CS57300
PURDUE UNIVERSITY

JANUARY 10, 2019

DATA MINING

ANNOUNCEMENTS

- TA office hours (Location: HAAS G50):
  - Mahak Goindani (Monday 6-7pm)
  - Omkar Patil (Thursday 1-2pm)
  - Hao Ding (Friday 3-4pm)
- Assignment 1 is out! **Due time: January 20 (Sunday) 11:59pm**
  - You can not apply any extension days on this assignment!
  - Please complete this assignment independently!

PROBABILITY AND STATISTICS BASICS

MODELING UNCERTAINTY

- Necessary component of almost all data analysis
- Approaches to modeling uncertainty:
  - Fuzzy logic: form of many-valued logic that reasons with partial truth values
  - Possibility theory: reasons about the possibility and necessity of events to deal with incomplete information
  - Rough sets: represents imperfect knowledge via upper and lower bounds on "certain" information
  - **Probability (focus in this course)**

PROBABILITY

- Probability theory (some disagreement)
  - Concerned with interpretation of probability
  - 17th century: Pascal and Fermat develop probability theory to analyze games of chance
- Probability calculus (universal agreement)
  - Concerned with manipulation of mathematical representations
  - 1933: Kolmogorov states axioms of modern probability

PROBABILITY BASICS

- Basic element: **Random variable (RV)**
  - A variable whose possible values are outcomes of a random phenomenon
  - *X* refers to random variable; *x* refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite set of possible values
    - e.g., Is there a storm warning $\in$ {Yes, No}
    - e.g., Tomorrow's weather $\in$ {sunny, rainy, cloudy, snow}
  - Continuous RV can take any value within an interval
    - e.g., Temperature

PROBABILITY BASICS

- **Sample space (S)**
  - Set of all possible outcomes of the random phenomenon
- **Event**
  - Any subset of outcomes contained in the sample space S
  - When events A and B have no outcomes in common they are said to be *mutually exclusive*

Random variable(s)	Sample space	Example event
Two coin tosses	HH, HT, TH, TT	At least one H
Select one card	2♥, 2♦, …, A♣ (52)	A card of hearts

Probability of 1 H
and a TT are mutually exclusive

A card of H and a black card

Q: Think of some mutually exclusive events of the above example events?

AXIOMS OF PROBABILITY

- For a sample space **S** with possible events $\mathbf{A_S}$:

A function that associates real values with each event A is called a *probability function* if the following properties are satisfied:

1. $0 \leq P(A) \leq 1$ for every A
2. $P(\mathbf{S}) = 1$
3. $P(A_1 \vee A_2 \dots \vee A_{n \in S}) = P(A_1) + P(A_2) + \dots + P(A_n)$

   if $A_1, A_2, \dots, A_n$ are pairwise mutually exclusive events

INTERPRETING PROBABILITIES

- Meaning of probability is focus of debate and controversy
- Two main views: Frequentist and Bayesian

FREQUENTIST VIEW

- Dominant perspective for last century
- Probability is an **objective** concept
  - Defined as the frequency of an event occurring under repeated trials in "same" situation

CALCULATING PROBABILITIES: FREQUENTIST

- Repeated experiments
  - Let n be the number of times an experiment is performed
  - Let n(A) be the number of outcomes in which A occurs
  - Then as $n \to \infty$ P(A) = n(A) / n
- When the various outcomes of an experiment are equally likely (e.g., toss a fair die), the task of computing probability reduces to counting
  - Let N be size of sample space (i.e., number of simple outcomes)
  - Let N(A) be the number of outcomes contained in A
  - Then: P(A) = N(A) / N
- Limitation: Restricts application of probability to repeatable experiment
  - What's the probability of Trump being re-elected in 2020?

BAYESIAN VIEW

- Increasing importance over last decade
  - Due to increase in computational power that facilitates previously intractable calculations
- Probability is a **subjective** concept
  - Defined as individual degree-of-belief that event will occur
  - E.g., belief that we will have a rainy day tomorrow
- Observed data helps us to update and inform our prior beliefs

CALCULATING PROBABILITIES: BAYESIAN

- Begin with *prior* belief estimates: **P(A)**
    - E.g., Bob believed that the chance of raining tomorrow is 0.2
      P(rainy)=0.2
- Update belief by conditioning on observed data through Bayes' theorem
  P(A|data) = P(data|A) P(A) / P(data)
    - But then Bob observed a storm warning on the weather channel. In the past the storm warning appeared on 40% of rainy days. Overall a storm warning was given on 1 out of 8 days. Bob updated his belief on the chance of raining tomorrow:
      P(rainy|warning) = 0.4 * 0.2 / 0.125 = 0.64
- Even when the same data is observed, if people have different priors, they can end up with different posterior probability estimates P(A|data)

PROBABILITY DISTRIBUTION

- **Probability distribution** (i.e., probability mass function or probability density function) specifies the probability of observing every possible value of a random variable
- Discrete
  - Denotes probability that *X* will take on a particular value:

$$P(X = x)$$

- Continuous
  - Probability of any particular point is 0, have to consider probability within an interval:

$$P(a < X < b) = \int_a^b p(x)dx$$

JOINT PROBABILITY

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables

  E.g., P(Weather, Warning) = a 4 × 2 matrix of values:

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

- *Every question about events can be answered by the joint distribution*

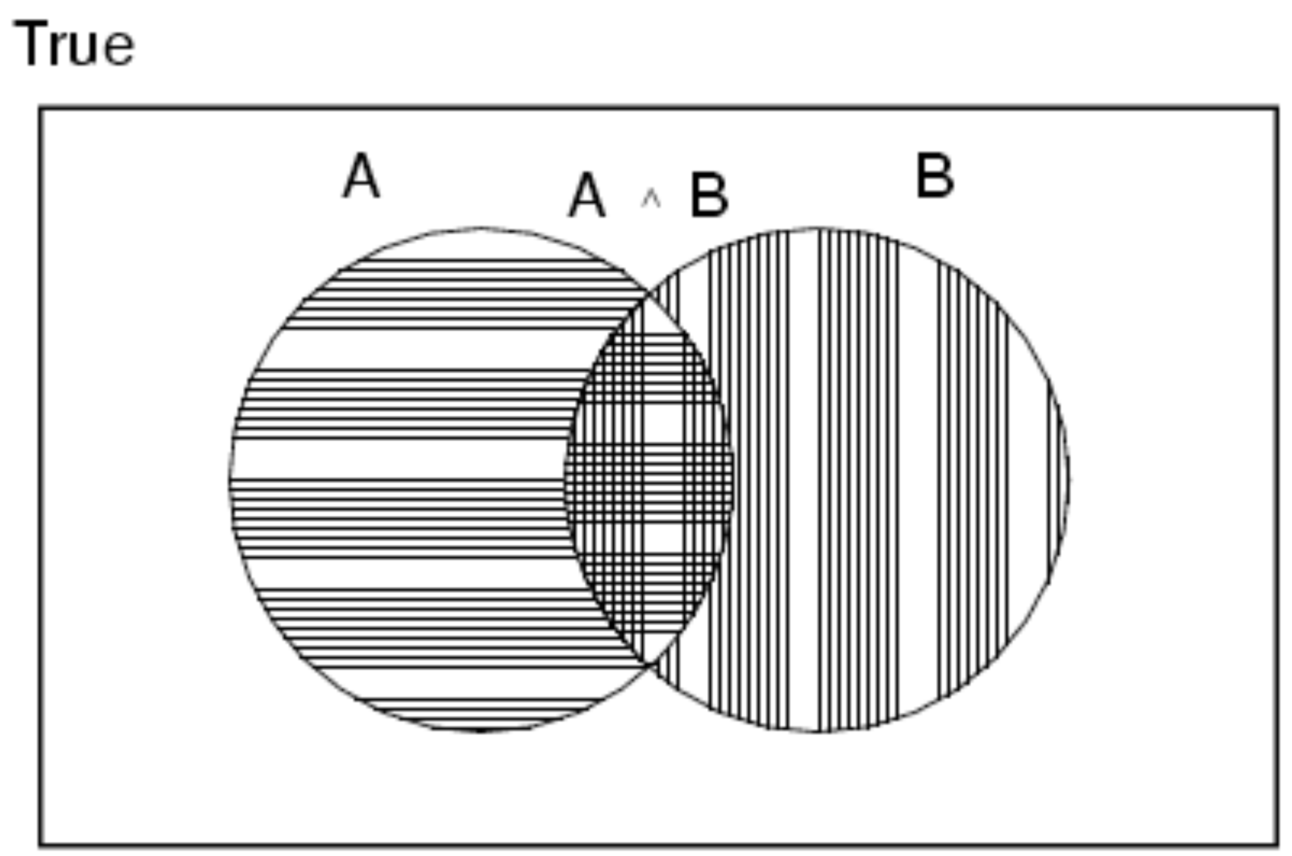
CONDITIONAL PROBABILITY

- **Conditional** (or posterior) probability: The probability of an event given that another event has happened
  - e.g., P(warning=Y | snow=T) = 0.4
  - Complete conditional distributions:
    P(warning | snow) =
    {P(warning = Y | snow = T), P(warning = N | snow = T)},
    {P(warning = Y | snow = F), P(warning = N | snow = F)}
- If we know more, then we can update the probability by conditioning on more evidence
  - e.g., if windy is also given then P(warning=Y | snow=T, windy=T) = 0.5

CONDITIONAL PROBABILITY

- Definition of conditional probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \qquad \text{if } P(B) > 0$$



- **Product rule** gives an alternative formulation:

$$\begin{aligned} P(A \wedge B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

- **Chain rule** is derived by successive application of product rule:

$$\begin{aligned} P(X_1, \ldots, X_n) &= P(X_n|X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1}) \\ &= P(X_n|X_1, \ldots, X_{n-1})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2}) \\ &= \ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

MARGINAL PROBABILITY

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events
- Marginalization:

$$P(A) = \sum_{b \in B} P(A, b)$$
$$= \sum_{b \in B} P(A|b)P(b)$$

- Example: What is P(cloudy)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

INDEPENDENCE

- Two variables A and B are independent if knowing B tells you nothing about A and vice versa:

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A)\,P(B)$$

- Two variables *A* and *B* are **conditionally** independent given *Z* iff for all values of *A, B, Z*:

$$P(A, B \mid Z) = P(A \mid Z)\,P(B \mid Z) \quad \text{or} \quad P(A \mid B, Z) = P(A \mid Z)$$

- *Note: independence does not imply conditional independence or vice versa*

EXAMPLE 1

- **Conditional independence does not imply independence**
- Gender and lung cancer are not independent

  $$P(C \mid G) \neq P(C)$$

- Gender and lung cancer are conditionally independent given smoking

  $$P(C \mid G, S) = P(C \mid S)$$

- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

EXAMPLE 2

- **Independence does not imply conditional independence**
- Sprinkler-on and raining are independent
  $P(S \mid R) = P(S)$
- Sprinkler-on and raining are not conditionally independent given grass is wet
  $P(S \mid R, W) \neq P(S \mid R)$
- Why? Because once we know the grass is wet, if it’s not raining, then the explanation for the grass being wet has to be the sprinkler

Ref material - https://en.wikipedia.org/wiki/Joint_probability_distribution

MULTIVARIATE RV

while a given person has a specific age, height and weight, the representation of these features of an unspecified person from within a group would be a random vector. Normally each element of a random vector is a real number

- A multivariate random variable $\boldsymbol{X}$ is a set $X_1, X_2, \ldots X_p$ of random variables
- **Joint** density function: $P(\mathbf{x})=P(x_1,x_2,\ldots,x_p)$
- **Marginal** density function: the density of any subset of the complete set of variables, e.g.,:

$$P(x_1) = \sum_{x_2, x_3} p(x_1, x_2, x_3)$$

e

- **Conditional** density function: the density of a subset conditioned on particular values of the others, e.g.,:

$$P(x_1|x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)}$$

EXPECTATION

- Denotes the expected value or mean value of a random variable X
- Discrete

$$E[X] = \sum_x x \cdot p(x)$$

- Continuous

$$E[X] = \int_x x \cdot p(x) dx$$

- Expectation of a function

$$E[h(X)] = \sum_x h(x) \cdot p(x)$$

$$E[aX + b] = a \cdot E[X] + b$$

- Linearity of expectation

$$E[X + Y] = E[X] + E[Y]$$

VARIANCE

- Denotes the squared deviation of X from its mean
- Variance

$$Var(X) = E[(x - E[X])^2]$$
$$= E[X^2] - (E[X])^2$$

Expand (x-E[X])^2

- Standard deviation

$$\sigma = \sqrt{Var(X)}$$

- Variance of a function

E[(aX+b-E[ax+b])^2]
E[(aX-aE[x])^2]
a^2E[X-E[X]^2]

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)$$

COMMON DISTRIBUTIONS

- Bernoulli
- Binomial
- Multinomial
- Poisson
- Normal

BERNOULLI

- Binary variable (0/1) that takes the value of 1 with probability p
  - E.g., Outcome of a fair coin toss is Bernoulli with *p=0.5*

$$P(x) = p^x (1-p)^{1-x}$$

$$E[X] = 1(p) + 0(1-p) = p$$

$$\begin{aligned} Var(X) &= E[X^2] - (E[X])^2 \\ &= 1^2(p) + 0^2(1-p) - p^2 \\ &= p(1-p) \end{aligned}$$

BINOMIAL

- Describes the number of successful outcomes in *n* independent Bernoulli(p) trials

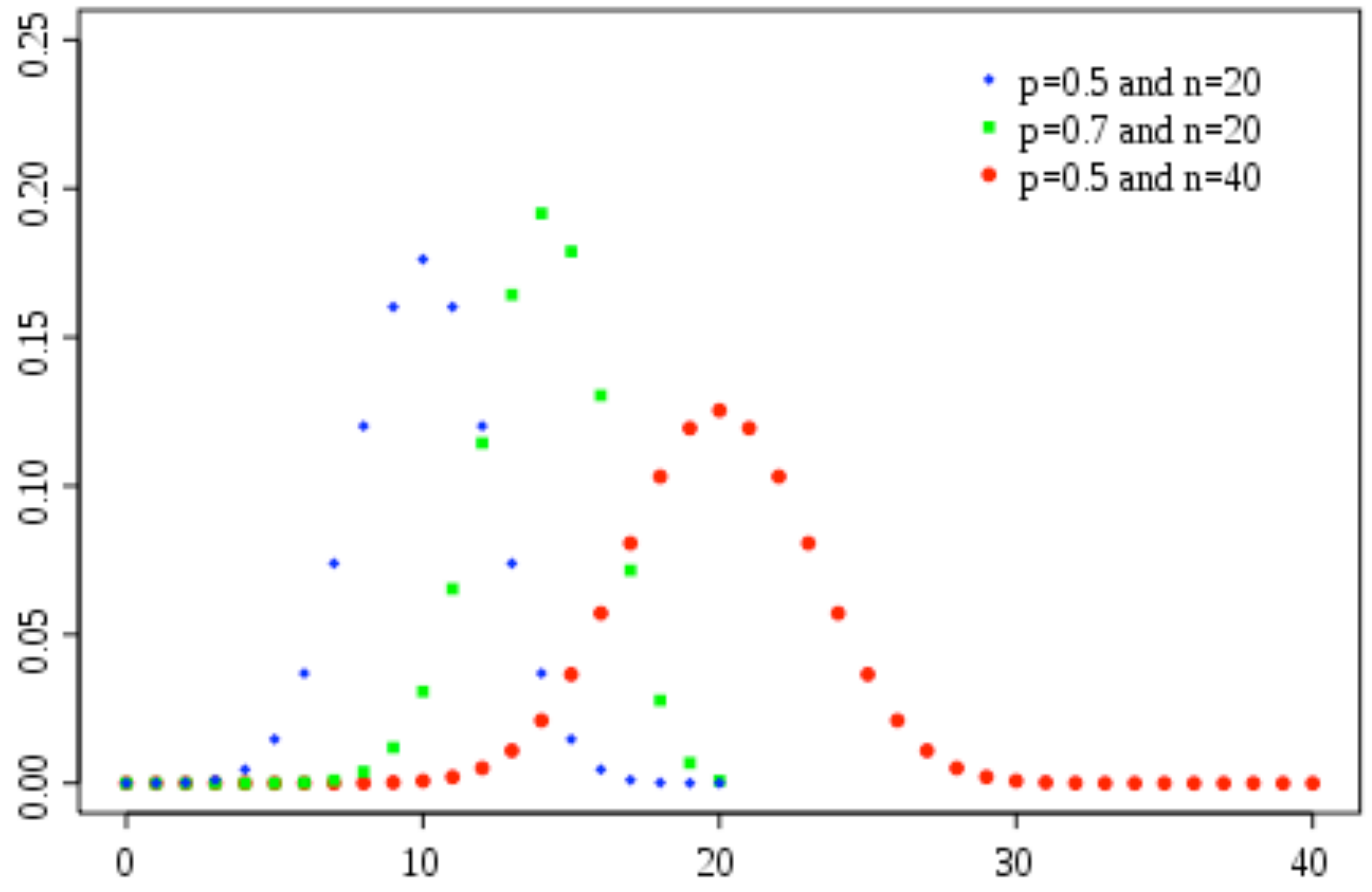
ref:https://mourafiq.com/2016/02/01/intuition-behind-binomial-distribution.html

  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with *n=10* and *p=0.5*

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

$$E[X] = np$$

$$Var[X] = np(1-p)$$



MULTINOMIAL

- Generalization of binomial to *k* possible outcomes; outcome *i* has probability p_i of occurring
  - E.g., Number of {diamonds, clubs, hearts, spades} in a sequence of 10 random draw of cards (with replacement) is multinomial
- Let X_i denote the number of times the *i*-th outcome occurs in *n* trials:

$$P(x_1, ...x_k) = \binom{n}{x_1, ...x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

$$E[X_i] = np_i$$

$$Var(X_i) = np_i(1 - p_i)$$

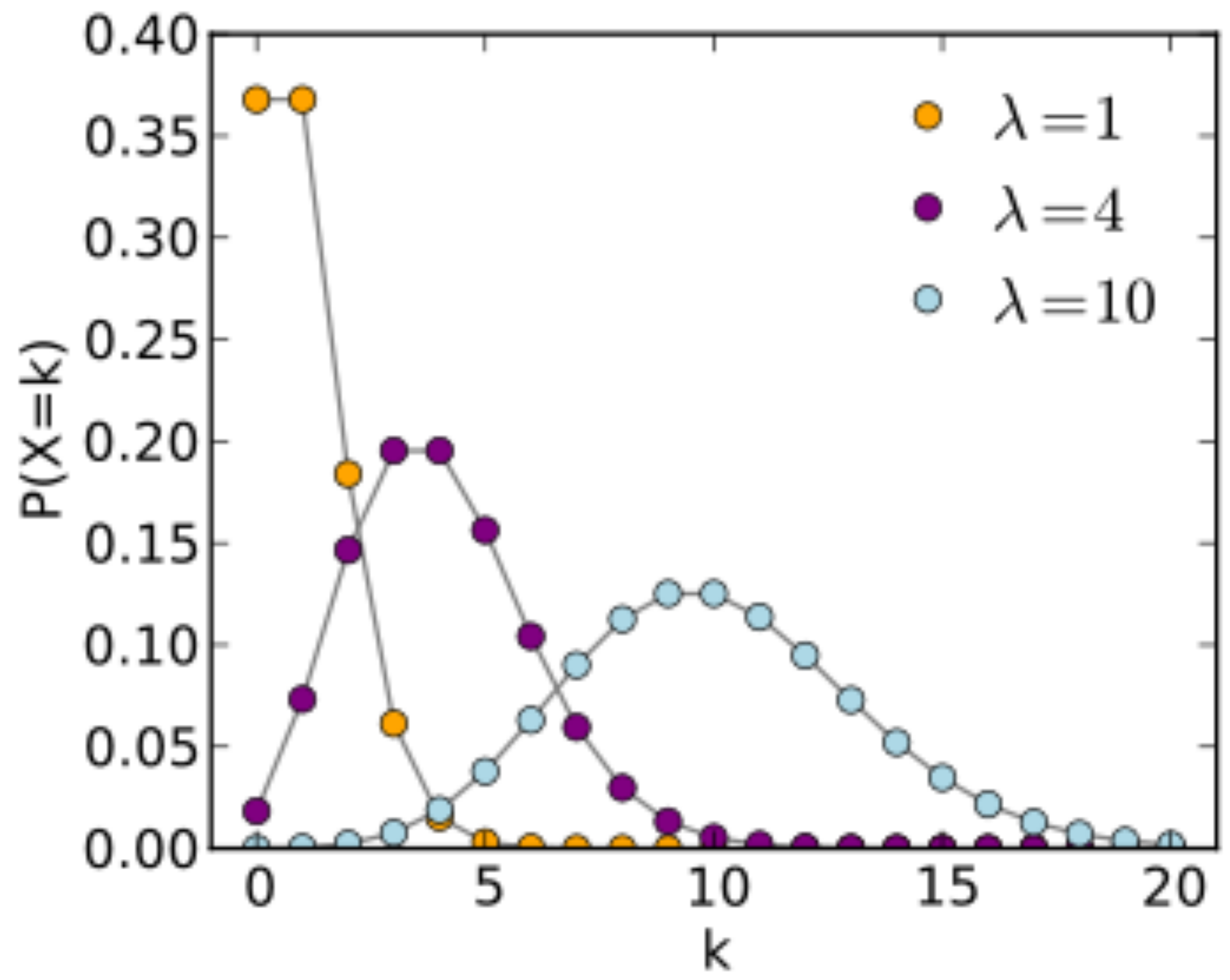
POISSON

https://mourafiq.com/2016/02/05/intuition-behind-poisson-distribution.html

- Describes the probability of a given number of events occurring in a fixed interval of time (or space), given an average arrival rate (λ) and independent events that occur randomly over time (or space)
  - E.g., Given an average of 4 power failures per winter, what is the probability that there will be more than 7 failures this winter?

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

$$\lambda = E[X] = Var[X]$$



NORMAL (GAUSSIAN)

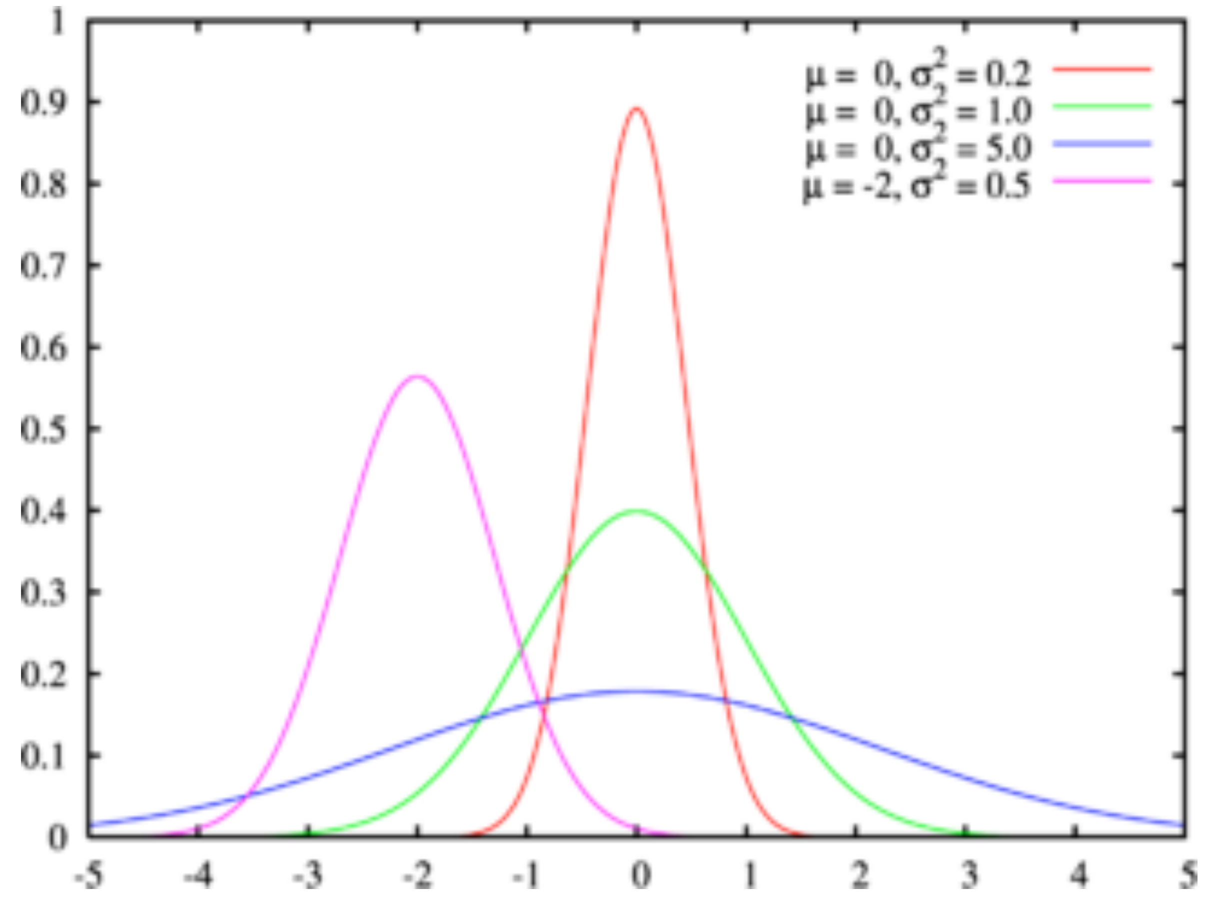
https://towardsdatascience.com/understanding-the-central-limit-theorem-642473c63ad8

- Important distribution gives well-known bell shape
- Central limit theorem:
  - Distribution of the mean of *n* samples becomes normally distributed as *n* ↑, regardless of the distribution of the underlying population

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$



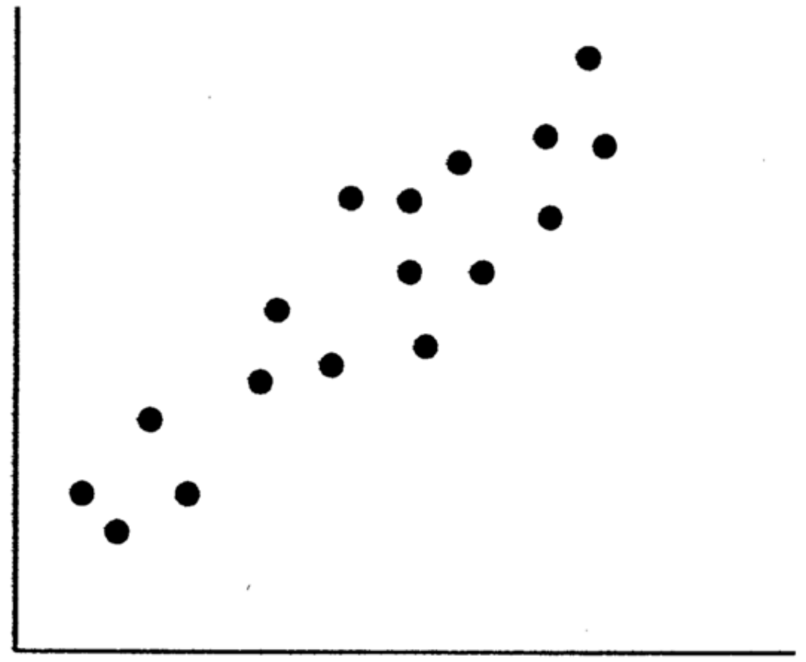
COVARIANCE AND CORRELATION

COVARIANCE

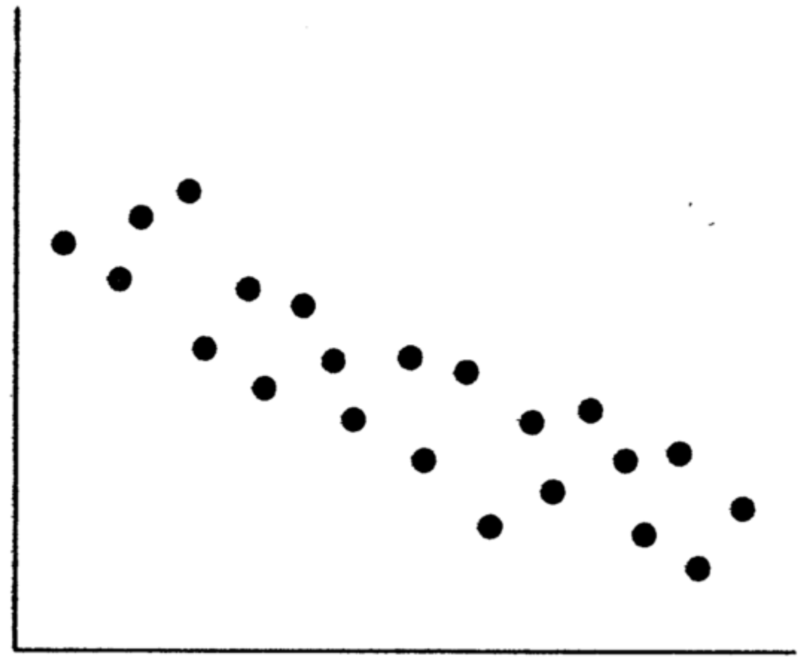
- Measures how variables X and Y vary together:

$$COV(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= E[XY] - E[X]E[Y]$$

  - Positive if large values of X are associated with large values of Y
  - Negative if large values of X are associated with small values of Y



(a)



(b)

Measures **linear** relationship

COVARIANCE

- For discrete random variable pair (X, Y) that can take on the values of (x_i, y_i) for i=1, ..., n with equal probabilities 1/n:

$$COV(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E[X])(y_i - E[Y])$$

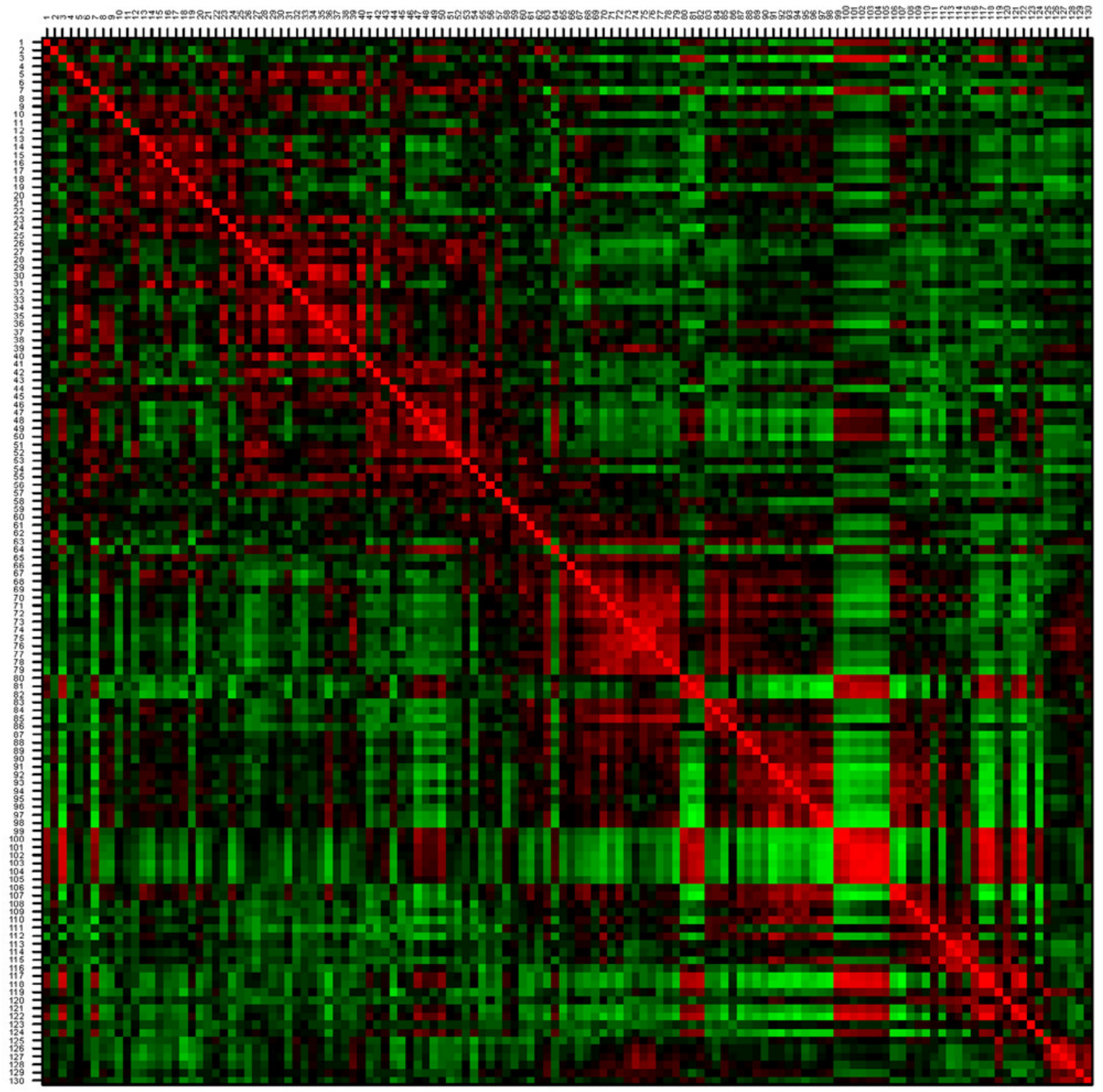
- Covariance matrix (Σ)
  - Symmetric matrix of covariances for p variables
  - Σ_{ij} = COV(X_i, X_j)

CORRELATION COEFFICIENT

- Covariance depends on ranges of X_j and X_k
- Correlation standardizes covariance by dividing through standard deviations

$$\rho(X_j, X_k) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{X}_j\right)\left(x_{ik} - \bar{X}_k\right)}{\sigma_{X_j}\sigma_{X_k}}$$

- Correlation matrix
  - Symmetric matrix of correlations for *p* variables
  - What values are on the diagonal?



NEXT CLASS

- Review basic knowledge on linear algebra