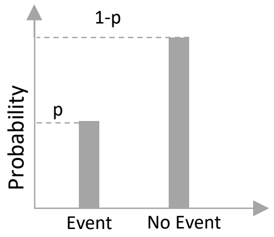
Mathematical Foundations for Data Science DA5000

Session 3 – Probability Distributions

Nandan Sudarsanam,
Department of Data Science and AI,
Wadhwani School of Data Science and AI,
Indian Institute of Technology Madras

Common distributions: Bernoulli

- A distribution where there are only two outcomes



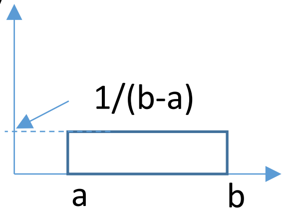
- State space (k) is {0,1}
- PMF is *(1-p)* for *k=0* and *p* for *k=1*
- CDF is 0 for k<0, (1-p) for k=0, and 1 for k = 1
- Mean is *p*
- Variance is *p(1-p)*

- Even if there are many outcomes, we can always split the state space into two-outcomes and define a new random variable which is Bernoulli.
- Examples: Success/failure, Yes/No, Male/Female, Greater 3% or not

Common distributions: Uniform

- Discrete
  - The six sided dice, coin toss, choosing from 4 different models, searching for a matching part from a given set.
  - Formula for pdf: $f(X = x) = \frac{1}{k}$ for all x that belongs to a specific set with k elements and $f(X = x) = 0$ for all other values of x.
- Continuous
  - number of seconds past the minute, searching for a part (over time).
  - Simplistic model of number of kilometres that a truck will last
  - Formula for PDF:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 \text{ for } x<a \text{ and } x>b \end{cases}$$



  - What is the CDF, mean and Variance? How would you derive them?

CDF = $\frac{x-a}{b-a}$

Mean = $\frac{1}{2}(b+a)$

Variance = $\frac{1}{12}(b-a)^2$

Common distributions: Binomial

- Binomial
  - Discrete distribution signifying number of successes of *n* independent Bernoulli experiments
  - Example Real-world: Probability of 4 out of 10 deep sea drills will fail. Probability of there being 5 purchases out of 20 demo invites
  - Formula for PMF: ${}^{n}_{k}C.p^{k}(1-p)^{n-k}$
  - Formula for CDF is just the summation. How would you derive it?
  - Mean: *np*, variance: *np(1-p)*. How would you derive it?

Common distributions: Poisson

- Poisson
  - Discrete distribution that signifies the probability of ‘x’ occurrences of a certain event over a certain period of time or space.
  - Examples: Number of failures per month, Number of purchases per square kilometre.
  - PMF $\frac{\lambda^k}{k!} e^{-\lambda}$
  - Mean and variance are λ (lambda >0). How would you derive it?
  - CDF: How would you derive it?
  - Relating it with the binomial: https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239

Common distribution: Geometric

- Geometric
  - Number of attempts before an event. How many sales calls before one successful sale.
  - The interarrival distributions counterpart of a binomial. Take any of the binomial examples.
  - PMF:
  - CDF:
    $(1-p)^{k-1}p$
  - Mean is $\frac{1}{p}$, and variance $\frac{1-p}{p^2}$
    $1-(1-p)^k$

Common distributions: Exponential

- Exponential
  - The interarrival times of the Poisson distribution
  - The equivalent of the geometric distribution for the Poisson process. Example: Time it takes for the next sales call
  - PDF: $\lambda e^{-\lambda x}$, where lambda>0
  - CDF: 1-$e^{-\lambda x}$
  - Mean: $\frac{1}{\lambda}$
  - Variance: $\frac{1}{\lambda^2}$

Summary of four distributions

The Environment/Context	Count per some unit frame	Interarrival distribution
Bernoulli Process	Binomial	Geometric
Poisson Process	Poisson	Exponential

- The Exponential is the only continuous distribution, whereas the other three are discrete distributions

Hypergeometric and Negative Binomial

- Hypergeometric:
  - We can think of the Binomial as a sampling with replacement
  - What if we sample from a finite population?
  - When will this distribution be similar/different from the binomial?
- Negative Binomial:
  - Trials till the k^{th} success
  - How does it help to think of this in terms of the Binomial and Geometric

Normal Distribution

- Also referred to as a Gaussian distribution
- The most widely used distribution of a random variable
- Characterized by a bell shaped curve
- Why?
  - Central Limit theorem
  - Examples: Various physical characteristics (height, weight, length, etc.), large counts (characters in a page, steps taken in day, etc.), and many approximations.

Normal Distribution

- PDF: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$; $N(\mu, \sigma^2)$
- CDF? Mean? Variance?
- Some useful results:
  - $P(X > \mu) = P(X < \mu) = 0.5$ (what is $P(X = \mu)$?)
  - $P(\mu - \sigma < X < \mu + \sigma) = 0.6827$
  - $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$
  - $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$
- Standard Normal: $N(0, 1^2)$

Normal and Poisson approximations

- Implications when *n* is large and *np* and *nq* is >5
  - Example: Making 10,000 sales calls
  - Computational implications
  - Value of determining the probability of each discrete outcome
  - Mean = *np* and variance = *np(1-p)*
- Poisson approximation of the Binomial when n>100, but *np* or *nq*<10
- Normal can also be used to approximate Poisson when $\lambda > 5$

Simulation

- Replicate the uncertainty in distributions to answer business questions. Uncertainty in Supply, demand, weather, flight timings, defaulting on loans, stock prices, competitors prices, etc.
- Why not use the distributions directly? What about complex dependencies?
- Toy example for binomial
- Assume that we looking at number of visitors for a one hour online offer. The per minute number of people visiting the site is a Normal distribution. However, this is a changing distribution. The mean number of people visiting the site changes as a function of time $8000 - t^2$(time t goes from 0 to 60). The standard deviation is 800 people (per minute). Finally we cannot support more than 9000 visitors at any given minute.