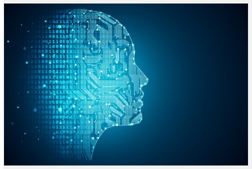


Data Science for Researchers and Scholars

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Random Variables

- A variable x is a random variable if the value that it assumes, corresponding to the outcome of an experiment is a chance or random event.
- Random variables can be discrete or continuous

Examples

- x = SAT score for a randomly selected student
- x = number of people who click on your website on a randomly chosen of the year 2023
- x = outcome of a die toss

Probability Distributions of Discrete Random Variables

- The probability distribution for a discrete random variable *x* is a graph, table or formula that gives the probability $p(x)$ associated with each value of x
- Note that
  - $\forall x \; 0 \leq p(x) \leq 1$
  - $\sum_x p(x) = 1$

Example

Toss a fair coin three times and define x = number of heads.

		x
HHH	1/8	3
HHT	1/8	2
HTH	1/8	2
THH	1/8	2
HTT	1/8	1
THT	1/8	1
TTH	1/8	1
TTT	1/8	0

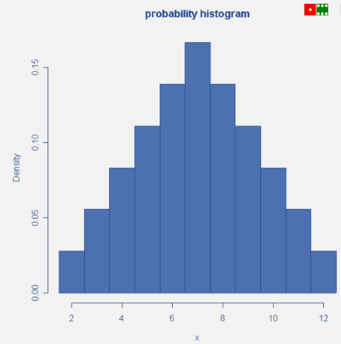
$P(x = 0) = 1/8$
$P(x = 1) = 3/8$
$P(x = 2) = 3/8$
$P(x = 3) = 1/8$

x	$p(x)$
0	1/8
1	3/8
2	3/8
3	1/8

Probability distribution of x

Example

Toss two dice and define
x = sum of two dice.



x	$p(x)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Probability Distributions

- Probability distributions can be used to describe the population, just as we described samples using statistics
- Shape: Symmetric, skewed, mound-shaped…
  - Outliers: unusual or unlikely measurements
  - Center and spread: mean and standard deviation. A population mean is called μ and a population standard deviation is called σ.
- Let x be a discrete random variable with probability distribution $p(x)$. Then the mean, variance and standard deviation of x are given as

$$\text{Mean}: \mu = \sum x p(x)$$
$$\text{Variance}: \sigma^2 = \sum (x-\mu)^2 p(x)$$
$$\text{Standard deviation}: \sigma = \sqrt{\sigma^2}$$

Example

Toss a fair coin 3 times and record x, the number of heads.

x	$p(x)$	$xp(x)$	$(x-\mu)^2 p(x)$
0	1/8	0	$(-1.5)^2(1/8)$
1	3/8	3/8	$(-0.5)^2(3/8)$
2	3/8	6/8	$(0.5)^2(3/8)$
3	1/8	3/8	$(1.5)^2(1/8)$

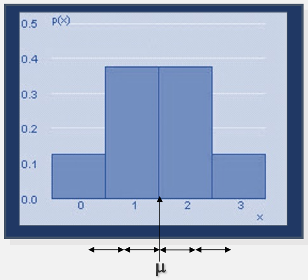
$$\mu = \sum xp(x) = \frac{12}{8} = 1.5$$

$$\sigma^2 = \sum (x-\mu)^2 p(x)$$

$$\sigma^2 = .28125 + .09375 + .09375 + .28125 = .75$$
$$\sigma = \sqrt{.75} = .688$$

Example

The probability distribution for *x* the number of heads in tossing 3 fair coins.



- Shape? – Symmetric
- Outliers? - None
- Center? $\mu = 1.5$
- Spread? $\sigma = .688$

Random Variable and Distribution

- A random variable X is an outcome of a random experiment
- The distribution of a random variable is a table, graph or a formula that gives the probability $P(X)$ associated with each possible value of X
  - In the case of discrete random variables, we write

$$P(X = x) = p_\theta(x)$$

  to mean probability that the random variable X takes the value x is $p_\theta(x)$, a function of x, parameterized by θ.
  - $\forall x \; 0 \leq p_\theta(x) \leq 1$
  - $\sum_x p_\theta(x) = 1$

Bernoulli distribution

- Bernoulli distribution is a discrete probability distribution
- It escribes the probability of achieving a "success" or "failure" from a random experiment (called Bernoulli trial) with only two possible outcomes (success or failure).
- Example: outcome of coin toss with two outcomes – heads (success denoted by 1) or tails (denoted by 0)

$$P(X = x) = \theta^x (1-\theta)^{1-x}$$

  - When $x = 1$, we have
    $P(X = 1) = \theta^1 (1-\theta)^{1-1} = \theta$
  - When $x = 0$, we have
    $P(X = 0) = \theta^0 (1-\theta)^{1-0} = 1 - \theta$
  - Note that $\forall \theta$ such that $0 \leq \theta \leq 1$,
    - $\forall x \; 0 \leq p_\theta(x) \leq 1$ and
    - $\sum_x p_\theta(x) = \theta + 1 - \theta = 1$

Exercise: Mean and variance of Bernoulli distribution

Exercise: Mean and variance of Bernoulli distribution

$$P(X=x)=\theta^x(1-\theta)^{1-x}$$

- Mean = expectation of x

$$\mu=\sum_x xP(X=x)=1(\theta)+0(1-\theta)=\theta$$

- Variance = expectation of the square of the difference between x and the mean of x

$$\sigma^2=\sum_x (x-\mu)^2 P(X=x)$$

$$\sigma^2 \;=(1-\mu)^2\theta+(0-\mu)^2(1-\theta)$$

$$\sigma^2 \;=\theta-\mu^2\theta-\mu^2+\mu^2\theta=\theta-\theta^2=\theta(1-\theta)$$

Categorical distribution generalizes Bernoulli distribution

- Instead of 2 outcomes, now we have k discrete outcomes $1, 2, \cdots k$ that occur with probabilities $p_1, p_2, \cdots p_k$
- Example: outcome of k-sided die toss

$$P(X = x) = p_1^{\mathrm{I}(x=1)}\, p_2^{\mathrm{I}(x=2)} \cdots p_k^{\mathrm{I}(x=k)}$$

where

$\mathrm{I}(x = v)$=1 iff $x = v$ and $\mathrm{I}(x = v)$=0 otherwise

Note that

$P(X = 1) = p_1$, $P(X = 2) = p_2$,$\cdots$ $P(X = k) = p_k$ as desired

We further require that $\forall k\ 0 \leq p_k \leq 1$ and $\sum_{v=1}^{k} p_v = 1$

Categorical distribution

- A convenient way to represent the outcome of a categorical random experiment is one hot encoding, a k-element vector with a 1 in the position corresponding to the observed outcome and 0s everywhere else.
  - Outcome $X = 1 = x_1$is encoded as $\mathbf{v}_1 \; = [1,0,0, \cdots 0]$
  - Outcome $X = 2 = x_2$ is encoded as $\mathbf{v}_2 \; = [0,1,0, \cdots 0]$ …
  - Outcome $X = k = x_k$ is denoted by $\mathbf{v}_k \; = [0,0,0, \cdots k]$
- Now
  - $\mathbf{v}_1$ occurs with probability p_1
  - $\mathbf{v}_2$ occurs with probability p_2 …
  - $\mathbf{v}_\mathrm{k}$ occurs with probability p_k
- The outcomes of the categorical random variable X have a 1-1 correspondence with one-hot vector valued random variable $\mathbf{V}$
- One hot encoding offers many conveniences
  - As an exercise, compute the mean of the categorical distribution with
    - Scalar discrete representation of the outcomes
    - One hot encoding of the outcomes

Categorical distribution

- A convenient way to represent the outcome of a categorical random experiment is one hot encoding, a k-element vector with a 1 in the position corresponding to the observed outcome and 0s everywhere else.
  - Outcome $X = 1 = x_1$is encoded as $\mathbf{v}_1 \ = [1,0,0, \cdots 0]$
  - Outcome $X = 2 = x_2$ is encoded as $\mathbf{v}_2 \ = [0,1,0, \cdots 0]$ …
  - Outcome $X = k = x_k$ is denoted by $\mathbf{v}_k \ = [0,0,0, \cdots k]$
- Now
  - $\mathbf{v}_1$ occurs with probability p_1
  - $\mathbf{v}_2$ occurs with probability p_2 …
  - $\mathbf{v}_\mathrm{k}$ occurs with probability p_k
- The outcomes of the categorical random variable X have a 1-1 correspondence with one-hot vector valued random variable $\mathbf{V}$

One hot encoding of categorical outcomes

x → One Hot Encoder → $\mathbf{v}$

Example: 6-sided die

3 → One Hot Encoder → $\begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}$

- The correspondence between x and $\mathbf{v}$ is a bijection
- x and $\mathbf{v}$ contain the identical information
  - Outcome of the categorical random experiment
  - Example, outcome of tossing a 6-sided die

Mean and variance of Categorical distribution

Discrete scalar representation of outcomes

$$P(X = x) = p_1^{\mathrm{I}(x=1)}\, p_2^{\mathrm{I}(x=2)} \cdots\, p_k^{\mathrm{I}(x=k)}$$

- Mean = expectation of X
- $\mu = \sum_i \; x_i\, P(X = x_i) = 1p_1 + 2\, p_2 + \cdots + kp_k$

One hot vector representation of outcomes

$\forall i \in \{1, \cdots k\}, P(\mathbf{V} = \mathrm{v}_i) = p_i$

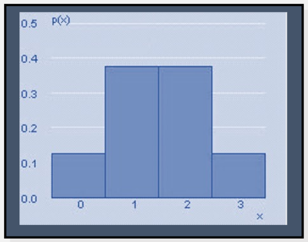
- Mean = expectation of $\mathbf{V}$
- $\boldsymbol{\mu} = \sum_i \; \mathrm{v}_i P(\mathbf{V} = \mathrm{v}_i) = \sum_i \; \mathrm{v}_i p_i = [p_1, p_2, \cdots p_k]$

- One hot encoding is elegant and offers many conveniences
- We will use it often in machine learning

The Binomial Random Variable

- Binomial random variable generalizes the Bernoulli variable
- Bernoulli – Toss a coin once and record the outcome
- Toss a coin *n* times and record x = number of heads

Binomial distribution of the number of heads in 3 tosses of a fair coin



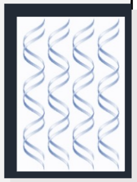
x	$p(x)$
0	1/8
1	3/8
2	3/8
3	1/8

Bernoulli versus Binomial

- The Bernoulli distribution represents the success or failure of a single Bernoulli trial.
- The Binomial Distribution represents the number of successes and failures in n independent Bernoulli trials for some given value of n.

The Binomial Random Variable

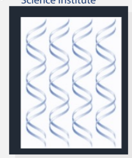
- Many situations in real life resemble the coin toss, but the coin is not necessarily fair, so that P(H) $\neq$ 1/2.
- Example: A geneticist samples 10 people and counts the number who have APOE-e4 a gene linked to Alzheimer's disease.
- Coin: Person
- Head: Has one or more copies of APOE-e4 gene
- Tail: Has no copy of APOE-e4 gene
- Number of coin tosses: $n = 10$
- P(Has Alzheimer's gene) = P(H) = fraction of the population that has at least 1 copy of the APOE-e4 gene ≈ 0.2 to 0.3



The Binomial Experiment

- The experiment consists of n identical trials.
- Each trial results in one of two outcomes, success (S) or failure (F).
  - The probability of success (or failure) on a single trial is p and the probability of failure is $q = 1 - p$.
- The probabilities p (and hence q) remain constant from trial to trial.
- The trials are independent.
- We are interested in x, the number of successes in n trials.

Binomial or Not?



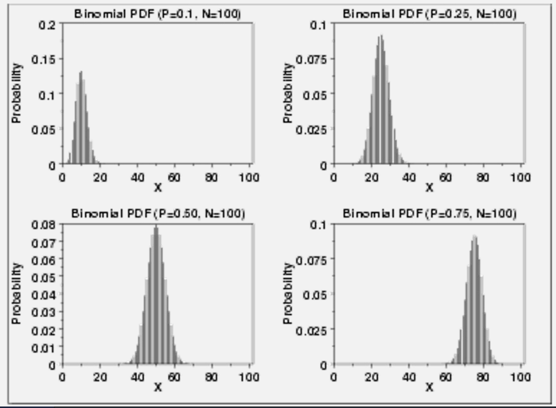
- Binomial distribution requires that the trials be independent
- Independence is often violated in real life applications

- Select two people from the U.S. population, and suppose that 20% of the population has the APOE-e4 Alzheimer's gene.
  - For the first person, $p = \text{P(gene)} = 0.20$
  - For the second person, $p \approx \text{P(gene)} = .20$, even though one person has been removed from the population.

Binomial or Not?

- 1 in 10 PCs are defective.
- We have 20 PCs in the lab
- We randomly select 3 for testing.
- Is this a binomial experiment?
  - The experiment consists of $n = 3$ identical trials
  - Each trial results in one of two outcomes
  - The probability of success (finding the defective PC) is 0.1 and it remains constant across trials
  - But there is a catch.
    - The trials are not independent.
    - P(success on the 2nd trial | success on the 1st trial) = $1/19$, not $2/20$
- Rule of thumb: if the sample size n is large relative to the population size N, say $n/N \geq .05$, the trials are likely not independent and the experiment not likely binomial.

Plots of Binomial Distribution



The Binomial Probability Distribution

- For a binomial experiment with n trials and probability p of success on any single trial, the probability of k successes in n trials is

Number of ways to choose k out of n items

$$P(x=k) = C_k^n p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} \text{ for } k = 0,1,2,\ldots n.$$

Recall $C_k^n = \dfrac{n!}{k!(n-k)!}$

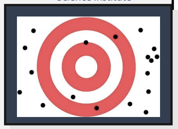
with $n! = n(n-1)(n-2)\ldots(2)1$ and $0! \equiv 1$.

Mean and Standard Deviation: Binomial Distribution

Exercise: For a binomial experiment with n trials and probability p of success on a given trial, show that

- Mean $\mu = np$
- Variance $\sigma^2 = npq$
- Standard deviation $\sigma = \sqrt{npq}$

Example



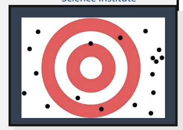
- Ukrainian missiles hit a target 80% of the time.
- The Ukrainian forces fire five missiles at a target.
- What is the probability that exactly 3 missiles hit the target?

$n = 5$ $\quad$ *success* = hit $\quad$ $p = .8$ $\quad$ $x =$ # of hits

$$P(x=3) = C_3^n p^3 q^{n-3} = \frac{5!}{3!2!}(.8)^3(.2)^{5-3}$$

$$= 10(.8)^3(.2)^2 = .2048$$

Example (Contd)



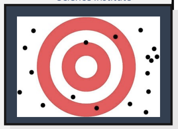
What is the probability that more than 3 missiles hit the target?

$$P(x > 3) = C_4^5 p^4 q^{5-4} + C_5^5 p^5 q^{5-5}$$

$$= \frac{5!}{4!1!}(.8)^4(.2)^1 + \frac{5!}{5!0!}(.8)^5(.2)^0$$

$$= 5(.8)^4(.2) + (.8)^5 = .7373$$

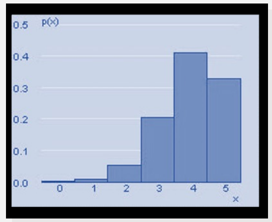
Exercise



- What is the probability that no missiles hit the target?
- What is the probability that fewer than 3 missiles hit the target?
- What is the probability that fewer than 4 but more than 1 missiles hit the target?

Exercise

- Plot the probability distribution for x = number of hits.
- What are the mean and standard deviation for x?



Mean $\mu = np = 5(0.8) = 4$
Standard deviation $\sigma = \sqrt{npq} = 0.89$

The Poisson Random Variable

- The Poisson random variable is often used to model the number of occurrences of a specified event in a given unit of time or space.

- Examples:
  - The number of calls received by a switchboard during 9am to 5pm.
  - The number of times a printer jams in a day
  - The number of traffic accidents at the intersection of Atherton Street and College Avenue in State College on a football weekend.

The Poisson Probability Distribution

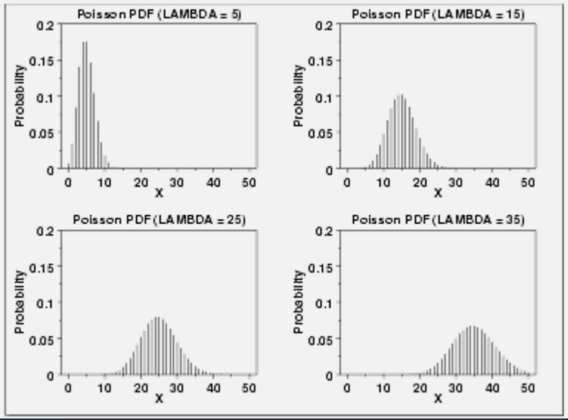
- Let x a Poisson random variable.
- The probability that $X = k$ (k occurrences of the event of interest) $for\ k = 0,1,2,\cdots$ is given by:

  $P(X = k) = \frac{\mu^k e^{-\mu}}{k!}$ for $k \geq 0$

  $P(X = k) = 0$ otherwise

  - Where μ is the mean of the distribution and standard deviation is $\sqrt{\mu}$ and $e \approx 2.718281828459$ is the base of the natural logarithm
  - We get Poisson by fixing the mean of the Binomial distribution, and letting the number of trials approach ∞

Plots of Poisson Distribution



Exercise

- The average number of traffic accidents on a certain intersection in New York city is two per week.
- What is the probability of exactly one accident during a one-week period?

$$P(x=1)=\frac{\mu^k e^{-\mu}}{k!} \quad =\frac{2^1 e^{-2}}{1!}=2e^{-2}=.2707$$

Exercise

- What is the probability that 8 or more accidents happen during a 1-week period?
- What is the probability that no accidents happen during a 1-week period?

The Hypergeometric Probability Distribution

- A bowl contains M red M&M candies and $N - M$ blue M&M® candies.
- Select n candies from the bowl (without replacement) and record x the number of successes - red M&Ms selected.
- Why can't we use the Binomial distribution? – trials are not independent
- Hypergeometric distribution is given by $P(X = k) = \frac{C_k^M C_{n-k}^{N-M}}{C_n^N}$
  - Where N is the population size
  - M is the maximum number of possible successes
  - n is the number of trials
  - k is the number of successes

Mean and Variance of Hypergeometric distribution

$$\text{Mean}: \mu = n\left(\frac{M}{N}\right)$$

$$\text{Variance}: \sigma^2 = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Exercise

- A package of 8 AA batteries contains 2 batteries that are defective.
- A student randomly selects 4 batteries and replaces the batteries in his calculator.
- What is the probability that all four batteries work?

Success = working battery

$$N = 8$$
$$M = 6$$
$$n = 4$$
$$k = 4$$

$$P(x=4) = \frac{C_4^6 C_0^2}{C_4^8}$$

$$= \frac{6(5)/2(1)}{8(7)(6)(5)/4(3)(2)(1)} = \frac{15}{70}$$

Exercise

- What are the mean and variance for the number of batteries that work?

$$\mu = n\left(\frac{M}{N}\right) = 4\left(\frac{6}{8}\right) = 3$$

$$\sigma^2 = n\left(\frac{M}{N}\right)\left(\frac{N-M}{N}\right)\left(\frac{N-n}{N-1}\right)$$
$$= 4\left(\frac{6}{8}\right)\left(\frac{2}{8}\right)\left(\frac{4}{7}\right) = .4286$$

Continuous Random Variables

- A random variable is continuous if it can assume the infinitely many values corresponding to points on a line interval.
- Examples
  - Height, weight
  - Scores on a test
  - Measurement error