Mathematical Foundations for Data Science
DA5000

Session 2 – Overview of Probability and Statistics in Data Science

Nandan Sudarsanam,

Department of Data Science and AI,

Wadhwani School of Data Science and AI,

Indian Institute of Technology Madras

Course Overview

- Objective: The course will introduce students to the basic concepts of data analysis and pave the way for understanding more advanced topics in analytics and data science.

- An opportunity for a new start on statistics and probability

- An imperative lens to view all forms of decision-making in management. It is a way of thinking, even if it is not practiced in a prescriptive way.

Why probability and Statistics for Data Science?

- Two reasons:

  1. Many real-world problems that involve data require only the use of concepts in probability and statistics.

     - Data Science is not only Machine Learning or AI. Examples: Data and graphical description, Inferencing, univariate and multivariate modeling

     - Even when the objective is the same (example: Prediction task), prob-stat methods are not dichotomous with AI/ML - They reflect a continuum in the solution approach.

     - Reading: Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." Statist. Sci. 16 (3) 199 - 231, August 2001. https://doi.org/10.1214/ss/1009213726

Why probability and Statistics for Data Science?

2. Probability and statistics is one of the mathematical pillars for Machine Learning

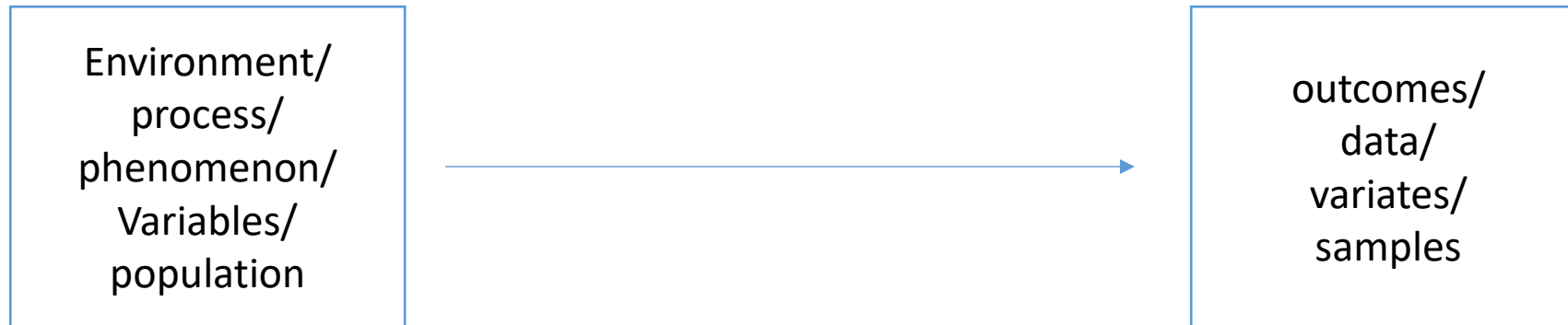   - Example 1: In a prediction problem we are dealing with two Random variables X and Y. A prediction task involves the solution to $E(Y|X=x)$. The concept of conditioning and expectation.

   - Example 2: Some forms of Data reflect the mixture of two underlying phenomena and we are sometimes tasked with separating them according to the possible sources. Examples: heights of students in the class, customers of for a certain car choose it as their primary vehicle or second vehicle. In order to group or cluster the data we assume each data source follows a probability distribution with specific means and variances (Gaussian Mixture Models).

   - Example 3: Bias in expected performance when comparing multiple models through cross-validation and picking the model with best performance. The need for a separate test data set based accuracy.

Course Overview

- Bertsekas, D., & Tsitsiklis, J. N. (2008). *Introduction to probability* (Vol. 1). Athena Scientific.

- Montgomery, D. C., & Runger, G. C. (2010). *Applied statistics and probability for engineers*. John wiley & sons.

Dovetailing with the pre-term

- The big picture.

| Environment/ process/ phenomenon/ Variables/ population | → | outcomes/ data/ variates/ samples |

- The role of descriptive statistics, probabilistic modelling, and inferential statistics
- We will not go in-depth into descriptive. We will revise it.
- The next few sessions will be on the basics of probability

Random Variables

- What are they? Revision
  - Discrete and continuous random variables
  - Moments
  - Joint distributions
  - Functions of Random Variables and simulation examples
- Specific distributions
  - Bernoulli, Uniform (discrete and continuous), Binomial, Poisson, Geometric, Exponential, Hypergeometric, Negative Binomial, Normal or Gaussian.

Basic definitions: Revision

- Random process or experiment: A process/phenomenon/experiment which can result in different outcomes, even though it is repeated in the same manner every time.

- Sample spaces: The set of all possible outcomes of a random experiment is called a sample space.

- Toy examples: Tossing a coin or dropping a ball from a window.

Basic definitions: Revision

- Random variables
  - Moving from tables to expressions: Let us toss a coin 5 times

Outcomes (# of heads)	Probability
0	3%
1	16%
2	31%
3	31%
4	16%
5	3%

$$f(x; n, p) = P(X = x) = {}^n_x C . p^x (1 - p)^{n-x}$$

  - Random variable: the uncertain outcome of a random process, often denoted by capital letters
  - What is small x?
  - The Probability Mass Function (PMF) is the expression itself ($f(x; n, p)$)

Basic Definitions: Revision

- ## Continuous and Discrete random variables
  - Take our examples of tossing coins and dropping balls. What is different?
  - Is it about integers vs decimals? Is it about infinite and non-infinite values? Countable vs non-countable.
  - In the discrete case, the PMF reflects the probability that a specific value or a countable set of values can occur.
  - In the continuous case, the Probability Density Function (PDF) reflects probability of occurrence within a given range or interval.
  - Absolute versus relative likelihood: A random variable is characterized by a mathematical function (PDF/PMF) that quantifies the likelihood (absolute or relative) of the outcomes in the sample space of a random experiment.
  - All RVs can be characterized by a Cumulative Distribution Function (CDF) which captures the probability of occurrences below a certain value x. $\mathrm{F}(x) = P(X \leq x)$.

Basic Definitions: Examples

- Having too many discrete states, and the discretization of a continuous random variable.

- Examples:

| Area	Discrete	Continuous
Manufacturing/ Engineering	Defective/Non Defective, Defects, Arrival in time.	Time to failure, performance/characteristic of a part (strength, stress, etc.), Demand
Corporate (Finance HR Marketing)	Letter based credit rating, delinquency/default, fraudulent transactions, Attrition/onboarding (small samples),Number of units sold, product likeability.	ROI as percentage, PnL, valuation Attrition/onboarding (large samples), Employee salary, time to join, time to leave, Time spent in a shop, Number of clicks/visits per site.
Miscellaneous	Grades in this course, Rain or not, Covid test (+ve or –ve), outcomes of test match in cricket, which team will win a basketball tournament.	Marks in this course, Chemical composition, mileage of a vehicle,, rainfall in mm, # of Covid cases, Body temperature.

Simple ways of describing Random Variables

- We looked at a mathematical expression in the form of PMF and PDF

- Some basics:

  - For PMFs $f(x_i) = P(X = x_i)$ and a) $0 \leq f(x_i) \leq 1$, b) $\sum_{i=1}^{n} f(x_i) = 1$

    - Example: $f(x) = \frac{2x+1}{25}, x = 0,1,2,3,4$

  - For PDFs $P(a \leq x \leq b) = \int_a^b f(x).dx$ and a) $f(x) \geq 0$, b) $\int_{-\infty}^{\infty} f(x) = 1$, c) The probability at a given point is 0

  - Other descriptions are CDF, and the four moments: Mean, Variance, Skew and Kurtosis

Four Moments

- Remember the histogram?
- Remember Mean and standard deviation?
  - You calculated that from data (empirical approach)
  - What if I gave you a PDF/PMF (theoretical)
- Measures of Central Tendency (Mean)
- Dispersion (Variance)
- Skew and Kurtosis

Data Set
10.04
9.31
11.15
11.22
10.19
10.49
8.38
10.32
8.14
7.89
10.07
10.42
11.55
9.63
9.05
8.96
12.57
.
.
.
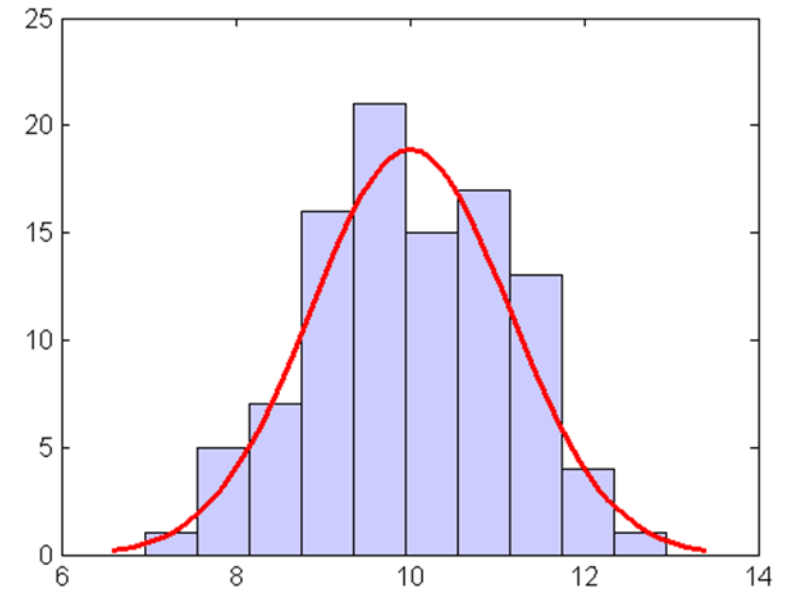
Revision: Measures of Central Tendency (from data)

- Data Set: 3,4,3,1,2,3,9,5,6,7,4,8

- Mean

$$\frac{3+4+3+1+2+3+9+5+6+7+8+4}{12} = 4.583 \qquad \frac{x_1+x_2+x_3+\cdots}{n} \quad \text{or} \quad \frac{\sum_{i=1}^{n} x_i}{n}$$

- Median

1,2,3,3,3,4,4,5,6,7,8,9 Hence Answer = 4

- Mode

The value 3 appears 3 times, and 4 appears 2 times and all other values appear once. Hence 3 is the mode

Measures of Central Tendency

- Where do we want to use Mean, Median and Mode

- Choosing between mean and median
  - Bad outliers
    - Errors
    - Do not provide a realistic picture of the story
  - Good outliers
    - The story is in the outliers

- Mode
  - Useful with nominal variables
  - Multi modal distributions

Revision: Measures of Dispersion

- Data set: 3,4,3,1,2,3,9,5,6,7,4,8
- Range (Max-Min) (9-1 = 8)
- Inter Quartile Range: 3rd quartile -1st quartile (75th Percentile – 25th Percentile) (6.5 – 3 = 3.5)
- Sample Standard deviation

$$\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{1}{12-1}\sum((3-4.58)^2 + (4-4.58)^2 \ldots)$$

Revision: Measures of Dispersion

- Questions that go with Standard deviation
  - Why do we use the square function on the deviations? What are its implications?
  - Why do we work on standard deviation and not the variance?
  - Why do we average by dividing by N-1 and not N?

- Mean absolute Deviation and its variants
  - Use $|x_i - \bar{x}|$ instead of $(x_i - \bar{x})^2$

We often use **standard deviation** instead of **variance** in statistical analysis for several practical reasons:

1. **Same Unit as the Data:**
   - **Variance** is the average of the squared differences from the mean, which means its unit is squared (e.g., if data is in meters, variance is in square meters).
   - **Standard deviation** is the square root of variance, bringing it back to the same unit as the data. This makes it easier to interpret and relate directly to the original data.

2. **Intuitive Interpretation:**
   - Standard deviation measures the **average distance of data points from the mean**, giving a more intuitive sense of how spread out the data is. Since it's in the same unit as the data, people find it easier to understand the variability in context.
   - Variance, because it's in squared units, is harder to interpret and compare with the original data.

3. **Direct Relationship with Normal Distribution:**
   - In a **normal distribution**, about 68% of the data falls within one standard deviation of the mean, and about 95% within two standard deviations. This direct relationship makes standard deviation a helpful measure for understanding data spread.
   - Variance doesn't offer this intuitive relationship because it's in squared units.

4. **Consistency Across Fields:**
   - Standard deviation is more commonly used in reports and scientific research, providing consistency across various fields of study. While variance is also important, especially in theoretical work, standard deviation is the preferred choice in many applied contexts.

In summary, standard deviation is favored because it is in the same units as the data, making it easier to interpret and more practical for everyday use.

Overview of Mean and Standard deviation

	Empirical	Theoretical/Conceptual	
		Discrete	Continuous
Mean	$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$	$E(X) = \mu = \sum_{i=a}^{b}(x_i \cdot p_i)$	$E(x) = \mu = \int_a^b x f(x) dx$
Standard Deviation/ Variance	$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ Or $s = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2}$	$\text{Var}(x) = \sigma^2 =$ $\sum_{i=a}^{b} p_i(x_i - \mu)^2$ or $\sum_{i=a}^{b} p_i(x_i)^2 - \mu^2$	$\text{Var}(x) = \sigma^2 =$ $\int_a^b (x-\mu)^2 f(x).dx$ or $\int_a^b (x)^2 f(x).dx - \mu^2$

- Theoretical standard deviation in terms of mean:

$$E\big((X-\mu)^2\big) =$$
$$E(X^2) - (E(X))^2$$

General Operations on single Random Variables

- Effect of adding or removing a constant
  - Mean $E(X \pm c) = E(X) \pm c$
  - Standard Deviation $Var(X \pm c) = Var(X); SD(X \pm c) = SD(X)$
- Effect of multiplying or removing a constant
  - Mean $E(cX) = c\,E(X)$
  - Standard Deviation $Var(cX) = c^2\,Var(X)$; $SD(cX) = |c|\,SD(X)$
- Putting these together
  - $E(a + bX) = a + bE(X)$
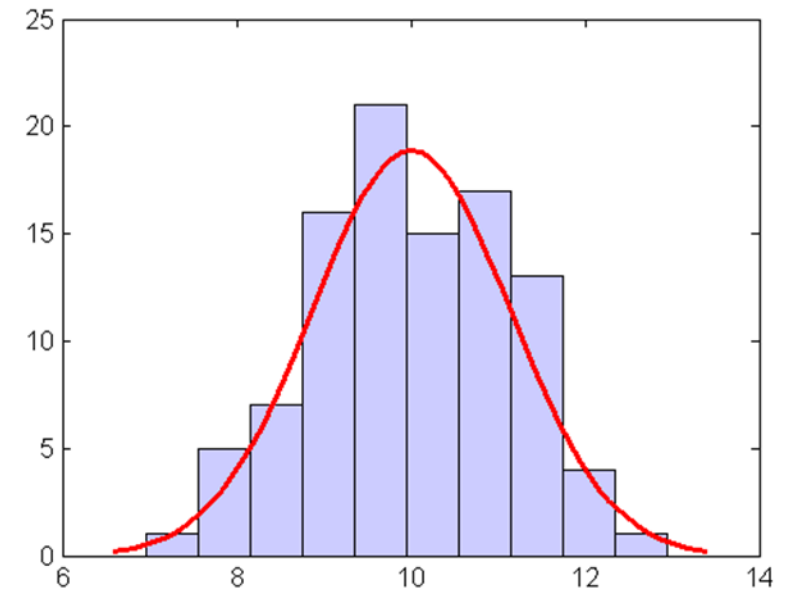 - $SD(a + b\,X) = |b|SD(X)$
  - $Var(a + bX) = b^2 Var(X)$

Skew and Kurtosis

- What about skew and kurtosis?
- Skew:
  - Left or right-tailed property
  - It reflects the asymmetric deviation from the mean
- Kurtosis
  - Fat tailed
  - If I make the tails fatter, am I not just increasing variance?

Data Set
10.04
9.31
11.15
11.22
10.19
10.49
8.38
10.32
8.14
7.89
10.07
10.42
11.55
9.63
9.05
8.96
12.57
.
.
.

Histogram

Joint Distributions

- Why do we need to look at distributions jointly.
  - Photocopy shop example
- Modified example from the text

		X (Revenue)			p(y)
		x = 3	x = 5	x = 7	
Y (cost)	y= 2	0.08	0.07	0.00	0.13
	y = 4	0.01	0.62	0.02	0.65
	y = 6	0.00	0.11	0.09	0.20
	p(x)	0.09	0.80	0.11	1

- Joint probability P(X=3, Y=2)
- Marginal probability (P(X=3))
- Conditional probability P(Y=2|X=3)

Going from tables to formulas

- Joint

PDF $f_{X,Y}(x,y)$ captures $P(a \leq X \leq b, c \leq Y \leq d)$

$= \int_c^d \int_a^b f(x,y).\,dx.\,dy$

PMF $f_{X,Y}(x,y)$ captures $P(X = x \text{ and } Y = y)$

- Marginals: $f_X(x) = \sum_{y=-\infty}^{\infty} f(x,y)$ or $\int_{-\infty}^{\infty} f(x,y).\,dy$

- Conditionals $f_{X|Y}(x|y) = \dfrac{f(x,y)}{f(y)}$

Functions of Random Variables

- Sometimes we are interested in a function of the random variable
  - Example1: If the number of people who visit our store is a random variable, our revenue is a function of this random variable.
  - Example 2: If the price of oil is random variable, then the price of petrol is a function dependent on this.
- The function of a random variable is another random variable which we can obtain from the original PMF or PDF

Functions of Random Variables

- Let $f_X(x)$ be the RV, and we are interested in the RV $Y = g(X)$

- Expected value can be straight forward $E(Y) = \int_{-\infty}^{\infty} g(x).f_X(x).dx$

- Steps for PDF:
  - Find CDF as $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y))$
  - If CDF is continuous then $f_Y(y) = \frac{d}{dy}(F_Y(y))$

- Example: Let X be uniform on [0,1]. Find PDF of $Y = \sqrt{X}$
$$F_Y(y) = P(Y \leq y) = P(\sqrt{X} \leq y) = P(X \leq y^2) = y^2$$

Therefore $f_Y(y) = 2y, 0 \leq y \leq 1$

Properties of combining two random variables

- P(X=3, Y=2) can be written $P_{i,j} = P_{1,1}$
- P(X=3, Y=2) is the same as P(X=3 ∩ Y=2). If X=3 is an event (A) and Y=2 is an event (B) then = P(A ∩ B).
- From what we understood in conditionals it is clear that

P(A|B) = P(A ∩ B)/P(B)

- Multiplication rule: $P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)$
- Bayes theorem:$P(B|A)= \dfrac{P(A|B)P(B)}{P(A)}$
- Two events are independent if $P(B|A) =$ P(B) or $P(A|B) = P(A)$ therefore $P(A \cap B) = P(A) \cdot P(B)$

Example problems using Bayes theorem

- We sell computers with optional insurance against damage. 90% of our customers do not take insurance. We find that those who don't purchase insurance are more careful with the product than those who do. There is a 30% chance that a customer with insurance will file a repair request, where as only 5% of those without insurance will file a request.

- If we receive a repair request, what is the probability that it is from a customer who has insurance?

$$\frac{0.3 \times 0.1}{0.3 \times 0.1 + 0.05 \times 0.9} = 40\%$$

Examples using Bayes theorem

- Let us say a stock will go up if a merger goes through (down if it does not). Your insightful market research shows that there is an 80% chance of the merger occurring. Historically your research team has been good. However, you do not want to take a bet on just this information. You hire an insider to overhear the meeting involving the merger discussions. However, this informant's listening is only 90% accurate (or she sometimes mishears the discussion). Following the meeting, she calls you and tells you that the merger is not happening. What are the odds that she is correct?

$$\frac{0.9 \times 0.2}{0.9 \times 0.2 + 0.8 \times 0.1} = 69\%$$

Decision-making using EVPI and EVSI

Expected Value of Perfect Information (EVPI)

- EVPI is the expected price that we are willing to pay to gain access to perfect information.

- EVPI = Expected value with perfect information - Expected value without perfect information

Expected value of sample information (EVSI)

- EVSI is the expected increase in monetary value by obtaining access to sample of additional observations.

- EVSI = Expected value with sample information – Expected value without sample information

Example 1

- Consider a situation in which there is 80% chance that stock price goes up, and 20% chance that it goes down. The profit we make by investing when price is up is Rs.10, and the loss we incur by investing when price is down is Rs.20. Calculate EVPI

- The payoff matrix can be represented as follows:

Decision	Stock up	Stock down
Buy	10	-20
Not buy	0	0
Probability	0.8	0.2

Example 1

- Expected value without perfect information:
  - Expected monetary value = (10*0.8) + (-20*0.2)

    (EMV)   = 4

- Expected value with perfect information (EV|PI):
  - If it is know that the stock is up, we make a decision to buy. If it is down, we make a decision to not buy.
  - EV|PI = (0.8*10) + (0.2*0) = 8
- EVPI = EV|PI – EMV

      = 8 - 4 = 4

Example 1: EVSI scenario 1

- Considering Example 1, assume we have the stock price status for previous 5 years. From this sample of 5 years, 90% of time we predict the current status correctly. 10% of the time the prediction is wrong.

- That is, if the stock is actually up, we predict that it is up with probability 0.9, and predict that it is down with probability 0.1

Example 1: EVSI scenario 1

Decision	Stock up	Stock down
Buy	10	-20
Not buy	0	0
Probability	0.8	0.2

Actual status

	Up	Down
Up	0.9	0.1
Down	0.1	0.9

Sample info

- EV|SI = (0.8*0.9*10) + (0.1*0.2*(-20)) = 6.8
- We know that, EMV = 4
- EVSI = EV|SI – EMV

    = 6.8 – 4 = 2.8
- Note: EVPI value was 4

Example 1: EVSI scenario 2

- Suppose for example 1, the probabilities from the sample is modified by the given matrix.

Actual status

Sample info		Up	Down
	Up	0.7	0.2
	Down	0.3	0.8

- EV|SI = (0.8*0.7*10) + (0.2*0.2*(-20)) = 4.8
- We know that, EMV = 4
- EVSI = EV|SI − EMV

    = 4.8 − 4 = 0.8

Decision	Stock up	Stock down
Buy	10	-20
Not buy	0	0
Probability	0.8	0.2

- A note on functions of random variables