Lecture 4: Probability theory, Bayes theorem, Probability functions, Neumann-Pearson lemma	Dr. Farbin 8/28/24

Probability Theory: The branch of mathematics concerned w/ the likelihood of events and their statistics.

Useful in:

- Experiments - distinguish/describe observations
  $9.8\ m/s^2 \pm 0.01\ m/s^2$
- Theory - theoretical predictions or models
- Quantifying Uncertainty - understanding sources of uncertainty
  1) Stochastic processes: underlying randomness
     - Quantum Mechanics
     - Shuffling cards
  2) Incomplete observations & Unknowns
     - Monty hall problem
     - When does a heart attack occur

Frequentist Approach: This interpretation argues that probability is the long-term frequency of an event occuring in repeated trials.

↳ Probability is viewed as a property of the physical world.

Examples include:

- Coin flip: If you flip a coin 1000 times and it lands on heads 510 times, the frequentist would say that the probability of heads based on observed frequency is $P(H) = 0.51$

- Drug effectiveness: Using clinical trials one would determine the frequency of patients that recover.

<u>Bayesian Approach:</u> This interpretation argues that probability is a degree of belief/confidence in a particular event or hypothesis based on prior knowledge and evidence, readily updated with new data.

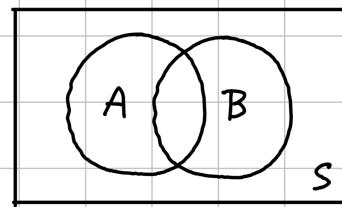
↳ Probability is viewed as a measure of belief or information, conditional on prior data/knowledge.

Examples include:

- Coin flip: A Bayesian starts with belief that $P(H) = 0.50$ indicates a fair coin. If they note $P(H) = 0.51$ after 1000 coin flips, they would argue that the coin might be biased towards heads
- Drug effectiveness: A Bayesian starts with prior belief (previous studies) about the drug's effectiveness, then update as new patient recovery data is collected.

<u>Note on the Approaches:</u> Using either result leads to the same equations of probability theory; they are just different in interpretation.

Basic Probability: Consider a set S w/ subsets A and B as shown in the figure:



By definition, the probability of choosing a point w/in S is

$$P(S) = 1$$

this implies that $\forall A \subset S$ ("for all A which is a subset of S")

$$P(A) \geq 0$$

The probability of not choosing a point w/in A is

$$P(\bar{A}) = 1 - P(A)$$

obviously

$$P(A \cup \bar{A}) = 1$$

The prob. of A or B is

$$P(A \cup B) = P(A) + P(B) - \overbrace{P(A \cap B)}^{\text{subtract the shared region to avoid double-counting}}$$

If A and B do not overlap,

$$P(A \cup B) = P(A) + P(B) \; ; \; P(A \cap B) = P(\overbrace{\emptyset}^{\text{Prob. of null set}}) = 0$$

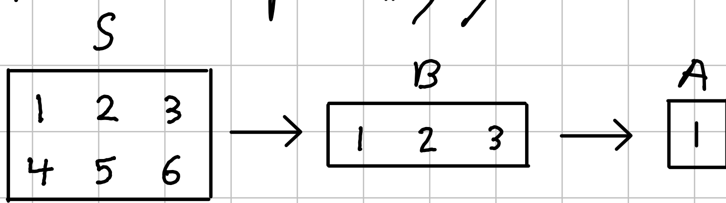
If A is a subset of B,

$$P(A) \leq P(B)$$

Conditional Probability: The probability of A given B is given by

$$P(A \underbrace{|B}_{\text{"given B"}}) = P(A \cap B) / P(B)$$

↳ Consider the example of a dice role. Given that you rolled a 3 or less, what is the probability you rolled a 1?



$$P(1 \mid 3 \text{ or less}) = P(1 \text{ and } 3 \text{ or less} - \text{Just } 1) \, / \, P(3 \text{ or less})$$

$$= 1/6 \, / \, 1/2$$

$$P(1 \mid 3 \text{ or less}) = 1/3$$

Using the formulas for $P(A|B)$ and $P(B|A)$ with the fact that $P(A \cup B) = P(B \cup A)$, we find

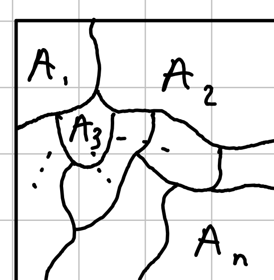
$$P(A|B) = P(B|A) \cdot P(A) \, / \, P(B)$$

If $P(A \cup B) = \emptyset = 0$, then

$$\underbrace{P(A ; B)}_{\text{"prob. of A and B together"}} = P(A) \cdot P(B)$$

The prob. of A given B now becomes

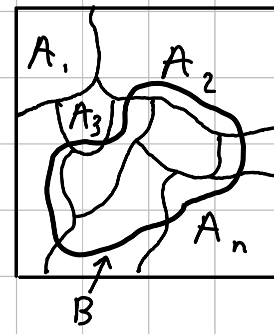
$$P(A|B) = P(A) P(B) / P(B) = P(A)$$



__Bayes Theorem:__ Consider the set given by the figure. The total set S can be viewed as the union of all subsets,

$$S = \cup_i A_i$$

Now consider the set B. We may write



$$P(B) = P(B \cap S)$$
$$= P(B \cap \cup_i A_i)$$
$$= P(\cup_i [B \cap A_i])$$

↳ Thus the prob. of B is the union of all the parts of the A_i's that B overlaps.

Using the definition of $P(B \mid A_i)$, we write

$$P(A \mid B) = P(B \mid A) \cdot [P(A) / \textstyle\sum_i P(B \mid A_i) P(A_i)]$$

This is Bayes theorem, which we will concretize with an example: Let's say there is a disease spreading and $P(\text{sick}) = 0.20$ and $P(\overline{\text{sick}}) = 0.80$. A testing kit has come out stating that if you are sick, this kit will mark you positive 98% of the time: $P(+ \mid \text{sick}) = 0.98$. They also claim that if one is not sick, the kit marks you negative 97% of the time: $P(- \mid \overline{\text{sick}}) = 0.97$. What is the probability of us being sick if we test positive?

$$A_i = \{\text{sick}, \overline{\text{sick}}\}$$

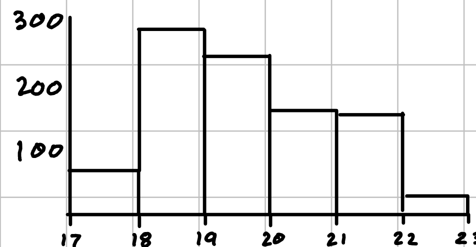
$$P(\text{sick} \mid +) = [P(+ \mid \text{sick}) P(\text{sick})] / [P(+ \mid \text{sick}) \cdot P(\text{sick}) + P(+ \mid \overline{\text{sick}}) P(\overline{\text{sick}})]$$
$$= 0.92$$

<u>Random Variables</u>: A variable whose numerical values are according to a frequency distribution. Consider the table below:

Student Name	Age	Major
Bob Bobbinson	19	Law
⋮	⋮	⋮

Bob's age may not seem Like a random variable, but it fits:

1) Obviously, it is a numeric value
2) The age of students at the University that Bob attends form a frequency distribution.



Hence age can be considded a random variable. We could say that the set of students and their ages are given by $\{19, 20, 18, \dots\}$ Thus the age of a single student is drawn from this set.

$$\text{Age of one student} \underbrace{\sim}_{\text{"drawn from"}} \{19, 20, 18, \dots\}$$

More gererally

$$x \sim \{x_1, x_2, \dots, x_n\}$$

<u>Probability of x:</u> The Likelihood of x having a specific value is given by that values probability

$$x \sim P(x)$$

If x is discrete

$P \rightarrow$ Probability mass distribution

If x is continuous

$P \rightarrow P(x \in [x, x+dx]) = p(x)dx$

probability density function

↳ Quick note on notation: $P(x)$ represents a function while $P(X = x_1) = P(x_1)$ represents a numerical value. If $P(x_1) = 1$, that means x will always be x_1, and if $P(x_1) = 0$, that means x will never be x_1

In a multivariate case

$P(x, y) \rightarrow$ Joint prob. distribution.

Properties of $P(x)$:

1) The Domain of P must be all probable values of x such that $\forall x_i \in x \rightarrow 0 \leq P(x_i) \leq 1$

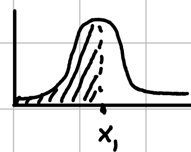
2) $P(x)$ must be normalized, meaning that the sum of the possible outcomes probabilities must be one

$$\sum_{x_i \in x} P(x_i) = 1 \quad ; \quad \text{if } x \text{ is discrete}$$

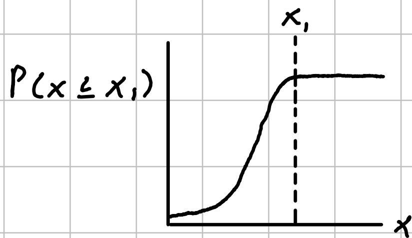
$$\int_{-\infty}^{\infty} p(x) d(x) = 1 \quad ; \quad \text{if } x \text{ is continuous}$$

Cummulative Probability: Usually, when one thinks of probability, they think of the likelihood of an event. But in many cases, one wants to consider the odds of multiple events. For example, consider the continuous variable x, $P(x \le x_1)$ can be represented by

$$\int_{-\infty}^{x_1} p(x)\,dx \rightarrow$$



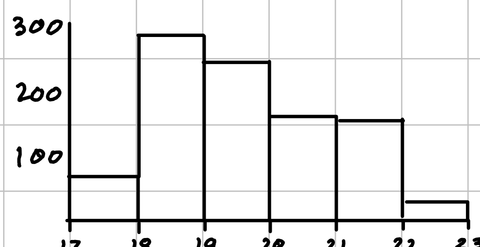
or by



<u>Histogram:</u> A graph plotting, within a series bins covering a numeric range, the distribution of a numeric variable's values.

↳ One of the most useful plot types

As an example we have the distribution of students by age at Bob's university.



<u>Probabilities and Histograms:</u> A rough estimate for an age value falling within some i^{th} bin b_i is

$$P(x \in b_i) \approx \frac{N_i}{N}$$

where N_i is the number of entries in bin i and $N = \sum_i N_i$ represents the total number of entries. In terms of the bin width δ_i, one may write

$$P(x \in b_i) = \frac{N_i \cdot S_i}{N \cdot S_i} = \frac{n_i(x)}{N \cdot S_i}$$

where $n_i(x)$ is the i^{th} bin's area. To obtain an exact probability, one would theoretically have to

$$\lim_{n_i, S_i \to 0} P(x \in b_i) = P(x)$$

<u>The Curse of Dimensionality</u>: When working with high-dimensional data, challenges arise that make the analysis and interpretation of data difficult.

- Data Sparsity: The volume of space between data points grows exponentially, making it harder to find clusters of data
- Distance metrics become less useful: distances become more similar between data points, which negatively impact algorithms that rely on distance metric such as K-nearest neighbor
- Computational complexity increases
- Greater risk of overfitting

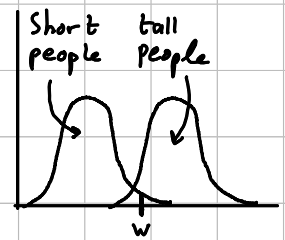
All of these downsides accumulate to make probability estimation difficult in high-dimensional data

<u>Neumann-Pearson lemma</u>: let's say we have two probability density (or mass) functions $f_0(x)$ and $f_1(x)$ We also have two hypothesis H_0 and H_1. The lemma states that the most powerful test to reject H_0 in

favor of H_1 at some significance level is

$$\Lambda(x) = f_1(x) / f_0(x) \quad ; \Lambda \text{ is the test statistic}$$

If the ratio is greater than some threshold value, we reject H_0, if not, we do not reject H_0.



<u>Example of lemma's use:</u> Let's say we have two distributions of people, sorted as short and tall respectively. For a particular point w, the best way to check if person w is short (hypothesis H_0) or tall (hypothesis H_1), is to compute

$$\Lambda(w) = P(x \mid \text{tall}) / P(x \mid \text{short})$$

with some threshold value we choose.

↳ It is clear that this gets complicated in higher dimensions, but that is what machine learning does!!!

$$ML(x) = \text{test statistic}$$