Fundamentals of Probability and Statistics for AI

Artificial Intelligence for Digital Health (AID)
M.Sc. in Digital Health – University of Pisa
Davide Bacciu (davide.bacciu@unipi.it)

Lecture Outline

- Refresher on Probability and Statistics
  - Fundamental principles and definitions
  - Common random variables and probability distributions
  - Some useful rules and concepts
- Statistical hypothesis testing
  - Methods for testing hypotheses
  - Drawing conclusions from data
- Statistical dependence and correlation
  - Linear correlation
  - Mutual information

Probability and Statistics Refresher

Probability

- Intuition
  - Probability as a measure of uncertainty
- Sample Space, Events, and Outcomes
  - All possible outcomes in a scenario
- On the use of probability in AI
  - Handling uncertainty
  - Making informed decisions
  - Learning distributions and generative processes
- Your classical frequentist estimate of discrete probabilities

$$P(A) = \frac{Number\ of\ favorable\ outcomes}{Total\ number\ of\ outcomes}$$

Healthcare Scenario Example

- Healthcare Scenario
  - Predicting whether a patient will develop a particular condition
- Sample Space
  - Includes all possible outcomes (e.g., "develops condition" vs. "does not develop condition")
- Probability Understanding
  - Each outcome has a probability
  - Quantifies the uncertainty in predicting patient's health

Random Variables (RV)

- Definition
  - Variables whose outcomes are determined by chance
  - A function describing the outcome of a random process by assigning unique values to all possible outcomes of the experiment

$$X: \Omega \rightarrow \mathbb{R} \quad \text{where } \Omega \text{ is the sample space}$$

- Types of Random Variables
  - Discrete Random Variables: X = x_i where (i = 1, 2, ..., n)
  - Continuous Random Variables: X $\in [a, b]$
  - Use uppercase to denote a RV, e.g. X, and lowercase to denote a value (observation), e.g. x

- A RV models an attribute of our data
  - Systolic blood pressure as a continuous random variable
  - Number of patients developing a condition as a discrete random variable

Probability Functions

- Discrete Random Variables
  - A probability function $P(X = x) \in [0,1]$ measures the probability of a RV X attaining the value x
  - Subject to sum-rule $\sum_{x \in \Omega} P(X = x) = 1$
- Continuous Random Variables
  - A density function $p(t)$ describes the relative likelihood of a RV to take on a value t
  - Subject to sum-rule $\int_{\Omega}^{t} p(t)dt = 1$
  - Defines a probability distribution, e.g. $P(X \leq x) = \int_{-\infty}^{x} p(t)dt$
- Shorthand $P(x)$ for $P(X = x)$ or $P(X \leq x)$

Common Distributions

- Binomial Distribution
  - Models positive response to a new drug
  - Each patient has a certain probability of responding p, with k patients responding positively over a population of n subjects

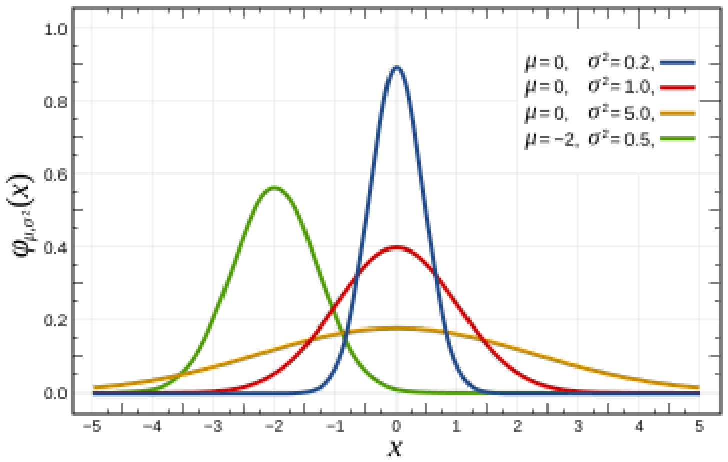
$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

  - Generalized to C different outcomes by the multinomial distribution
- Poisson distribution
  - Models the number of (independent) events occurring within a fixed interval of time or space
  - Modelling the number of patients X admitted to the ER in a given time

$$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

  - With λ average number of arrivals (e.g. patients/hour)
- Normal Distribution
  - Models continuous data such as height or weight of patients
  - Data tends to cluster around a mean value μ with a spread σ^2

$$\mathrm{p}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{\sigma^2}}$$



Joint and Conditional Probabilities

If a discrete random process is described by a set of RVs $X_1, \dots, X_N$, then the joint probability writes

$$P(X_1 = x_1, \dots, X_N = x_n) = P(x_1 \wedge \cdots \wedge x_n)$$

The joint conditional probability of $x_1, \dots, x_n$ given y

$$P(x_1, \dots, x_n | y)$$

measures the effect of the realization of an event y on the occurrence of $x_1, \dots, x_n$

A conditional distribution $P(x|y)$ is actually a family of distributions

- For each y, there is a distribution $P(x|y)$

Chain Rule

Definition (Product Rule a.k.a. Chain Rule)

$$P(x_1, \dots, x_i, \dots, x_n | y) = \prod_{i=1}^{N} P(x_i \,\middle|\, x_1, \dots, x_{i-1}, y)$$

Definition (Marginalization)

Using the sum and product rules together yield to the complete probability

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1 | X_2 = x_2) P(X_2 = x_2)$$

Bayes Rule

Given hypothesis $h_i \in H$ and observations $\boldsymbol{d}$

$$P(h_i|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|h_i)P(h_i)}{P(\boldsymbol{d})} = \frac{P(\boldsymbol{d}|h_i)P(h_i)}{\sum_j P(\boldsymbol{d}|h_j)P(h_j)}$$

- $P(h_i)$ is the prior probability of h_i
- $P(\boldsymbol{d}|h_i)$ is the conditional probability of observing $\boldsymbol{d}$ given that hypothesis h_i is true (likelihood).
- $P(\boldsymbol{d})$ is the marginal probability of $\boldsymbol{d}$
- $P(h_i|\boldsymbol{d})$ is the posterior probability that hypothesis is true given the data and the previous belief about the hypothesis

Expectation of a Random Variable

- The expectation (or expected value) is the long-term average or mean value of a random variable over many trials or instances.
- Represents the 'center of mass' of a probability distribution
- Discrete Random Variables

$$E_{x \sim P}[X] = \sum_{x \in \Omega} x \cdot P(X = x)$$

- Continuous Random Variables

$$E_{x \sim P}[X] = \int_{x \in \Omega} x \cdot p(x) dx$$

- Expectation is a linear operator and works on functions of RVs

$$E_{x \sim P}[f(X)] = \sum_{x \in \Omega} f(x) \cdot P(X = x)$$

Example – Expectation of Discrete RV

Example: Number of Patients Arriving at an ER per Hour

- Let X represent the number of patients arriving at an ER.
- Possible outcomes: $x = 0, 1, 2, \dots, 5$
- Probabilities: $P(X = x) = \{0.1, 0.2, 0.3, 0.25, 0.1, 0.05\}$

- Calculation of $E[X]$:
  $(0 * 0.1) + (1 * 0.2) + (2 * 0.3) + (3 * 0.25) + (4 * 0.1) + (5 * 0.05) = 2.2$

- Interpretation: On average, 2.2 patients are expected to arrive at the ER per hour

Example – Expectation of Continuous RV

Example: Blood Pressure Distribution

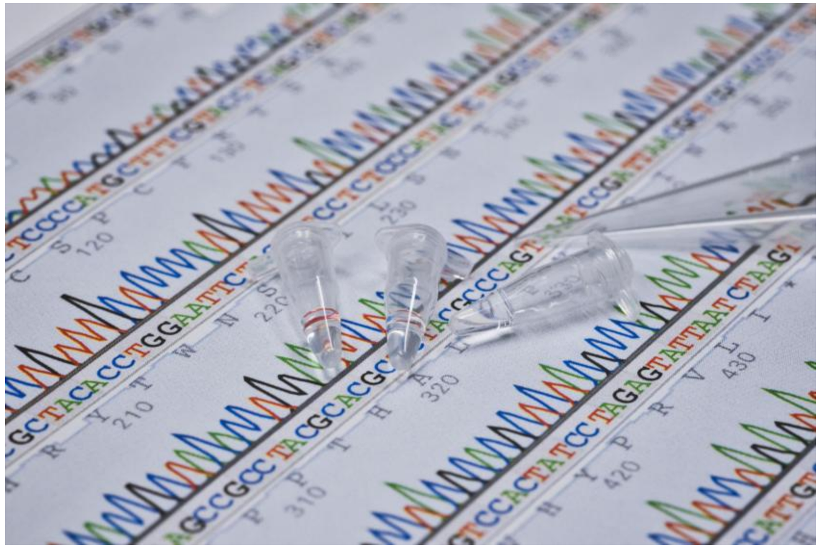
- Let X represent systolic blood pressure in a population.
- Assume X follows a normal distribution with:
  - $\mu = 120$ (mean), $\sigma^2 = 15^2$ (variance)

- Expected Value for a normal distribution: $E[X] = \mu$

- Interpretation: the average systolic blood pressure in this population is 120 mmHg

Statistics Refresher

- Tool for data analysis and inference
- Types of Statistics
  - Descriptive statistics
  - Inferential statistics
- Example in Clinical Study
  - Descriptive: Summarize average age, gender distribution, baseline health
  - Inferential: Draw conclusions about treatment effectiveness based on data from a sample of participants
- Role in AI
  - Summarizing data
  - Drawing conclusions
  - Learning is inference



Descriptive Statistics

- Measures of Central Tendency
  - Mean: Average value of data ⇒ $\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$ with sample size N
  - Median: Middle value when data is sorted
  - Mode: Most frequent value
- Measures of Variability
  - Range: Difference between highest and lowest values
  - Variance: Measure of data spread as squared difference from mean

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{\mu})^2$$

  - Standard Deviation: Measure of data dispersion (square root of variance)
- Example: Patient Ages in Hospital Ward
  - Mean age indicates central tendency
  - Variance and standard deviation show age spread
  - Helps tailor healthcare to demographic

Inferential Statistics

- Making inferences about a population from a sample
  - Importance of sampling (size and coverage) and estimation
- Example: Evaluating a new drug
  - Testing on a sample of patients
  - Using inferential statistics to draw conclusions about the drug's effectiveness

Statistical Hypothesis Testing

Statistical Significance

- Confidence interval
  - Interval which is expected to contain the quantity being estimated

$$\bar{\mu} \mp z \frac{\sigma}{\sqrt{N}} \longrightarrow \text{Error margin}$$

with z being a (critical) value associated to the expected confidence level (e.g. for 95% $z = 1.96$)

- Hypothesis testing
  - Testing assumptions about data: does a test statistics of the population fall into the confidence interval I expect under my hypothesis?
  - In healthcare: assessing effectiveness of new treatment vs. existing one

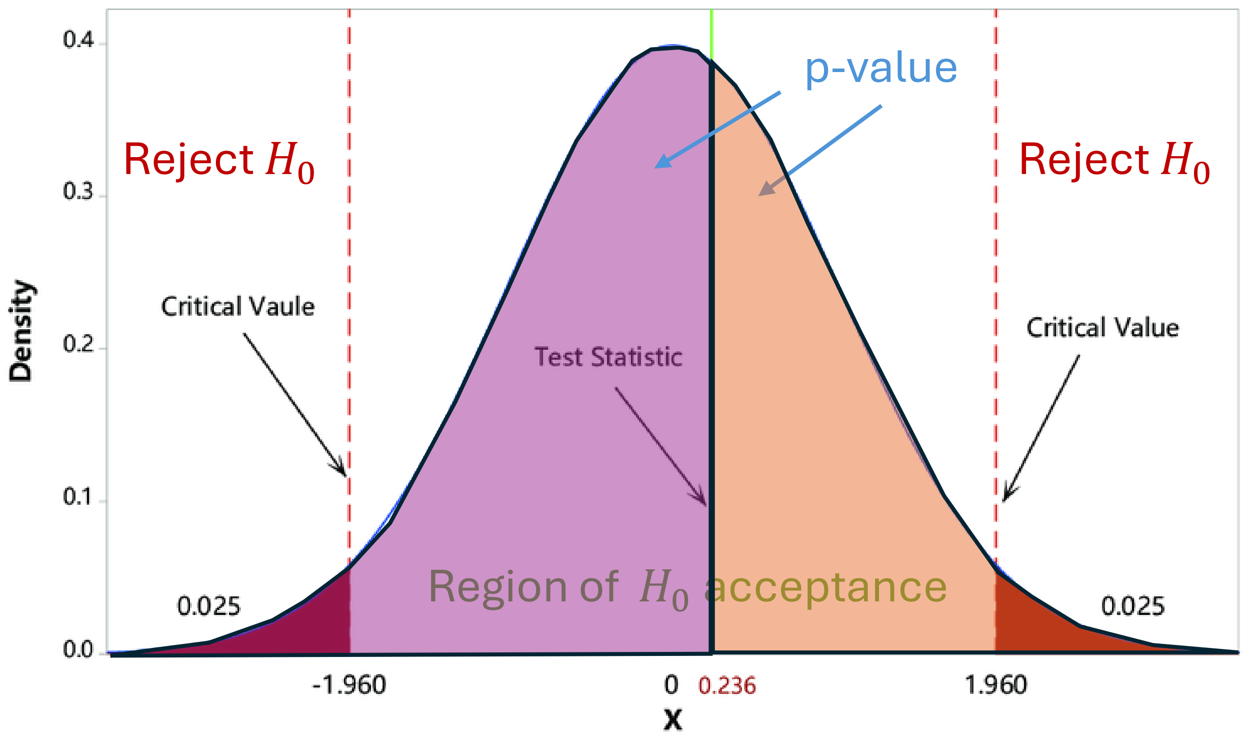
Hypothesis Testing

- Statistical hypothesis: a statement about the parameters describing a population
- Null Hypothesis vs. Alternative Hypothesis
  - Null hypothesis H_0 : e.g. no difference in effectiveness between treatments
  - Alternative hypothesis H_1 : e.g. new treatment is more effective
- P-value
  - Probability of obtaining a result as extreme as the one observed, assuming the null hypothesis is true
  - A very small P-value means that such extreme observed outcome will be highly unlikely under the null hypothesis
  - Else said: P-value less than threshold indicates statistical significance of the alternative hypothesis

Testing Statistical Hypotheses in Brief

1. Define a test statistic (numerical summary) that can be computed from observed data
2. Derive the distribution of the test statistics under the null hypothesis (e.g. a Normal)
3. Select a significance level α defining the maximum acceptable false positive rate (e.g. 5%) and map this to values of the test statistic (critical values)
4. Compute the test statistic for the data and check in which regions it falls (acceptance or critical/rejection regions)

$$p = 2 * \min\{P(T \geq t|H_0), P(T \leq t|H_0)\}$$



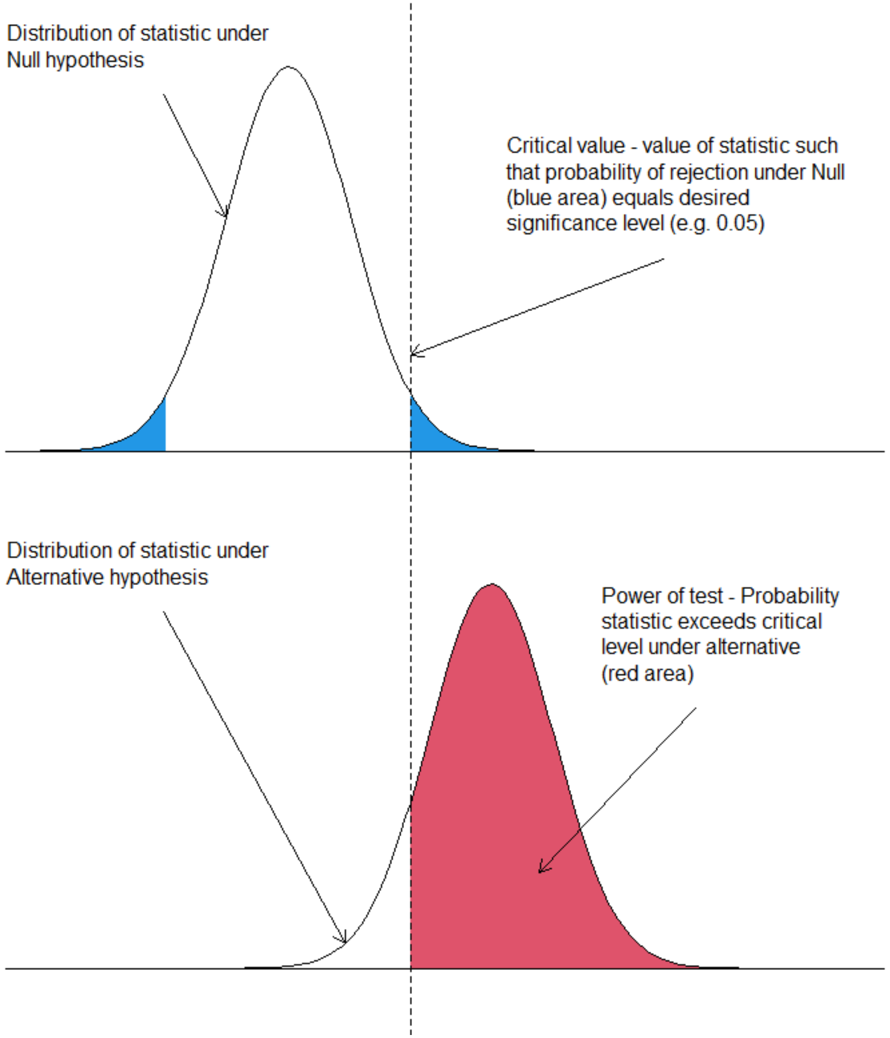
Hypothesis testing with significance level $\alpha = 0.05$
(critical values for $1 - \alpha/2$)

Power of a Statistical Test

The test power $1 - \beta$ is the probability that the test correctly rejects the null hypothesis when the alternative hypothesis is true

	Probability to reject H_0	Probability to not reject H_0
If H_0 is true	α (significance)	$1 - \alpha$
If H_1 is true	$1 - \beta$ (power)	β

Statistical power measures the sensitivity of hypothesis testing to detect a true effect

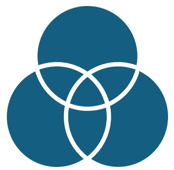


(Image credit to Wikipedia)

Statistical Dependence and Correlation

Understanding Correlation and Dependence

Correlation measures the strength and direction of a linear relationship between two random variables.



Dependence explores how one random variable changes in relation to another, capturing non-linear relationships.

Both are essential in healthcare for analyzing relationships between variables such as symptoms, biomarkers, and outcomes

We will see further in the course probabilistic models specialised to represent dependence between relevant RVs

Independence and Conditional Independence in Probability

- Two RV X and Y are independent if knowledge about X does not change the uncertainty about Y and vice versa

$$I(X,Y) \Leftrightarrow P(X,Y) = P(X|Y)P(Y)$$
$$= P(Y|X)P(X) = P(X)P(Y)$$

- Two RV X and Y are conditionally independent given Z if the realization of X and Y is an independent event of their conditional probability distribution given Z

$$I(X,Y|Z) \Leftrightarrow P(X,Y|Z) = P(X|Y,Z)P(Y|Z)$$
$$= P(Y|X,Z)P(X|Z) = P(X|Z)P(Y|Z)$$

- Shorthand $X \perp Y$ for $I(X,Y)$ and $X \perp Y|Z$ for $I(X,Y|Z)$

Measuring Correlation and Dependence

- Linear correlation analysis uses Pearson's correlation coefficient for quantitative data.
- Mutual information quantifies shared information between random variables.
- Conditional mutual information measures the dependence of two variables given a third (or more).

Linear Correlation Analysis

- Pearson's correlation coefficient ranges from -1 to +1
  - Positive values indicate a direct relationship
  - Negative values indicate an inverse relationship
- Ratio between the covariance of two variables X, Y and the product of their standard deviations

$$\rho_{X,Y} = \frac{\mathbb{E}_{x,y\sim P}[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y}$$

- For sample data it becomes the infamous r coefficient (for its friends)

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{\mu_x})(y_i - \overline{\mu_y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{\mu_x})}\sqrt{\sum_{i=1}^{n}(y_i - \overline{\mu_y})}}$$

- Example: Analyzing the relationship between blood pressure (X) and cholesterol levels (Y)

You can combine correlation analysis with confidence intervals and hypothesis testing to assess uncorrelation

Mutual Information

- Mutual information (MI) measures the information gained about one variable by knowing another
- For discrete RVs this writes as

$$MI(X,Y) = \sum_{x \in \Omega_x, y \in \Omega_y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

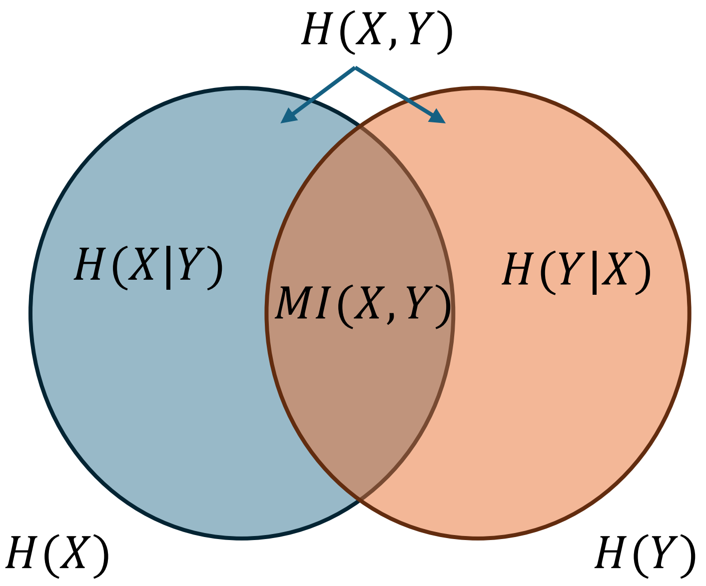
- Higher mutual information indicates more dependence between variables
- Example: Assess how a patient's age (X) influences disease presence (Y)

Mutual Information - Visually

Mutual information can be interpreted as expectation

$$MI(X,Y) = \sum_{x \in \Omega_x, y \in \Omega_y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

$$MI(X,Y) = \mathbb{E}_{x,y\sim P}[\log \mathrm{P}(\mathrm{x},\mathrm{y})] \underbrace{- \mathbb{E}_{x\sim P}[\log \mathrm{P}(\mathrm{x})]}_{\text{Entropy } H(X)} \underbrace{- \mathbb{E}_{y\sim P}[\log \mathrm{P}(\mathrm{y})]}_{\text{Entropy } H(Y)}$$



$MI(X,Y)$ is the intersection of information in X with information in Y

Estimating MI from a Sample

X	Y	$P(X,Y)$
0	0	0.2
0	1	0.3
1	0	0.2
1	1	0.3

- Mutual information just like correlation can be estimated from observed data through statistical methods.
- Use empirical distributions derived from data samples.

Marginal probabilities

$$P(X=0) = 0.2 + 0.3 = 0.5$$
$$P(X=1) = 0.2 + 0.3 = 0.5$$
$$P(Y=0) = 0.2 + 0.2 = 0.4$$
$$P(Y=1) = 0.3 + 0.3 = 0.6$$

$$MI(X,Y) = \sum_{x \in \Omega_x, y \in \Omega_y} P(x,y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

$$MI(X,Y) = 0.2 * \log_2 \frac{0.2}{0.5 * 0.4} + 0.3 * \log_2 \frac{0.3}{0.5 * 0.6} + 0.2 * \log_2 \frac{0.2}{0.5 * 0.4} + 0.3 * \log_2 \frac{0.3}{0.5 * 0.6}$$

Conditional Mutual Information

- Quantifies the information shared between two variables, given a third variable
- For discrete RV X, Y and conditioning variable Z

$$MI(X,Y|Z) = \sum_{x \in \Omega_x, y \in \Omega_y, z \in \Omega_z} P(x,y,z) \log\left(\frac{P(x,y|z)}{P(x|z)P(y|z)}\right)$$

- Useful for controlling confounding factors
- Example: Understanding the relation between smoking and lung cancer while controlling for age

Wrap-up

Take Home Lessons

- Descriptive Vs inferential statistics are central to AI and to biomedical applications
  - Describe population
  - Allow to draw conclusions supported by the data
- Confidence Intervals
  - Range of values within which a population parameter is expected to lie
  - Provides an estimate of the uncertainty around the parameter
- P-Values
  - Probability of obtaining test results at least as extreme as the observed results
  - Used to determine statistical significance
- Statistical Significance
  - Helps in deciding whether to reject the null hypothesis
  - Influenced by confidence intervals and p-values

Next Lecture Preview

- Understand basic concepts of machine learning
- Differentiate between learning paradigms and tasks
- Discuss data types and their roles
- Statistical Learning Theory
- How to evaluate a model and robustly assess its generalization