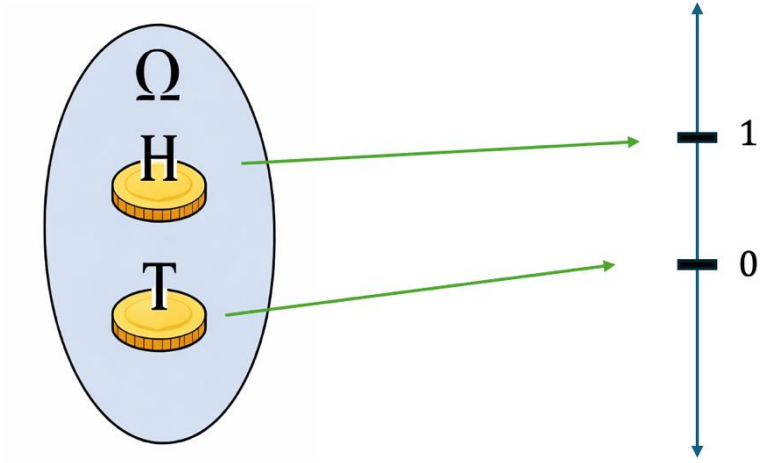
This is called the **distribution** of X.

Beware: Very different RV's can have the same distribution!

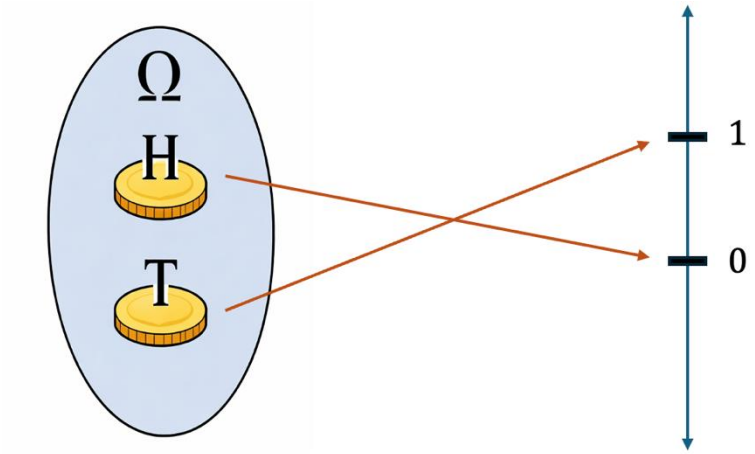
The RV contains more information than just the distribution…

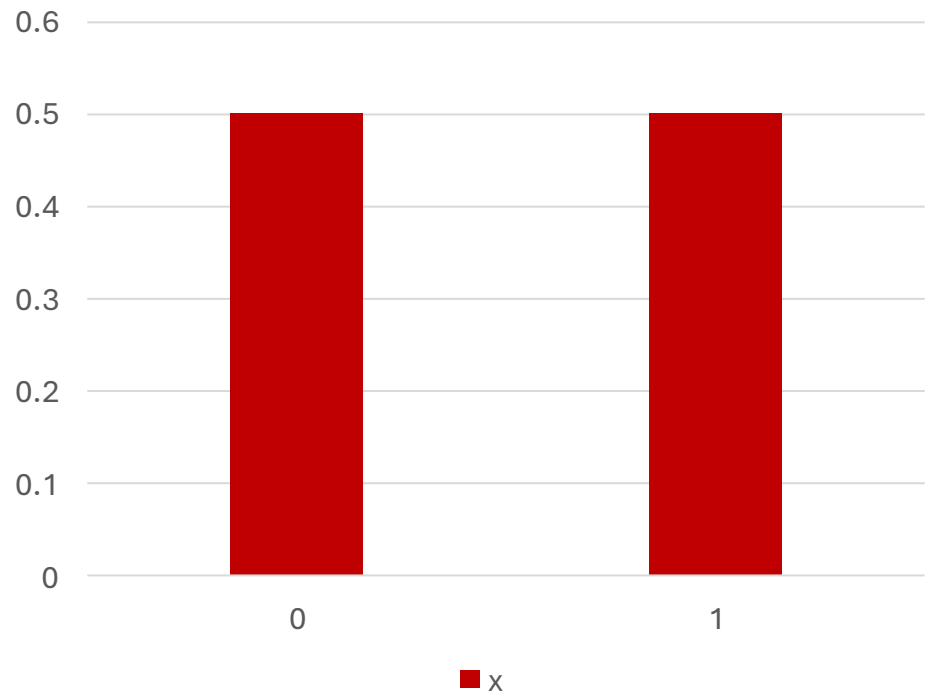
2 RV's, 1 Distribution

$X: \Omega \to \mathbb{R}$



$Y: \Omega \to \mathbb{R}$





Conditioning

For events A, B with $\mathrm{P}(B) \neq 0$, we define the **conditional probability**

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

Clearly, $P(A \mid B) \cdot P(B) = P(A \cap B)$, by multiplying out $P(B)$.

Conditioning

For events A, B with $\mathrm{P}(B) \neq 0$, we define the **conditional probability**

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

Clearly, $P(A \mid B) \cdot P(B) = P(A \cap B)$, by multiplying out $P(B)$.

This immediately yields **Bayes' rule**:

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

Simple to prove, but very valuable!

Independence

Two events A, B are **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

In this case,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A).$$

Independence

Two events A, B are **independent** if

$$P(A \cap B) = P(A) \cdot P(B).$$

In this case,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A).$$

Similarly, two random variables X and Y are **independent** if

$$P(X = x \ \wedge Y = y) = P(X = x) \cdot P(Y = y).$$

Equivalently, $P(X = x \mid Y = y) = P(X = x)$.

Exercise

Q A bag contains 2 red balls and 3 blue balls. First, Alice draws a ball from the bag randomly (and removes it from the bag). Then, Bob draws a ball randomly too. 1) What is the probability that Alice gets a red ball and Bob gets a blue ball? 2) What is the probability that Alice gets a blue ball given that Bob gets a blue ball?

Exercise

Q For events A, B, C and $Z_1, \ldots, Z_T$, which of the following identities are correct?

(a) $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

(b) $\frac{P(A|B,C)}{P(A|C)} = \frac{P(B|A,C)}{P(B|C)}$

(c) $P(\bigcap_{t=1}^{T} Z_t) = \Pi_{t=1}^{T} P(Z_t)$

(d) $P(\bigcap_{t=1}^{T} Z_t) = \Pi_{t=1}^{T} P(Z_t|Z_1, \ldots, Z_{t-1})$

Exercise

Q: When training a machine learning model, why isn't it a good idea to evaluate your model's performance on the training set?

After all, for any given predictor/classifier f, its error on a dataset S is an unbiased estimate of its true error?

Exercise

Q: When training a machine learning model, why isn't it a good idea to evaluate your model's performance on the training set?

After all, for any given predictor/classifier f, its error on a dataset S is an unbiased estimate of its true error?

A: But it stops being unbiased when we condition upon the fact that f was trained on S!

E.g., Imagine 10,000 people flip a fair coin 10 times. On average, each person's performance is reflective of their true performance.

But the ~10 people who flipped all Heads are unlikely to repeat this!

Exercise

Q: When idea to evalu

After is an unbia

A: Bu t f was t

E.g., ach perso

But the is!

TLDR: You must condition on all available information! Not allowed to cherry-pick and ignore (un)favorable information!

Understanding Train/Test Split

How *should* we evaluate the performance of a trained model?

Key idea: Use a separate, “fresh” test dataset!

Understanding Train/Test Split

How *should* we evaluate the performance of a trained model?

Key idea: Use a separate, “fresh” test dataset!

Workflow:

1. Sample all data, S
2. Uniformly at random, split S into S_{train} and S_{test} (say, 80%:20%)

Understanding Train/Test Split

How *should* we evaluate the performance of a trained model?

Key idea: Use a separate, "fresh" test dataset!

Workflow:

1. Sample all data, S
2. Uniformly at random, split S into S_{train} and S_{test} (say, 80%:20%)
3. Train your model by minimizing loss on S_{train},

$$\hat{f} \approx \arg\min_{f \in F} \hat{R}_{S_{\text{train}}}(f) = \arg\min_{f \in F} \frac{1}{n} \sum_{i \leq n} \ell(f(x_i), y_i).$$

Understanding Train/Test Split

How *should* we evaluate the performance of a trained model?

Key idea: Use a separate, "fresh" test dataset!

Workflow:

1. Sample all data, S
2. Uniformly at random, split S into S_{train} and S_{test} (say, 80%:20%)
3. Train your model by minimizing loss on S_{train},

$$\hat{f} \approx \arg\min_{f \in F} \hat{R}_{S_{\text{train}}}(f) = \arg\min_{f \in F} \frac{1}{n} \sum_{i \le n} \ell(f(x_i), y_i).$$

4. Evaluate your model via its performance on S_{test}, i.e.,

$$\hat{R}_{S_{\text{test}}}(f) = \sum_{j \le m} \ell\left(f\left(x_j\right), y_j\right)$$

Refresher: Discrete vs. Continuous RV

Recall that a **discrete** random variable X places all its mass on a countable set S, i.e.,

$$P(X \in S) = 1.$$

In this case, the distribution of X is determined by the *probability mass function* (**PMF**) p_X giving the mass of each point:

$$p_X(x) = P(X = x).$$

Refresher: Discrete vs. Continuous RV

Recall that a **discrete** random variable X places all its mass on a countable set S, i.e.,

$$P(X \in S) = 1.$$

In this case, the distribution of X is determined by the *probability mass function* (**PMF**) p_X giving the mass of each point:

$$p_X(x) = P(X = x).$$

A **continuous** random variable has a probability density function (**PDF**) $f_X(x)$ and

$$P(X \in [a, b]) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Expectation & Variance

Random variables contain a lot of information! We can often summarize/analyze using 2 key values: *expectation* and *variance*.

The **expectation**, or mean, of a random variable X is

$$\mathbb{E}[X] = \sum_x x \cdot p_X(x), \quad \text{or} \quad \mathbb{E}[X] = \int_x x \cdot f_X(x)\mathrm{d}x.$$

Intuitively, the average value of X – a measure of "center."

Expectation & Variance

Random variables contain a lot of information! We can often summarize/analyze using 2 key values: *expectation* and *variance*.

The **expectation**, or mean, of a random variable X is

$$\mathbb{E}[X] = \sum_x x \cdot p_X(x), \quad \text{or} \quad \mathbb{E}[X] = \int_x x \cdot f_X(x)\mathrm{d}x.$$

Intuitively, the average value of X – a measure of "center."

The **variance** of a random variable X with mean μ is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Intuitively, measures the "dispersion" of X, how greatly it varies.

Properties of Expectation and Variance

Linearity of Expectation

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
   - Holds for **any** random variables X, Y – regardless of dependence!

2. $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$
   - Can always pull out constants

Properties of Expectation and Variance

Linearity of Expectation

1. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
   - Holds for **any** random variables X, Y – regardless of dependence!
2. $\mathbb{E}[c \cdot X] = c \cdot \mathbb{E}[X]$
   - Can always pull out constants

Variance Properties

1. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, **if** X and Y are independent!
   - Can fail without independence!
2. $\mathrm{Var}(cX) = c^2 \cdot \mathrm{Var}(X)$
   - Intuition: Variance uses "squared units." Square root to recover **std deviation**

Gaussian Distributions

Recall one-dimensional **Gaussian** distributions:

$$X \sim N(\mu, \sigma), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Gaussian Distributions

Recall one-dimensional **Gaussian** distributions:

$$X \sim N(\mu, \sigma), \qquad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Equivalently, a random variable X is Gaussian if

$$X = a\,Z + b, \qquad Z \sim N(0, 1).$$

Key idea: **Gaussian = affine function of a standard normal**

Multivariate Gaussian Distributions

A random vector $X \in \mathbb{R}^d$ is
multivariate Gaussian if

$$X = A\,Z + \mu, \qquad Z \sim N(0, I_\ell),$$

where:

- $Z \in \mathbb{R}^\ell$ is a vector of ℓ independent standard normals,
- $A \in \mathbb{R}^{d \times \ell}$,
- $\mu \in \mathbb{R}^d$.

Multivariate Gaussian Distributions

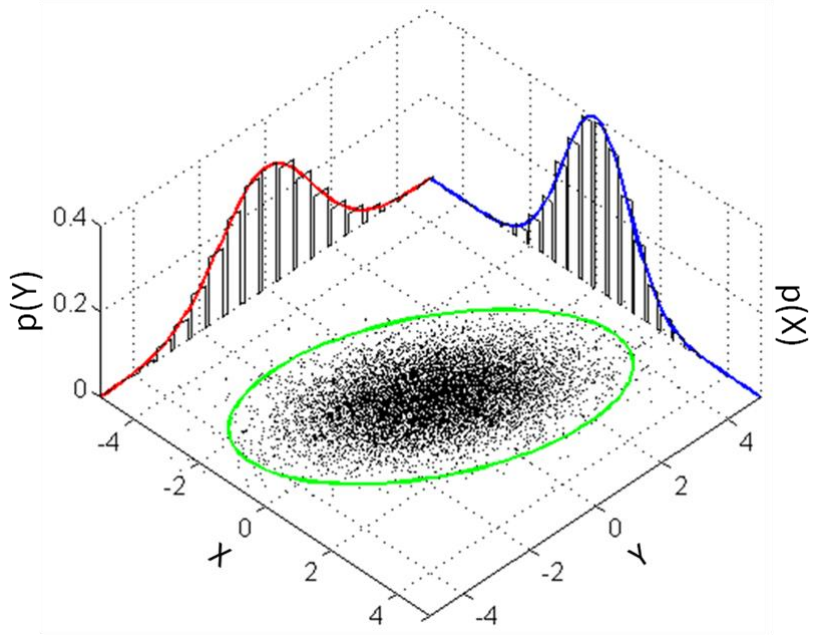
A random vector $X \in \mathbb{R}^d$ is **multivariate Gaussian** if

$$X = A\,Z + \mu, \qquad Z \sim N(0, I_\ell),$$

where:

- $Z \in \mathbb{R}^\ell$ is a vector of ℓ independent standard normals,
- $A \in \mathbb{R}^{d \times \ell}$,
- $\mu \in \mathbb{R}^d$.

Intuition: Z is a spherical Gaussian cloud. So every multivariate Gaussian is a stretched, rotated, shifted sphere.



Joint Gaussians

Random vectors X and Y are **jointly Gaussian** if $\binom{X}{Y}$ is distributed as a multivariate Guassian.

Equivalently,

$$\binom{X}{Y} = A\,Z + \mu$$

This is stronger than X and Y each being distributed Gaussian!

Key Characterization

For random variables $X_1, \ldots, X_n$, the random vector $(X_1, \ldots, X_n)$ is multivariate Gaussian **if and only if** $\sum_{i=1}^{n} a_i X_i$ is Gaussian for all $a \in \mathbb{R}^n$.

Exercise

Q Which of the following statements are true?

(a) Suppose X and Y are two jointly Gaussian random variables. Then $Z = X - 2Y$ is also Gaussian.

(b) Suppose X and Y are two jointly Gaussian random variables. Then the marginal distribution of X is also Gaussian.

(c) Suppose X and Y are two jointly Gaussian random variables. Then the conditional distribution of X given Y is also Gaussian.

Exercise

Q Suppose your spam classification software gives the guarantee that (1) if an email is spam, it will mark it as spam with probability 90%, (1) if an email is not spam, it will only mark it as spam with probability 10%. Suppose you know that 1% of all your emails are spam. If your spam classification software marks a certain email as spam, what is the probability that it is actually spam?