Modeling Risk and Realities: Week 3

Senthil Veeraraghavan
Operations, Information and Decisions Department

Week 3: Choosing Distributions that fit your Data

- Data and visualization: Graphical representation

- Choosing among the family of distributions: Discrete and continuous distributions.

Session 2

- How good does a certain distribution fit? Hypothesis testing and goodness of fit.

Distributions

- In this session, we will look at some families of distributions that are often used to model realities.

- First we examine discrete distributions, and then continuous distributions.

A simple random variable

- First, we will see three simple examples of random variables, which will help us examine other realistic examples.
- A coin is tossed. You will see either “heads” or “tails”. The outcome is a random variable.
- A team plays against another slight weaker opposing team. The team’s probability of winning is 60% (and the probability of losing is 40%).
  - The outcome is a random variable. Note some similarity with the coin toss example.
- A fair “die is cast” in a game of dice. The outcomes can be 1, 2, 3, 4, 5 or 6.
  - The probability of “6” turning up = 1/6.
  - The probability of “1” turning up = 1/6
  - Note that for a fair die, these probabilities are the same for any outcome

Bernoulli Distribution

- Bernoulli distribution has only two outcomes
  - Each outcome occurring with some probability.
  - The two probabilities add up to 1.

- Lets look at some realistic scenarios that can be modeled with Bernoulli distribution.

  - Will a firm from Europe enter the market in Asia?

  - Will a team ranked fourth in the middle of the season win the English Premier League when the season concludes?

  - Will a ride-share company buy its smaller startup competitor?

- What if the number of outcomes is more than two?

A Distribution with Three outcomes

- Suppose a firm enters a new market.
- The managers see three possible “scenarios” to model with a distribution.
- Firm’s market share next year could be “low”, “medium” or “high”.
  - Likelihood of “high” market share is 20%
  - Likelihood of “medium” market share is 70%
  - Likelihood of “low” market share is 10%

Three Outcomes and Probability Distribution

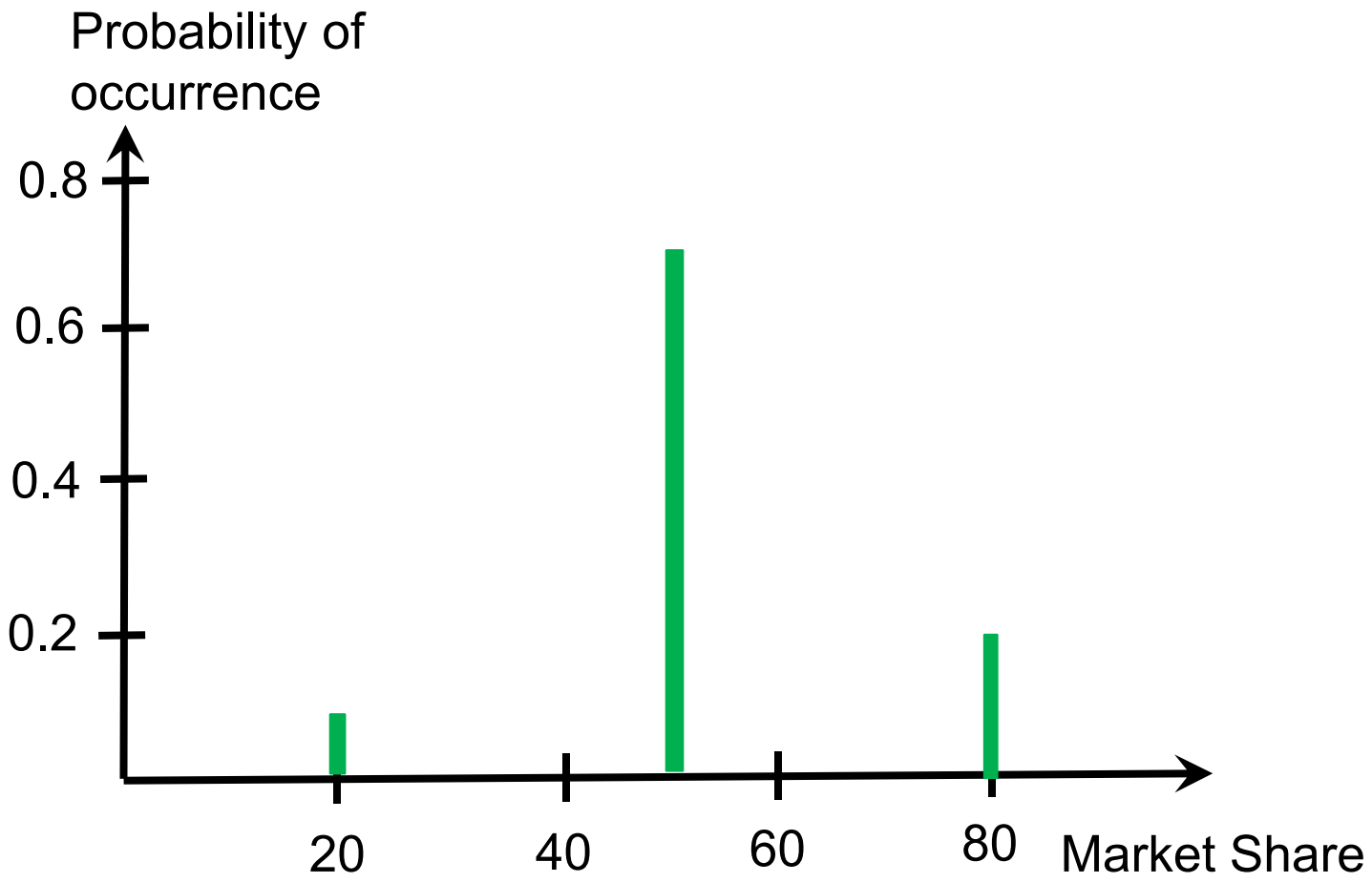
- For example, assume we have a probability distribution for the future market share (based on estimates from experts). Only the following three outcomes are possible.
  - Market share D_1= 80% with probability $p_1 = 0.2$
  - Market share D_2= 50% with probability $p_2 = 0.7$
  - Market share D_3= 20% with probability $p_3 = 0.1$

- Note that the probabilities are
  - greater than zero, and
  - they sum up to 1.

- Probability distributions like that one, described by a number of distinct scenarios with attached probabilities, are called **discrete**

- The probabilities of various outcomes are typically characterized by **a probability density function (pdf).**

A Graphical Representation of a pdf



Describing the Distribution: Mean and Standard Deviation

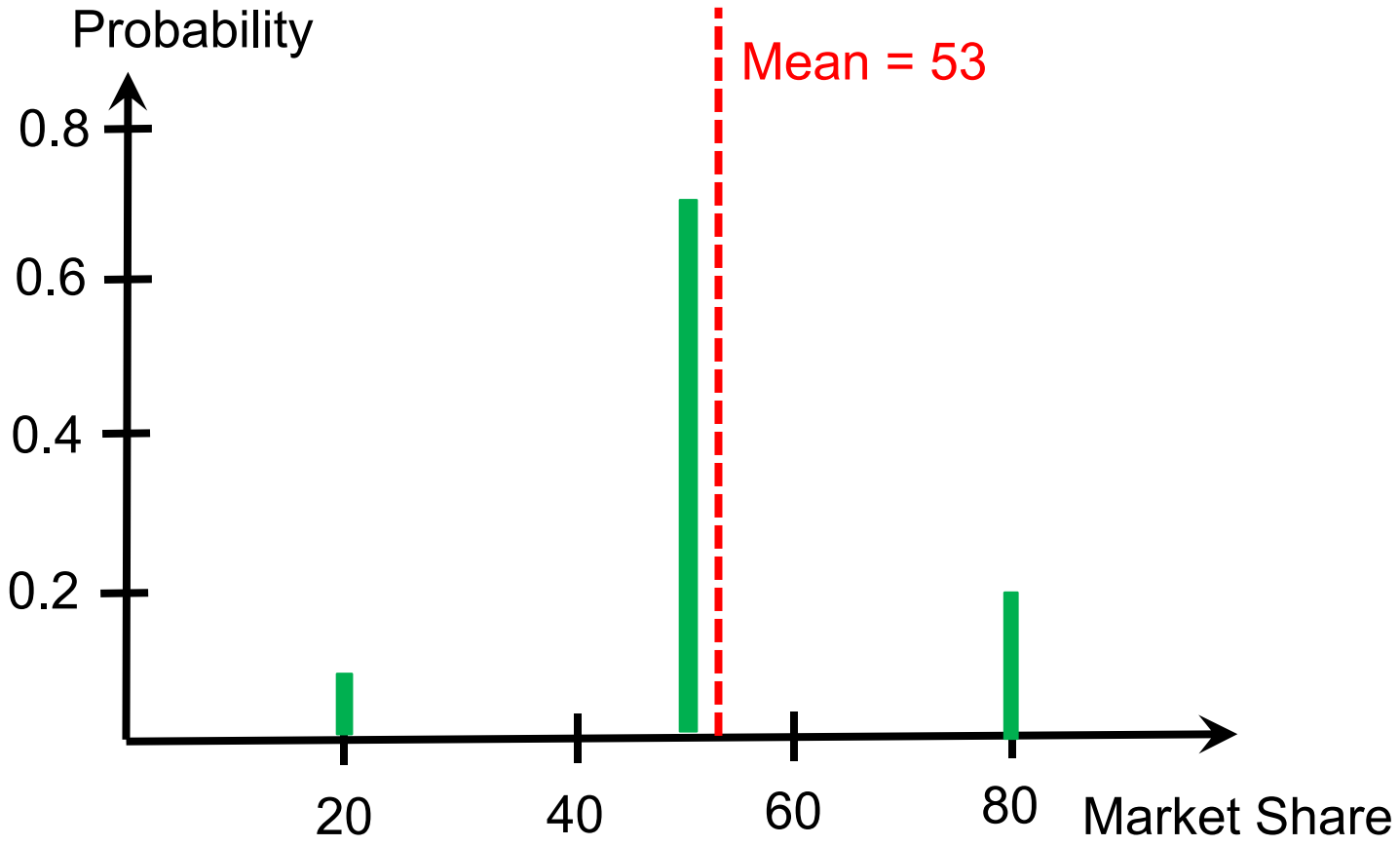
- For any probability distribution, including a simple one reflecting three demand scenarios, two useful descriptors are often calculated: **mean** (also called **expected value**) and **standard deviation**

- For a discrete probability distribution, the mean is defined as a sum of the products of scenario values and their probabilities

- For our market share distribution, the mean $\overline{\mathrm{D}} = p_1\mathrm{D}_1 + p_2\mathrm{D}_2 + p_3\mathrm{D}_3 = 0.2 * 80 + 0.7 * 50 + 0.1 * 20 = 53.$

- Mean, in this case, reflects the average market share we would see, if the firm had a chance to try the same action infinite times.

Three-Outcomes Probability Distribution: Mean



Describing Probability Distribution: Standard Deviation

- Standard deviation describes, roughly speaking, how far away actual random variable values are from the mean, on average. Colloquially speaking, it describes how "spread out" the distribution is around its mean.

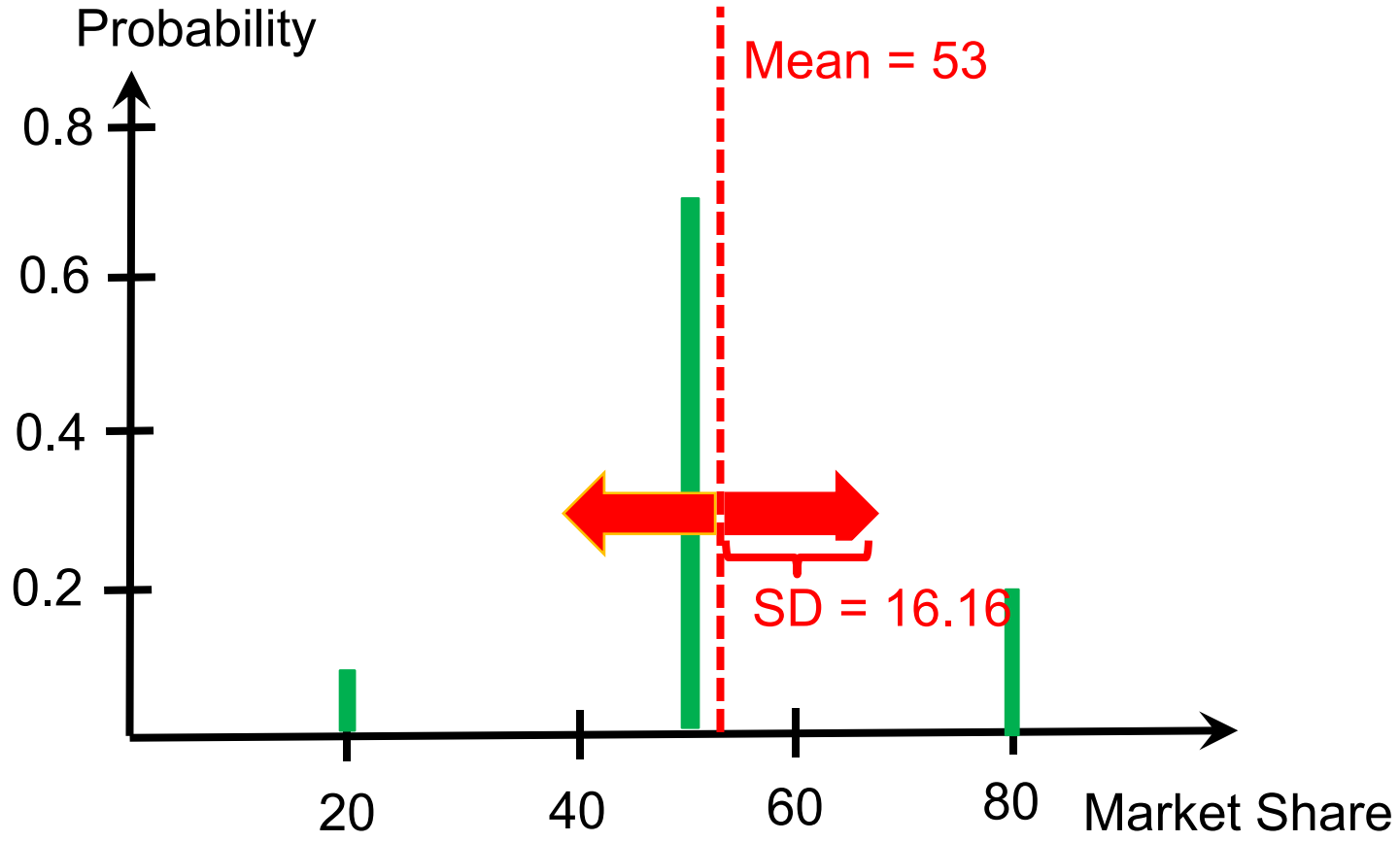
- Standard deviation is defined as a square root of the sum of products of scenario probabilities and the squares of the difference between scenario value and the mean value

- For example, for the three-scenario demand probability distribution we consider, the standard deviation is calculated as

$$\mathrm{SD} = \sqrt{p_1 * (\mathrm{D}_1 - \overline{D})^2 + p_2 * (\mathrm{D}_2 - \overline{D})^2 + p_3 * (\mathrm{D}_3 - \overline{D})^2}$$
$$= \sqrt{0.2 * (80 - 53)^2 + 0.7 * (50 - 53)^2 + 0.1 * (20 - 53)^2} \approx 16.16$$

Three-Outcomes Probability Distribution: Mean and Standard Deviation

- Mean and standard deviation help to support a general intuition about the nature of a random variable



Multiple Outcomes: pdf

- What if we have more than three outcomes?
- Suppose there are n outcomes.
  - D_1 with probability p_1
  - D_2 with probability p_2
  - D_3 with probability p_3

  .....
  - D_n with probability p_n

  and $p_1 + p_2 + p_3 + \cdots + p_n = 1$

- The probability density function (pdf) of the random variable X is typically written as

$$f(k) = Prob(X = D_k) = p_k$$

Multiple Outcomes: Cumulative Distribution Function

- Distributions are also described by the CDF, or cumulative distribution function.
- CDF serves the same purpose as pdf and is just another way of describing the random variable.
- Represented by capital letters (and pdf is often represented by small letters).
- The cumulative distribution function is the sum of pdfs up to the point of interest.

e.g. $F(D_3) = \mathrm{Prob}(X \leq D_3) = p_1 + p_2 + p_3$

Multiple Outcomes: Mean and Standard Deviation

- What about mean and standard deviation of the distribution for *n* discrete outcomes?

- Mean = $\overline{\mathrm{D}} = p_1 \mathrm{D}_1 + p_2 \mathrm{D}_2 + p_3 \mathrm{D}_3 + \cdots + p_n \mathrm{D}_\mathrm{n}$

- Standard deviation =

$$\sqrt{p_1 * (\mathrm{D}_1 - \overline{\mathrm{D}})^2 + p_2 * (\mathrm{D}_2 - \overline{\mathrm{D}})^2 + \cdots + p_n * (\mathrm{D}_n - \overline{\mathrm{D}})^2}$$

Multiple outcomes: Dice Example

- A die is cast. Random variable X = Number that shows up. There are 6 possible outcomes for the random variable X.
- The face could show 1 or 2 or 3 or 4 or 5 or 6.
- All the outcomes have the same probability of occurring. (This is a "fair" die).
- The probability density function (pdf):

$$\mathrm{f(n)} = \mathrm{Prob(X} = n) = \frac{1}{6} \qquad \text{for} \quad \mathrm{n} = 1, 2, 3, 4, 5, 6$$

- This distribution of this random variable is a **discrete uniform** distribution.

Dice Example: Mean and Standard Deviation

- Mean = $\overline{\mathrm{D}} = p_1 \mathrm{D}_1 + p_2 \mathrm{D}_2 + p_3 \mathrm{D}_3 + \cdots + p_n \mathrm{D}_\mathrm{n}$
- Mean = (1/6)(1) + (1/6)(2) + (1/6)(3)+ (1/6)(4) +(1/6)(5) +(1/6)(6)
  = 3.5
- Standard deviation =

$$\sqrt{p_1 * (\mathrm{D}_1 - \overline{\mathrm{D}})^2 + p_2 * (\mathrm{D}_2 - \overline{\mathrm{D}})^2 + \cdots + p_n * (\mathrm{D}_n - \overline{\mathrm{D}})^2}$$

- Standard deviation = 1.708

Discrete Uniform: Cumulative distribution Function

- Probability that Random variable, X is less than equal to n

$$\mathrm{F(n)} = \mathrm{Prob(X} \leq n) = p_1 + p_2 + \cdots + p_n$$

- CDF gives the probability the outcome is less than or equal to some value, so that…
  - F(1) = 1/6
  - F(2) = 2/6 = 0.333
  - F(3) = 3/6 = 0.5
  - F(4) = 4/6 = 0.666
  - F(5) = 5/6 = 0.833
  - F(6) = 1

- The cumulative distribution always takes the value of 1 at the highest outcome.

A Discrete Uniform Distribution

- Consider random variable X with a discrete uniform distribution.
- Discrete uniform distribution is completely described by the total number of possible outcomes N.
  - Suppose, the possible outcomes are numbered by $n = 1,2, \dots, N$
- Probability density function
  - $\mathrm{f(n)} = \mathrm{Prob(X} = n) = \frac{1}{N} \quad for \quad n = 1,2,3, \dots, N$
- Cumulative distribution function
  - $\mathrm{F(n)} = \mathrm{Prob(X} \leq n) = \frac{n}{N} \quad for \quad n = 1,2,3, \dots, N$
- $\mathrm{Mean} = (\mathrm{N} + 1)/2$ and $\mathrm{Standard\ Deviation} = \sqrt{\frac{N^2-1}{12}}$

Another Discrete Example

- Let’s go through another example.
- Example: Binomial Distribution

Multiple Outcomes: Binomial Distribution

- Suppose you are deciding whether to invest in a new medical drug to cure a difficult-to-cure ailment. The drug is currently in clinical trials phase.
- Based on lab tests, the new drug has 60% probability of success in curing the ailment for a patient. With probability 40%, the drug fails (i.e., there is no effect).
- The drug is tested on 10 patients during clinical trials.
- How many successes would be there after clinical trials?
- The number of successes is a random variable with 11 possible outcomes: The number of successes can be 0, 1, 2, 3,…, 8, 9 or 10.

Discrete Example: Binomial Distribution

- The random variable X here is the number of successes.
- The number of successes is distributed “binomially”.
- pdf of the distribution is given on the right.
  - e.g. $f(6) = \text{Prob}(X = 6) = 25.08\%$
- Calculate the mean and standard deviation as an exercise (using the slide titled “Multiple Outcomes: Mean and Standard Deviation")

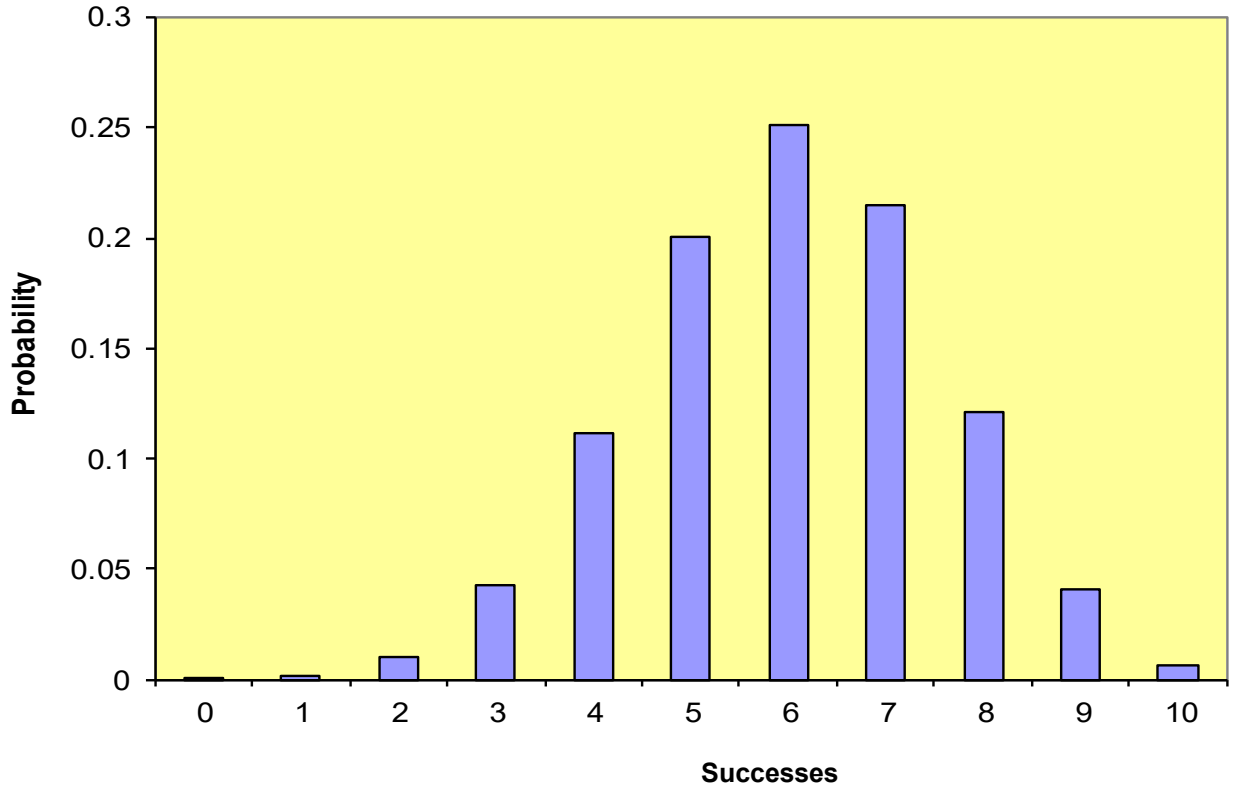
Successes	Probability
0	**0.000105**
1	**0.001573**
2	**0.010617**
3	**0.042467**
4	**0.111477**
5	**0.200658**
6	**0.250823**
7	**0.214991**
8	**0.120932**
9	**0.040311**
10	**0.006047**

Mean = 6

Standard deviation = 1.549

Discrete Example: Binomial Distribution

- Here is the Graphical representation of pdf of the Binomial distribution (for the drug trials example).



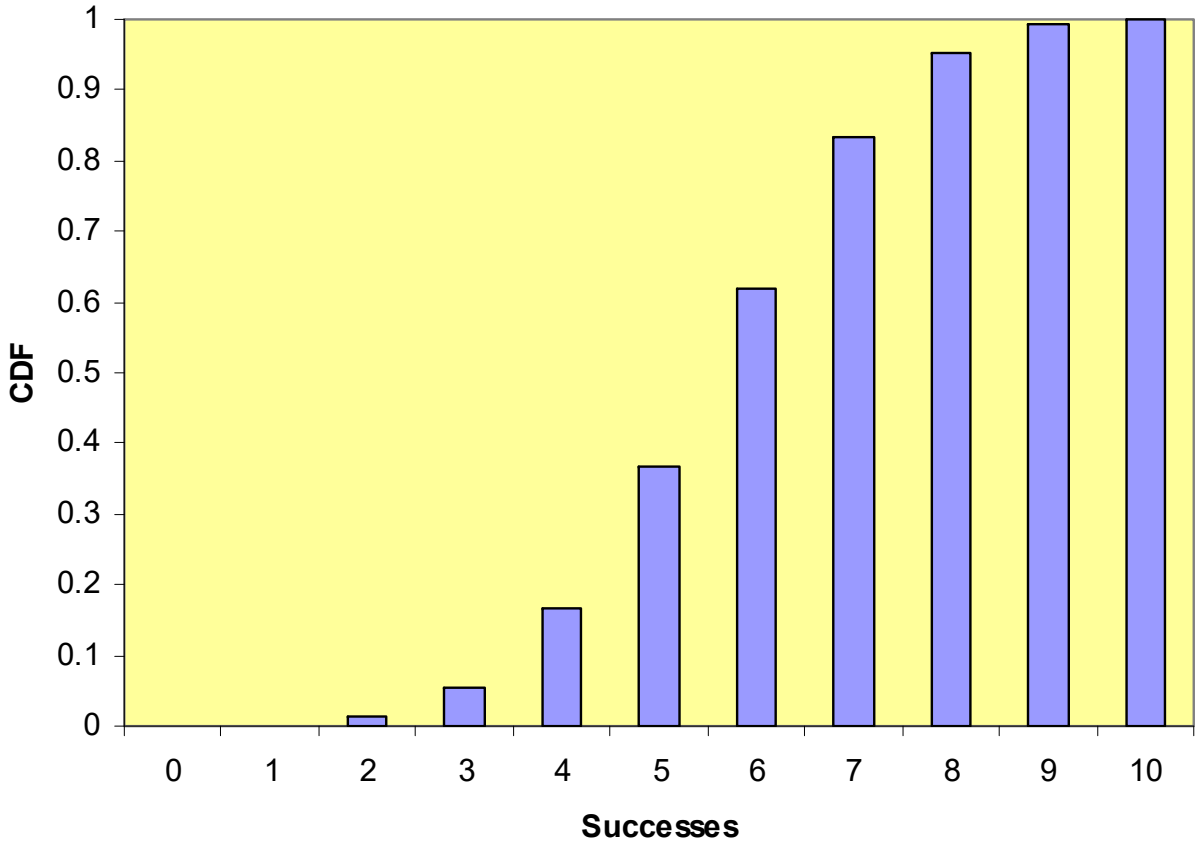
Discrete Example: Binomial Distribution

- The random variable X is the number of successes.
- pdf
  - e.g. $f(6) = \text{Prob}(X = 6) = 25.08\%$
- Cumulative Distribution Function:
  - e.g. $F(6) = \text{Prob}(X \leq 6) = 61.77\%$
- The probability that there will be 6 or fewer successes is 61.77%

n	pdf Pr[X=n]	CDF Pr[X≤ n]
0	0.000105	0.000105
1	0.001573	0.001678
2	0.010617	0.012295
3	0.042467	0.054762
4	0.111477	0.166239
5	0.200658	0.366897
6	0.250823	0.617719
7	0.214991	0.83271
8	0.120932	0.953643
9	0.040311	0.993953
10	0.006047	1

Binomial Cumulative Distribution Function

- Graphical representation of cumulative distribution function of the Binomial distribution
- This is the probability that number of successes (random variable) is less than or equal to n.



Binomial Distribution

- Consider a random variable X distributed according to a Binomial distribution.

- Binomial distribution is completely described by the total number of "trials" *N* and probability of success in each trial *p*
  - The possible outcomes are numbered by $n = 0, 1, 2, 3, \dots, N$

- Probability density function
  - $\mathrm{f(n)} = \mathrm{Prob(X} = n) = \binom{N}{n} p^n (1-p)^{N-n}$ where $\binom{N}{n} = \frac{N!}{(N-n)!n!}$

    e.g. 4!=1x2x3x4=24.

- Cumulative distribution function
  - $\mathrm{F(n)} = \mathrm{Prob(X} \leq n) = \sum_{k=0}^{n} \binom{N}{k} p^k (1-p)^{N-k}$

- Binomial: $\mathrm{Mean} = \mathrm{Np}$ and $\mathrm{Standard\ Deviation} = \sqrt{\mathrm{Np(1-p)}}$

Continuous Distributions

- Now let's take a look at continuous distributions.

Discrete vs. Continuous Probability Distributions

- So far, we have looked at discrete probability distributions with a countable number of outcomes or scenarios.

- Sometimes…
  - a) random variable being modeled has a large number of scenarios on any small interval of values and
  - b) the probability that any one exact scenario is realized is very small

- Think of examples such as stock prices or the amount of rainfall in a region.

- In such cases, it makes sense to describe such a probability distribution using groups of scenarios rather than focusing on individual scenarios

Continuous Random Variables

- A random variable X may take any value on the continuous real line.
- We can describe pdf and CDF as before…
- f(x) – density function refers to the probability that random variable X is in the infinitesimal region around x.
- F(x): Cumulative Distribution Function. This is cumulative and can be defined as the probability that random variable X is less than or equal to x.

$$\mathrm{F(x)} = \int_{-\infty}^{x} f(u)du$$

Continuous Random Variables: Mean and Standard Deviation

- Mean or Expectation of the random variable:

$$\mu = E[X] = \int_{-\infty}^{\infty} uf(u)du$$

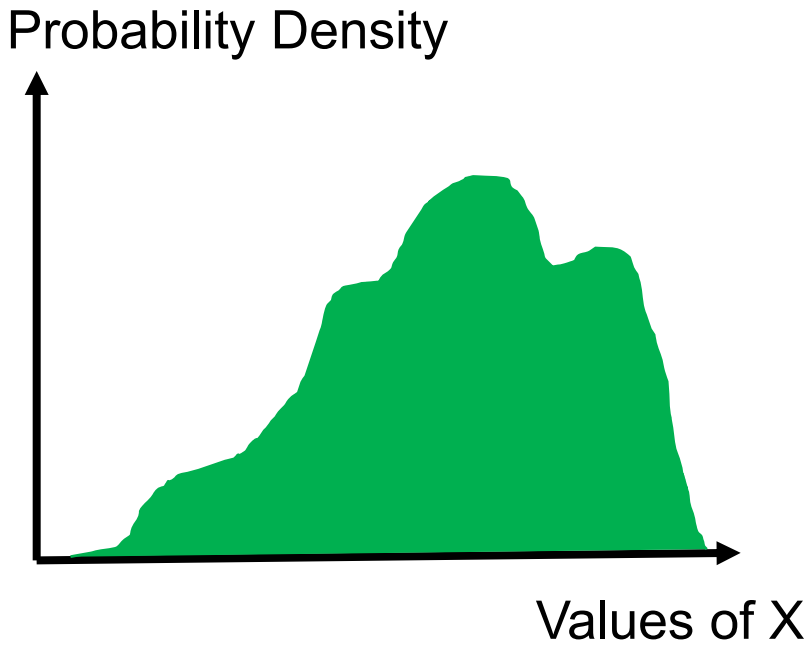
- Variance of the random variable:

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (u-\mu)^2 f(u)du$$

- Standard deviation is the square root of the variance.

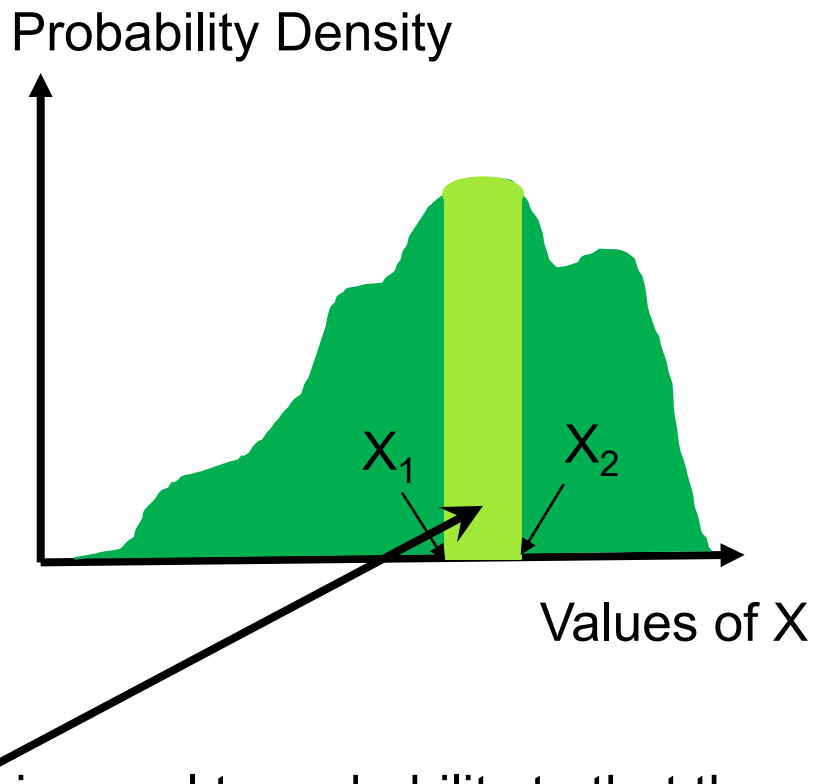
Continuous Distribution: Random Variable X

- Distributions like this are called continuous



Continuous Distribution: Random Variable X

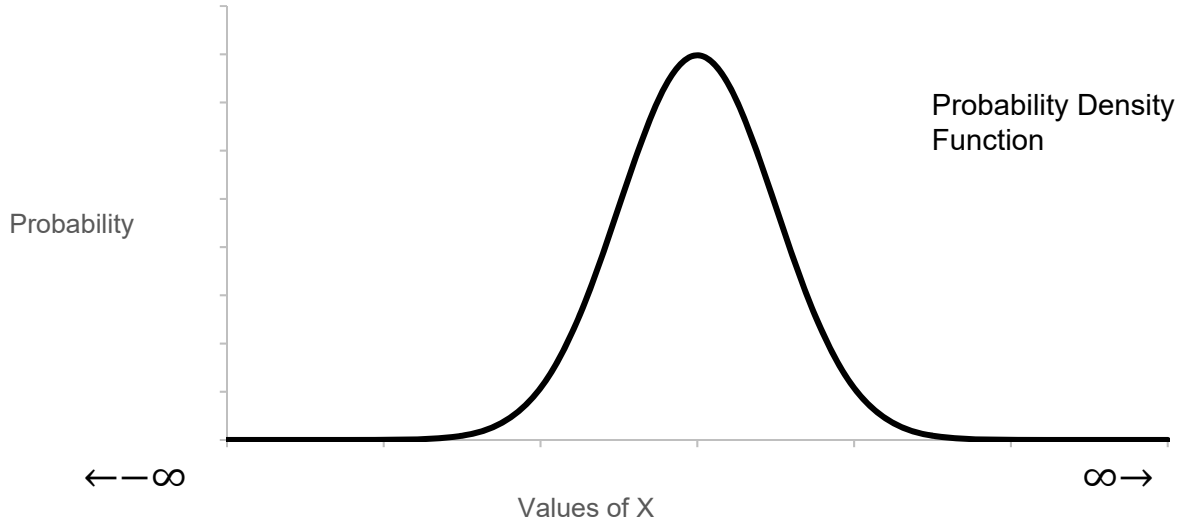
- Distributions like this are called continuous



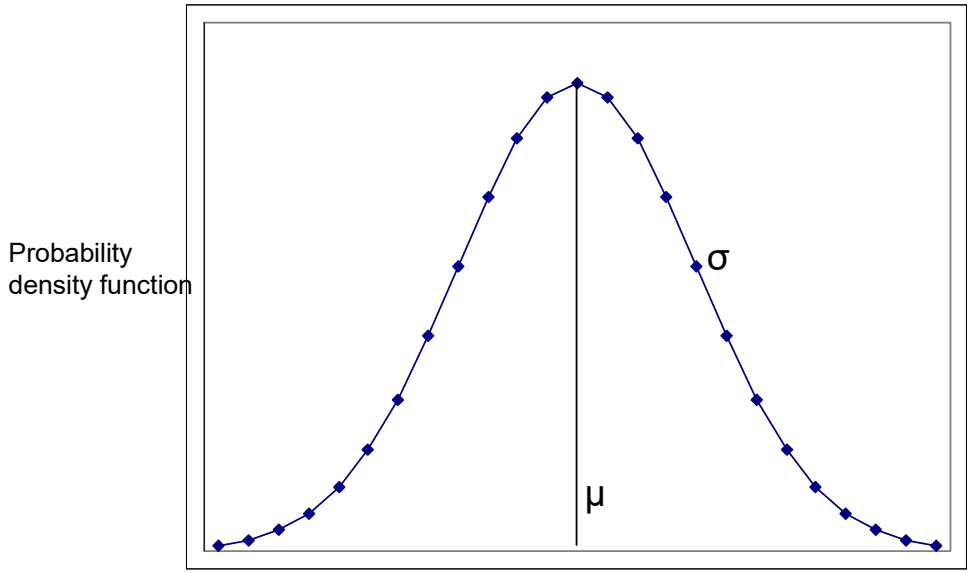
- The area is equal to probability to that the random variable X takes values in the interval between X_1 and X_2
- The area under the entire curve is equal to 1

Example 1: Normal Distribution

- One of the most popular examples of a continuous probability distribution is normal distribution
- Allows the underlying random variable to take any value from negative infinity to positive infinity, and
- is completely characterized by two parameters – mean μ and standard deviation σ.



Normal Distribution



- Probability density function for x on real line

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

- Cumulative distribution function for any real value of x

$$\mathrm{F(x)} = \int_{-\infty}^{x} f(u)du$$

Normal Distribution

- The statistical formulas (can be implemented in Excel) to calculate the pdf and CDF
  - for a normal random variable X with given mean μ and standard deviation σ.

- pdf $\mathrm{f(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$ In excel use $normdist(x, \mu, \sigma, 0)$

- CDF $\mathrm{F(x)} = \int_{-\infty}^{x} f(u)du$ In excel use $normdist(x, \mu, \sigma, 1)$

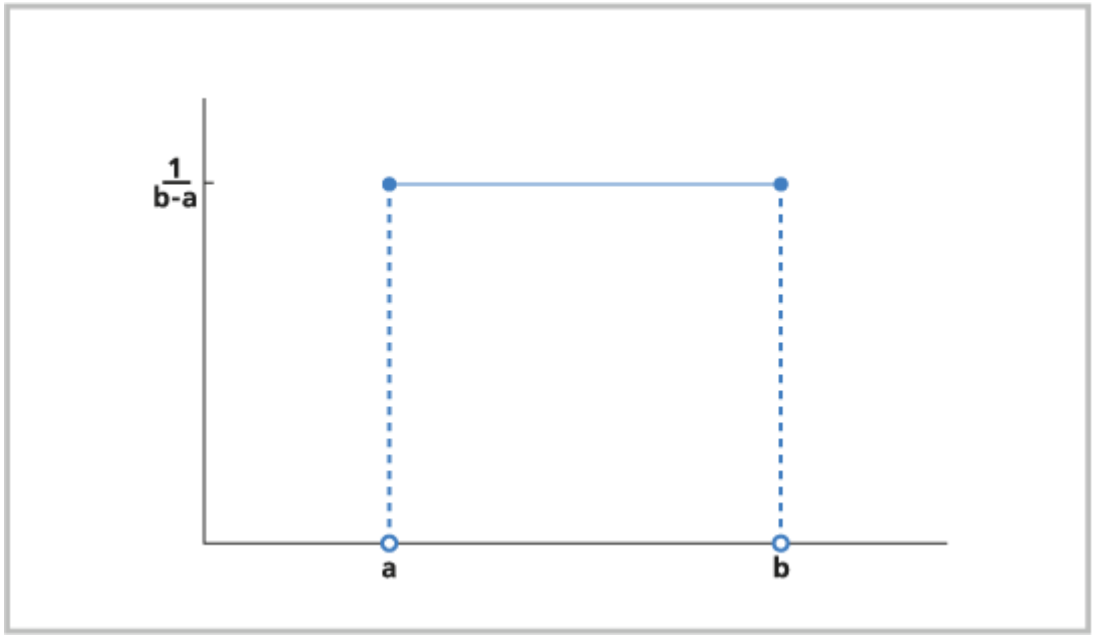
Example 2: Uniform Distribution

- Allows the underlying random variable to take any value between two points – a minimum point (say, a) to a maximum point (say, b) and all outcomes are equally likely to occur.

- Is completely characterized by two parameters – minimum and maximum value.

Example 2: Uniform Distribution

Probability density Function



- pdf: $\mathrm{f(x)} = \frac{1}{(b-a)}\ for\ a \leq x \leq b,\ and\ 0\ otherwise$
- CDF: $\mathrm{F(x)} = \begin{cases} 0\ for\ x < a \\ (x-a)/(b-a)\ \ for\ a \leq x \leq b \\ 1\ for\ x > b \end{cases}$

Other Continuous Distributions

- We focused on the two example distributions. However, many other continuous distributions are often used.
  - Exponential distribution, e.g. used to model loan processing times.
  - Beta distribution, e.g. for modeling project completion times over fixed intervals.
  - Gamma distribution, e.g. for modeling time between events in insurance risk.
  - Lognormal distribution, e.g. to model events with low probabilities of large values.
- Which distribution fits well? We will learn about goodness of fit tests (for normal and uniform distributions) in the next session.

See you in Week 3 Session 3

Senthil Veeraraghavan
Operations, Information and Decisions Department