Indian Institute of Technology Madras

Probability and Statistics for Data Science and AI

Nirav Bhatt
Email: niravbhatt@iitm.ac.in

Module: Probability and Statistics for DSAI

- Probability for DSAI
  - Revisiting: High school probability topics
  - Conditional Probability
  - Random Variables, Expectation, Distributions, and Random Processes
  - Applications of concepts in Probability for DSAI
- Statistics for DSAI
  - Descriptive statistics and Visualization
  - Inferential Statistics: Frequentist inference, Bayesian inference, Hypothesis testing
  - Applications of concepts in Statistics for DSAI

Outline

- Experiments and Random Phenomena
- Revisit to high school probability topics
- Random variables
- Important probability distributions and properties
- Descriptive statistics

Statements we make!

- Mom, I will be there in 10 minutes.
- Today, it is very humid, it will definitely rain.
- Amazon parcel will most likely to deliver today.
- I may make it to your party tomorrow.
- Machchi, You are less likely to get Corona.
- It is impossible to get 100% in this exam

Statements we make!

- How do we make these statements?
  - Past experience?
  - Some information available to you?
- Collected data all the time about different phenomena around us
- Do reasoning subconsciously
- Make these statements

Experiments

- **Experiment I:** Tossing a coin
  - Fair coin, Fair and same tossing machine, constant wind speed, temperature or any other variables that can affect the outcome of tossing a coin
  - Outcome at each tossing:

  - Observations

Experiments

- **Experiment II:** Measuring concentration of 1 M glucose in bottle by a set of students
  - Same day, same equipment, constant temperature or any other variables that can affect the outcome
  - Outcomes: 0.98, 1, 1.2, 0.85, 0. 97, 1.02, 1.1, 0.95, 0.89, 1.06, …
  - Observations

Experiments

- **Experiment III**
  - Determine (integer) age of a particular student in the class room
    Using date of birth on ID card
  - Outcomes: 18, 18, 18, 18, 18, 18, 18, 18, 18
  - Observations

Experiments and Variability

- Experiments I and II
  - Same experimental conditions
  - Different outcomes
  - Outcome cannot be predicted with certainty
  - Encountered variability
- Experiment III:
  - Same experimental conditions
  - Identical outcomes
  - Outcome can be predicted with certainty
  - Deterministic in nature

Variability and Random Phenomena

- Variability
  - Something that cannot be controlled
  - “Identical" conditions: Different outcomes of the phenomenon
  - Do not product exactly the same outcome
- Experiments I and II
  - Each toss either Head or Tail
  - Student measured different glucose concentrations for 1 M bottle
- Random Phenomenon:
  A phenomenon in which “randomness” is involved and outcomes cannot be predicted or controlled

Random Phenomena

- Characterising random phenomena
  - Different outcomes for same experimental conditions
  - Variability associated with different outcomes

    Is it possible to characterize?
  - Example I : Number of students attending BT5450 class on Friday?

    Finite number of outcomes: (0,. . . ., total number of students taking the course) — Discrete
  - Example II: Average height of students attending BT5450 class on Friday in cm?

    Infinite number of outcomes: Any number in an interval — Continuous

Random Phenomenon

- Deterministic Phenomenon
  - A phenomenon whose outcome can be predicted with a very high degree of confidence for the same experimental conditions
  - Example: Age of a person in year based on birth date in driving licence
- Random Phenomenon
  - A phenomenon which can have several possible outcomes for the same experimental conditions
  - Outcome can be predicted with limited confidence
  - Example: Arrival time of 9.30 am bus at bus-stand in India

Characterizing random phenomena

- Sources of error in observed outcomes
  - Model error: Lack of knowledge of generating process

    e.g. Development of unstructured kinetic models
  - Measurement error: Errors in sensors used for observing outcomes

    e.g. Measuring weight on different weighting machines
- Types of random phenomena
  - Discrete: Outcomes are finite
  - Continuous: Infinite number of outcomes

Random Phenomena

- Our interest: Variability associated with model and measurement errors
- Need a framework allows to understand and quantitatively characterize random phenomena

Random Experiment

- An experiment that can result in different outcomes when it is repeated under identical conditions every time
- Analysis of such a system: need to understand set of all possible outcomes

Sample space and events

- Sample space: The set of all possible outcomes of a random experiment
- Event: Subset of the sample space is an event

Concept of Counting

Permutation

- Example: A,C,G,T. How many ways they can be arranged without repeating them?
  ACGT,ACTG,ATCG,ATGC AGCT,AGTC
  CAGT,CATG,CGAT,CGTA,CTGA,CTAG
  TAGC,TACG,TCAG,TCGA,TGAC,TGCA
  GCAT,GCTA,GATC,GACT,GTAC,GTCA
- Permutations : Arrangement of items in an orderly manner
- n=4 (A,C,G,T), $4! = 4 \times 3 \times 2 \times 1 = 24$
- **Distinct *n* objects can be arranged in n ! (factorial) ways**

Concept of Counting

Permutation

- Example: How many 4 digits number can be created from 0,1,...,9 without repetition of digits?
- **Permutations (without repetition): From distinct n objects choose r:** $P_r^n = \dfrac{n\,!}{(n-r)\,!}$
- $P_4^{10} = \dfrac{10!}{(10-4)!} = \dfrac{10!}{6!} = 10 \times 9 \times 8 \times 7 = 5040$
- What about permutation with repetition?
- $10 \times 10 \times 10 \times 10 = 10^4 = 10000$
- **Permutations (with repetition): From distinct n objects choose r,** n^r

Concept of Counting

Permutation

- How many the same length arrangements of “Machchi" are possible?

Concept of Counting

Permutation

- **Permutations with multiple items:** n **objects with items of** k_1 **type,** k_2 **type,**$\ldots, k_k$**type**

$$\frac{n!}{k_1!k_2!\ldots k_k!}$$

- How can a family 5 can be arranged on circular table?
- **Permutations (Clockwise and anti-clockwise are different)**
- **Permutations (Clockwise and anti-clockwise are same)**

Combinations

order doesn't matter

- Choosing a hockey team from 17 players?

Combinations

order doesn't matter

- Combination (without repetitions): the number of combinations in which r items can be chosen from a set of n items $C_r^n = \frac{n!}{r!(n-r)!} = \frac{P_r^n}{r!}$
- six Gelato (Sorbet) flavors in the shop, the number of combination in which three scoops of Gelato (Sorbet) can be chosen?

$$C_3^6 = \frac{(6+3-1)!}{3!(6-1)!} = 56$$

- Combinations with repetition (n items, choosing r items)

$$C_r^n = \frac{(n+r-1)!}{r!(n-1)!}$$

Exclusive and Independent Events

- Mutually exclusive events: Two events are mutually exclusive if occurrence of one precludes occurrence of the other
- Independent events: Two events are independent if occurrence of one has no influence on occurrence of other

Different types of events

- Consider two events E and F of a sample space Ω

New Events

Union ($E \cup F$): A set consisting all elements in either E or F (or both)
Intersection ($E \cap F$): A set consisting all elements in both E and F
Null event (Mutually exclusive): $E \cap F = \{\} \rightarrow$ Events E and F cannot both occur

Different types of events

- Consider two events E and F of a sample space Ω

New Events

Complement event (E^c): E^c is the set of all the elements in Ω but are not in E, i.e. $E \cap E^c = \Phi$

Difference event ($E - F$): The set of all the elements in Ω that are in E but are not in F, i.e. $E \cap F^c$

Introduction to Probability

Algebra of Events

- Consider two events E and F of a sample space $\mathbf{\Omega}$

Rules

Commutative law: $E \cup F = F \cup E$, $E \cap F = F \cap E$
Associative law: $(E \cup F) \cup G = E \cup (F \cup G)$, $(E \cap F) \cap G = E \cap (F \cap G)$
Distribution law: $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$
$(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

Probability Measure

- Probability measure is a function that assigns a real value to every outcome of a random phenomena which satisfies following axioms
  - $P(S) = 1$ (one of the outcomes should occur)
  - $0 \leq P(A) \leq 1$ (Probabilities are non-negative and less than 1 for any event A)
  - For two mutually exclusive events A and B

$$P(A \cup B) = P(A) + P(B)$$

Conditional Probability

Conditional Probability

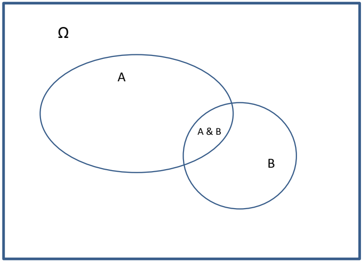
- If two events A and B are not independent, then information available about the outcome of event A can influence the predictability of event B

Definition

The probability of the occurrence of an event subject to the hypothesis that another event(s) has occurred

- $P(E|F)$: the event E will occur given that the event F has occurred.
- Assumption: The event F is not impossible event.
- Computation of $P(E|F)$:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}, \ P(F) > 0$$

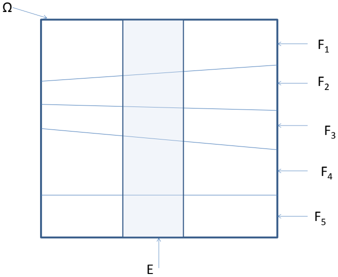


- Example: two (fair) coin toss experiment
  - Event A : First toss is head = {HT, HH}
  - Event B : Two successive heads ={HH}
  - P(B)=0.25 (no information)
  - Given event A has occurred, P(B/A)=0.5

Bayes' formula

- Direct computing the probability of an event is difficult. How?
- Consider mutually exclusive events F_1, F_2,...,F_n of $\boldsymbol{\Omega}$ such that

$$\bigcup_{i=1}^{n} F_i = \boldsymbol{\Omega}$$



- For an event E,

$$E = \bigcup_{i=1}^{n} (E \cap F_i)$$

Bayes' formula contd.

- The probability of the event E:

$$
\begin{aligned}
P(E) &= \sum_{i=1}^{n} P(E \cap F_i) \\
&= \sum_{i=1}^{n} P(E|F_i)P(F_i)
\end{aligned}
$$

- $P(F_i)$ given the event E has occurred?
- Bayes' formula:

$$
P(F_i|E) = \frac{P(E \cap F_i)}{P(E)} = \frac{P(E|F_i)P(F_i)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}
$$

-

Independent Events

- Two events E and F are said to be *independent* if

$$P(E \cap F) = P(E)P(F)$$

- If they are not independent, they are said to be dependent
- Some results:
  - If E and F are independent, then so are E and F^c.
  - The events F_1, F_2,...,F_n are independent if for every subset $F_{1'}$, $F_{2'}$,...,$F_{r'}$, $r' \leq n$, of these events

$$P(F_{1'} \cap F_{2'} \cap \ldots F_{r'}) = P(F_{1'})P(F_{2'}) \ldots P(F_{r'})$$

Random Variables

- Experiment 1: The testing of three components for quality (*N*: Non-defective, and *D*: Defective)
  $\boldsymbol{\Omega} = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$
- Each point in $\boldsymbol{\Omega}$ assigned a numerical value (0,1,2,3)
  $X(NNN)$ =0; $X(NND)$=1,
  $X(NDN) = 1$,$X(DNN)$=1;
  $X(NDD) = 2$, $X(DND) = 2$,
  $X(DDN) = 2$, $X(DDD)$=3

Random Variables

- Experiment 1: The testing of three components for quality (N: Non-defective, and D: Defective) $\mathbf{\Omega} = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$
- Each point in $\mathbf{\Omega}$ assigned a numerical value (0,1,2,3) $X(NNN)$ =0; $X(NND)$=1, $X(NDN) = 1$,$X(DNN)$=1; $X(NDD) = 2$, $X(DND) = 2$, $X(DDN) = 2$, $X(DDD)$=3

- Experiment 2: Choosing three fruits from the basket having apples (A) and oranges (O) $\mathbf{\Omega} = \{AAA, AAO, AOA, OAA, AOO, OAO, OOA, OOO\}$
- Each point in $\mathbf{\Omega}$ assigned a numerical value (0,1,2,3) $X(AAA)$ =0; $X(AAO)$=1, $X(AOA) = 1$,$X(OAA)$=1; $X(AOO) = 2$, $X(OAO) = 2$, $X(OOA) = 2$, $x(OOO)$=3

Random Variables

- Experiment 1: The testing of three components for quality (N: Non-defective, and D: Defective) $\mathbf{\Omega} = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$
- Each point in $\mathbf{\Omega}$ assigned a numerical value (0,1,2,3) $X(NNN)$ =0; $X(NND)$=1, $X(NDN) = 1$,$X(DNN)$=1; $X(NDD) = 2$, $X(DND) = 2$, $X(DDN) = 2$, $X(DDD)$=3

- Experiment 2: Choosing three fruits from the basket having apples (A) and oranges (O) $\mathbf{\Omega} = \{AAA, AAO, AOA, OAA, AOO, OAO, OOA, OOO\}$
- Each point in $\mathbf{\Omega}$ assigned a numerical value (0,1,2,3) $X(AAA)$ =0; $X(AAO)$=1, $X(AOA) = 1$,$X(OAA)$=1; $X(AOO) = 2$, $X(OAO) = 2$, $X(OOA) = 2$, $x(OOO)$=3

Definition

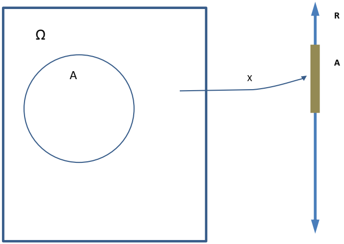
A function that associates a unique real number corresponding to every outcome of sample space.

Random Variables

- Formal definition:

Random Variable

A function $X : \mathbf{\Omega} \rightarrow \mathbf{R}$ is a random variable if and only if $P(\{\omega \in \mathbf{\Omega} : X(\omega) \leq y\})$ exists for all choices of $y \in \mathbf{R}$.

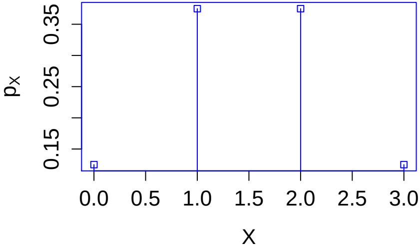


- Coin toss sample space [H T] mapped to [0 1].
- If the sample space outcomes are real valued no need for this mapping (eg. throw of a dice)
- Allows numerical computations such as finding expected value of a RV
- Discrete RV (throw of a dice or coin) Continuous RV (sensor readings, time interval between failures)
- Associated with the RV is also a probability measure

Random Variables

- Example: Probability measure

$$P_X(X=0)=\frac{1}{8},$$
$$P_X(X=1)=\frac{3}{8},$$
$$P_X(X=2)=\frac{3}{8},$$
$$P_X(X=3)=\frac{1}{8}$$



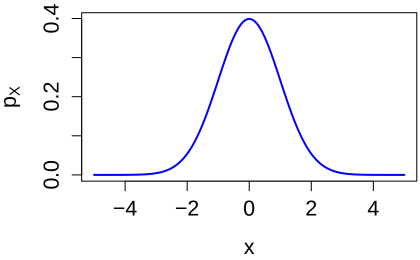
- Some observations: X is a finite integer number, and discrete
- Discrete random variable: X *whose set of possible values forms a discrete set*
- Set of possible values : a finite or infinite sequence
- P_X: Probability mass function (PMF): $P_X(X=x_i)=p_i$
- PMF also follows properties of a general probability

Random Variables

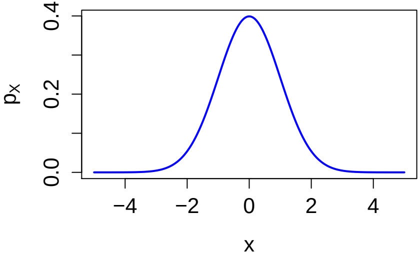
- Examples:
  - X: Length of a randomly selected telephone call
  - X: Life of a randomly selected bicycle in IIT
  - X: Height, weight of a randomly selected students etc
- Observation: X: continuous in nature, and can take on any value in some interval (a, b)

Random Variables

- Examples:
  - X: Length of a randomly selected telephone call
  - X: Life of a randomly selected bicycle in IIT
  - X: Height, weight of a randomly selected students etc
- Observation: X: continuous in nature, and can take on any value in some interval (a, b)



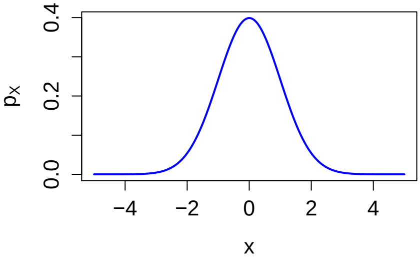
Random Variables



- PMF will not provide any useful information
- X continuous variable if there is a function $f(x) \geq 0$ so that any interval $-\infty \leq a \leq b \leq \infty$,

$$P_X(a \leq X \leq b) = \int_a^b f(x)dx$$

Random Variables

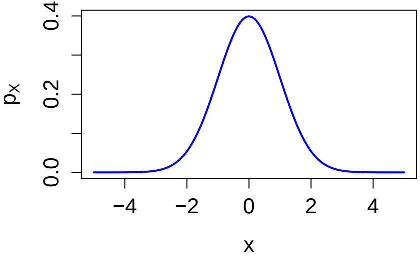


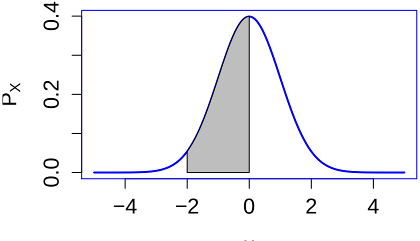
- PMF will not provide any useful information
- X continuous variable if there is a function $f(x) \geq 0$ so that any interval $-\infty \leq a \leq b \leq \infty$,

$$P_X(a \leq X \leq b) = \int_a^b f(x)dx$$

- $P(-2 \leq X \leq 0)$?

Random Variables



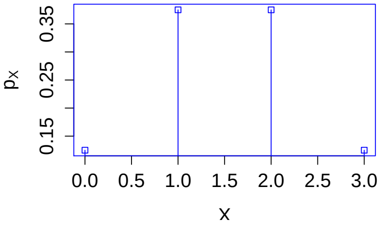


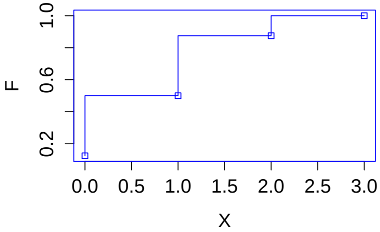
- PMF will not provide any useful information
- X continuous variable if there is a function $f(x) \geq 0$ so that any interval $-\infty \leq a \leq b \leq \infty$,

$$P_X(a \leq X \leq b) = \int_a^b f(x)dx$$

- $P(-2 \leq X \leq 0)$= Area under the curve

Random Variables





- Cumulative distribution function, c.d.f. of X for any real number a

$$F(a) = P(X \leq a) = \sum_{all\ x \leq a} p(x)$$

- Example: F(1) =0.5, F(2)=7/8
- c.d.f. for continuous X

$$\begin{aligned} F(a) &= P(X \leq a) \\ &= P(X \in (-\infty, a])) \\ &= \int_{-\infty}^{a} f(x)dx \end{aligned}$$

- Relationship between c.d.f and p.d.f

$$\frac{dF(a)}{da} = f(a)$$

Random Variables

- Game of rolling a balanced die:
  - Result: 2, 3 or 4; *Win: Rs. 10*
  - Result: 5; *Win: Rs. 20*
  - Result: 1, 6; *Win: Rs. -30*
- Should you play the game?
- $\boldsymbol{\Omega} = \{1, 2, 3, 4, 5, 6\}$
- X: the payoff amount in the game
  $X(1) = X(6) = -30$;
  $X(2) = X(3) = X(4) = 10$; $X(5) = 20$
- Probability mass function
  - $P_X(X = -30) = 1/3$
  - $P_X(X = 20) = 1/6$
  - $P_X(X = 10) = 1/2$

- Expected the payoff amount after the n games

$$
\begin{aligned}
E(Payoff) &= (20\,\frac{1}{6} + 10\,\frac{1}{3} - 30\,\frac{1}{3})\,n \\
E(Payoff) &= -\frac{10\,n}{6}
\end{aligned}
$$

Random Variables

- Expectation or expected value of X ($E[X]$): the average value of X in a large number of trails
- Not a value X could possible take
- For discrete random variables:

$$E[X] = \sum_i x_i P(X = x_i) \tag{1}$$

- For continuous random variables:

$$E[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx \tag{2}$$

Random Variables

- $E[X]$: The weighted average of the possible values of X, often denoted as μ

  No information about the spread of these values

- If we quantify the spread using a function $\phi(x)$

$$\phi(x) = (x - \mu)^2 \tag{3}$$

Definition: Variance

If X is a random variable with mean μ, the variance of X, denoted by $Var(X)$, is defined by

$$\begin{aligned} Var(X) = E[(X-\mu)^2] &= \sum_{x \in X(S)} (x-\mu)^2 f(x) \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

- The standard deviation of X is the square root of its variance

Random Variables

- Some properties of variance
  - For constant c, $Var(X + c) = Var(X)$
  - $Var(cX) = c^2 Var(X)$
  - $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$
- Variance for a continuous random variable

$$Var(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \qquad (4)$$

Random Variables

- Some properties of variance
  - For constant c, $Var(X + c) = Var(X)$
  - $Var(cX) = c^2 Var(X)$
  - $Var(X_1 + X_2) = Var(X_1) + Var(X_2)$

Random Variables

- For two random variables X and Y

Definition: Covariance

If $E[X] = \mu_x$ and $E[Y] = \mu_y$, the covariance of X and Y,

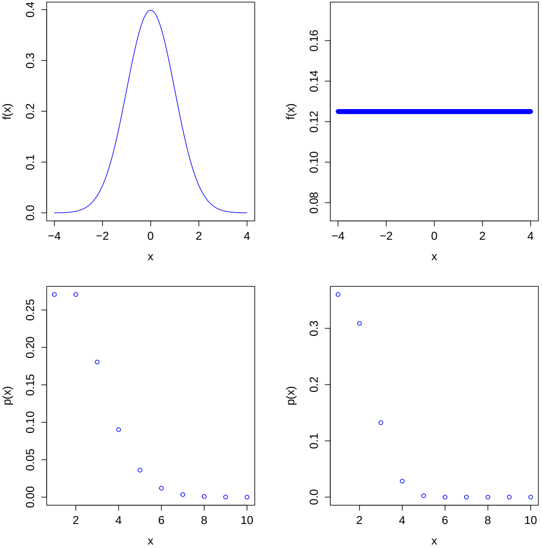
$$\begin{aligned} Cov(X, Y) &= E[(X - \mu_x)(Y - \mu_y)] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

- Properties:

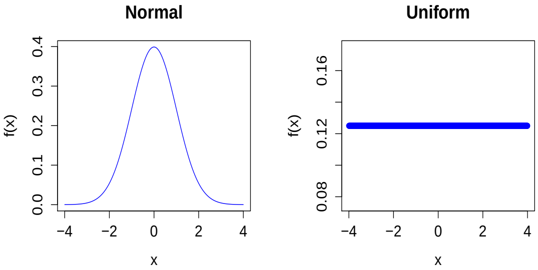
$$\begin{aligned} Cov(X + Z, Y) &= Cov(X, Y) + Cov(Z, Y) \\ Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) &= \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_j) \\ Var(\sum_{i=1}^{n} X_i) &= \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{m} Cov(X_i, Y_j) \end{aligned}$$

- X and Y are independent, then $Cov(X, Y) = 0$

Random Variables

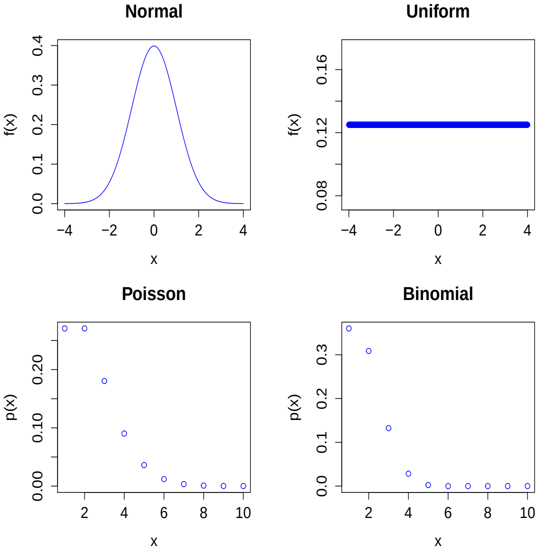


Random Variables



$\leftarrow$ Continuous distributions

Random Variables



$\leftarrow$ Continuous distributions

$\leftarrow$ Discrete distributions

Random Variables

- Distributions:
  - Discrete: Poisson, Binomial, Geometric, hyper-geometric etc.
  - Continuous: Normal and its variants,lognormal Exponential, Beta etc.

Random Variables

- Distributions:
  - Discrete: Poisson, Binomial, Geometric, hyper-geometric etc.
  - Continuous: Normal and its variants,lognormal Exponential, Beta etc.

Binomial Distribution

- Example:
  - Every minute 24 births in India, X be the number of female births=12
  - In pass and fail course, X is either pass or failure
- Experiment's outcome: Either Success or failure.
- If $X = 1$ for success, and $X = 0$ for failure,

$$P(X = 0) = 1 - p$$
$$P(X = 1) = p$$

- For n independent trails, X represents $\#$ of successes in the n trails, $X \sim b(n, p)$, $b(n, p)$ is Binomial distribution with parameters n and p.
- PMF

$$P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i},\ i = 0, 1, \ldots, n$$

Binomial Distribution

- Examples
  - Every minute 24 births in India, X be the number of female births=12
  - In pass and fail course, X is either pass or failure
- Experiment's outcome: Either Success or failure.
- If $X = 1$ for success, and $X = 0$ for failure,

$$P(X = 0) = 1 - p$$
$$P(X = 1) = p$$

- For n independent trails, X represents $\#$ of successes in the n trails, $X \sim b(n, p)$, $b(n, p)$ is Binomial distribution with parameters n and p.
- PMF

$$P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \; i = 0, 1, \ldots, n$$

- Mean and variance for X

$$\begin{aligned} E[X] &= np \\ Var[X] &= np(1 - p) \end{aligned}$$

Poisson DistributionRandom Variables

- If X is a Poisson random variable with parameter $\lambda > 0$, PMF of X
$$P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!},\ i = 0, 1, 2, \ldots, n$$
- Mean and variance of X

$$\begin{aligned} E[X] &= \lambda \\ Var[X] &= \lambda \end{aligned}$$

NormalRandom Variables Distribution

- Consider a density function $f(x)$, $-\infty < x < \infty$:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5}$$

  with parameters μ and σ.
- $X \sim \mathbb{N}(\mu, \sigma^2)$ if $f(x)$ is the density of X
- Normal distribution is symmetric about μ
- Examples
  - The height, weight of students in IIT Madras
  - The velocity of molecule in the gas
- Some results:
  - $E[Y] = E[a + b\,X] = a + bE[X] = a + b\,\mu$
  - $Var(Y) = Var(a + bX) = b^2\,Var(X) = b^2\,\sigma^2$

Normal Distribution

- Properties of normal desntiy function
  - $f(x) > 0$, and a nonempty set (a, b)

$$P(X \in (a,b) = Area_{(a,b)} f(x) \tag{6}$$

  - $f(x)$ is symmetric about μ, i.e., $f(\mu + x) = f(\mu - x)$
  - $f(x)$ decreases as $|x - \mu|$ increases
  - For all choice of *mu* and σ

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= 0.683 \\ P(\mu - 2\,\sigma < X < \mu + 2\,\sigma) &= 0.954 \\ P(\mu - 3\,\sigma < X < \mu + 3\,\sigma) &= 0.997 \end{aligned}$$

i.e., Given a probability distribution function for any one normal distribution, One can compute probability for any RV with μ and σ

Normal Distribution

- Define a random variable Z:

$$Z = \frac{X - \mu}{\sigma} \tag{7}$$

- $X \sim \mathbb{N}(\mu, \sigma^2)$, Z - distribution?
- $Z \sim \mathbb{N}(0, 1)$ (Standard normal distribution)
- Distribution function

$$\mathbf{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz \tag{8}$$

- Some results:

$$\begin{aligned} P(X < b) &= P\left(\frac{X - \mu}{\sigma} < \frac{X - \mu}{\sigma}\right) = \mathbf{\Phi}\left(\frac{b - \mu}{\sigma}\right) \\ P(a < X < b) &= \mathbf{\Phi}\left(\frac{b - \mu}{\sigma}\right) - \mathbf{\Phi}\left(\frac{a - \mu}{\sigma}\right) \end{aligned} \tag{9}$$

Normal Distribution: Chi-square distribution

- $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables
- Define a random variable P:

$$P = \sum_{i=1}^{k} Z_i^2 \tag{10}$$

- $P \sim \chi^2(k)$, χ^2 - distribution?
- Useful in hypothesis testing and other testing

Distribution of Sample statistics

- $X_1, X_2, \ldots, X_n$ random variables from $\mathcal{N}(\mu, \sigma^2)$
- Average of these RVs (mean) $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$
- Question: Distribution of $\bar{X}$ (sample stistics) ?

$$E[\bar{X}] = \mu$$
$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

- $\bar{X}$ $\mathcal{N}(\mu, \sigma^2/n)$
- $\sigma/\sqrt{n}$: Standard error in $E[\bar{X}]$ estimate
- As $n \to \infty$, $\bar{X} \to \mu$

Normal Distribution: Chi-square distribution

- $Z_1, Z_2, \ldots, Z_k$ are independent standard normal random variables
- Define a random variable P:

$$P = \sum_{i=1}^{k} Z_i^2 \tag{11}$$

- $P \sim \chi^2(k)$, χ^2 - distribution?
- Useful in hypothesis testing and other testing

Distribution of Sample statistics

- $X_1, X_2, \ldots, X_n$ random variables from $\mathcal{N}(\mu, \sigma^2)$
- Average of these RVs (mean) $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$
- Question: Distribution of $\bar{X}$ (sample stistics) ?

$$E[\bar{X}] = \mu$$
$$Var[\bar{X}] = \frac{\sigma^2}{n}$$

- $\bar{X}\ \mathcal{N}(\mu, \sigma^2/n)$
- $\sigma/\sqrt{n}$: Standard error in $E[\bar{X}]$ estimate
- As $n \to \infty$, $\bar{X} \to \mu$

Central Limit Theorem

- Interpretation: The sum of a large number of independent RV has a distribution that is approximately normal
- it captures the fact that the empirical frequencies of so many natural populations follows a normal curve
- $X_1, X_2, \ldots, X_n$ random variables , independently and identically distributed (i.i.d.) with finite μ and σ.
- Then for large n ($n \to \infty$), the distribution of $X_1 + X_2 + \ldots + X_n \to \mathcal{N}(n\mu, n\sigma^2)$
- The central limit theorem that

$$\frac{X_1 + X_2 + \ldots + X_n - n\,\mu}{\sigma\sqrt{n}}$$

is approximately a standard normal RV; thus, for n large,

$$P\left(\frac{X_1 + X_2 + \ldots + X_n - n\,\mu}{\sigma\sqrt{n}}\right) \approxeq P(Z < x)$$

where Z is a standard normal random variable.

Sample Statistics

- Central tendencies: , Sample median and Sample mode
- Spread (or variability)of data: Sample variance and sample standard deviation

Sample Statistics

- Central tendencies: Sample mean, Sample median and Sample mode
- Spread (or variability)of data: Sample variance and sample standard deviation

Sample Statistics

- Data set (n observation, numerical values): $x_1, x_2, \ldots, x_n$

Sample mean

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

- Data set(n observation, numerical values): $x_1, x_2, \ldots, x_k$
  corresponding frequencies $f_1, f_2, \ldots, f_k$

Sample mean

$$\bar{x} = \sum_{i=1}^{k} \frac{f_i\, x_i}{n}$$

Sample Statistics

- Salary data of a SME company

Employ	Salary/month (Rs.)
1	20,000
2	15,000
3	12,000
4	35,000
5	8,000
6	11,000
7	20,000
8	25,000
9	1,20,000
10	3,15,000

Sample Statistics

- Salary data of a SME company

Employ	Salary/month (Rs.)
1	20,000
2	15,000
3	12,000
4	35,000
5	8,000
6	11,000
7	20,000
8	25,000
9	1,20,000
10	3,15,000

- Mean Salary: Rs. 53909.09

Sample Statistics

- Salary data of a SME company

Employ	Salary/month (Rs.)
1	20,000
2	15,000
3	12,000
4	35,000
5	8,000
6	11,000
7	20,000
8	25,000
9	1,20,000
10	3,15,000

- Mean Salary: Rs. 53909.09
- Mean Salary: Rs. 17555.56 (after removing Employ No. 9 and 10)

Sample Statistics

- Salary data of a SME company

Employ	Salary/month (Rs.)
1	20,000
2	15,000
3	12,000
4	35,000
5	8,000
6	11,000
7	20,000
8	25,000
9	1,20,000
10	3,15,000
11	12,000

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

- Middle Salary: Rs. 20,000
- Most Common salary?

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

- Most Common salary?
- Common Salaries: Rs. 12,000; Rs. 20,000

Sample Statistics, Cont.

- Data set (n ordered observation from smallest to largest, numerical values): $x_1, x_2, \ldots, x_n$

Sample median

If n: odd, $x_{md} = (n+1)/2$ numerical value

If n: even, $x_{md} = \frac{(n/2)\text{value} + (n/2+1)\text{value}}{2}$

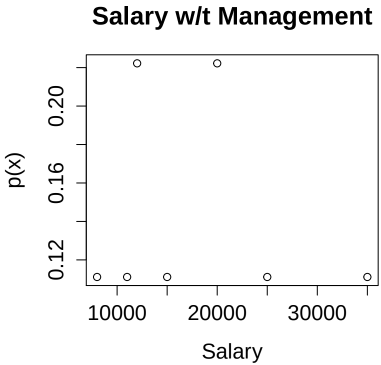
- Data Set (n observations,numerical values): $x_1, x_2, \ldots, x_k$ corresponding frequencies $f_1, f_2, \ldots, f_k$

Sample mode

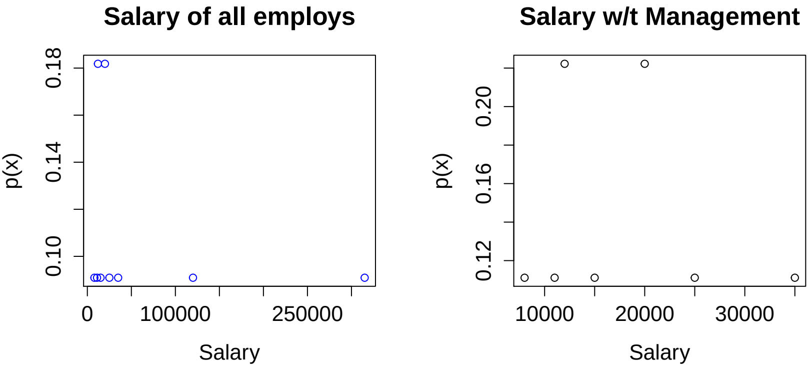
$x_{mode} = x_j$ with f_j is the greatest frequency

- Note: No single value occurs most frequently, i.e. the greatest frequency = $f_j = f_l = \ldots = f_p$, then $x_{mode} = x_j = x_l = \ldots = x_p$.

Mean, Median, and Mode



Mean, Median, and Mode



A rough Guideline

Measurement Scale	Best Measure of the "Middle"
Nominal (Categorical)	Mode
Symmetrical data	Mean
Skewed data	Median

Sample Statistics, Cont.

- Data Set (n observations,numerical values): $x_1, x_2, \ldots, x_n$

Sample variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{or}$$

$$\sigma^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2\right)$$

- Standard deviation: +ve square root of sample variance

SD

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

- Standard deviations: Rs. 92198.11

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

- Standard deviations: Rs. 92198.11
- Standard deviations (w/t Employs 9,10): Rs. 8472.18

Sample Statistics

- Salary data of a SME company

Employ	Ascending Order	Salary/month (Rs.)
5	1	8,000
6	2	11,000
3	3	12,000
11	4	12,000
2	5	15,000
1	6	20,000
7	7	20,000
8	8	25,000
4	8	35,000
9	10	1,20,000
10	11	3,15,000

- Standard deviations: Rs. 92198.11
- Standard deviations (w/t Employs 9,10): Rs. 8472.18

Sample Statistics, Cont.

- Data set: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Question: Any statistics to determine a relationship between x_i and y_i?
- Answer: Sample correlation coefficient (r_{xy})

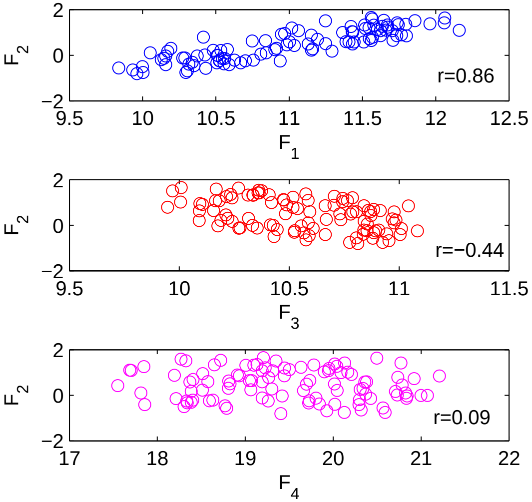
Sample variance

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{(n-1)\sigma_x \sigma_y}$$

- $r_{xy} > 0 \rightarrow$ +ve correlation, $r_{xy} > 0 \rightarrow$ -ve correlation
- $|r_{xy}$= measure of the strength of the linear relationship between x and y variables

Sample Statistics, Cont.

- Four flow variables: F_1, F_2, F_3, F_4



Percentiles

- Sample $100p$, $0 \leq p \leq 1$ percentile:
  100p percent of the data values are less or equal to it

Example

Course grades: 10% E, 16% D, 24% C, 32% B, 15% A, 5% S
I got B grade, what is percentile for A?

- Computation of percentile:
  - For Group data: **Add up all percentages below the particular group, plus half the percentage at the group**
  - For a data set of size n, $100p$ percentile:
    - At least np of the values are less than equal to it
    - If two data values satisfy this condition, then the arithmetic average of these two values

Quartiles

- Values that divide a list of numbers into quarters:
- Computation of Quantiles:
  - Put the list of numbers in order
  - Then cut the list into four equal parts
  - The Quartiles are at the "cuts"

Example

6, 7, 4, 5, 4, 2, 8
Order them: 2,4,4,5,6,7,8
Quartile 1 (Q1) = 4
Quartile 2 (Q2) = 5
Quartile 3 (Q3) = 7