# Statistical Inference and Predictive Analytics

IS 6489

Session 2

1

## Outline

- Discuss Assignment 1
- Going from population to individual unit
  - Outcome
  - Probability
  - Random variable
  - Expectation
- Simulation



2

## Probability – Terminology

- *Experiment*: any procedure that can be repeated infinitely
  - Rolling die **once**
  - Maximum temperature tomorrow
- *Sample space*: All possible outcomes in an experiment
  - Rolling die: 1, 2, 3, …, 6
  - Temperature: [10, 100]
- *Event*: any collection of outcomes in the sample space
  - The outcome is even (2, 4, 6); outcome is 1; …
  - Max temperature is above 50; is lower than yesterday's temperature; …





## What is Probability?

- *Probability* is a **function** that maps events to a number between [0,1].

*Events* ⟶ [0, 1]

- Experiment: rolling two dice
- Outcomes: {(1,1), …, (1,6), (2,1), …, (2,6),…}
- Events: Sum of dice face is even, both dice have identical outcomes, the outcomes are 3 and 1, …

## Review of Probability – Random Variable

- The outcome of an experiment, which is unknown, is represented by a *random variable*
  - usually represented by a capital letter.
  - Temperature, demand, stock price, number of clicks on a website
- Random variables are categorized as either *discrete* or *continuous*
  - Temperature?
  - Is it going to snow tomorrow?
  - Demand?
  - Number of clicks on a website?
- Example: Random variables X denotes the outcome of coin toss
  - Takes value 0 (corresponding to Heads) or 1 (otherwise).
  - What is Prob(X = 0)?





## Review of Probability – Distributions

- The *Probability Distribution* of a random variable tells us the probability that the variable can take one or more of the possible values
  - What is the probability maximum temperature tomorrow is between 35F and 45F?
    - Let the random variable Maximum Temperature be denoted by Tmax: Prob(Tmax $\in$ [35,45]).
  - What is the probability tomorrow's demand for a product is at least 100 units.
    - Let the random variable Demand be denote by D: Prob(D $\geq$ 100)

## Review of Probability – Discrete Distribution

- When the random variable is discrete, the associated probability distribution is also discrete
- Let the random variable R denote the outcome of a dice roll.
  - R takes values 1, 2, 3, 4, 5, or 6
  - Probability distribution (more accurately, *mass function*) is given by

| Outcome of dice roll | Probability of outcome |
|---|---|
| 1 | 1/6 |
| 2 | 1/6 |
| 3 | 1/6 |
| 4 | 1/6 |
| 5 | 1/6 |
| 6 | 1/6 |



7

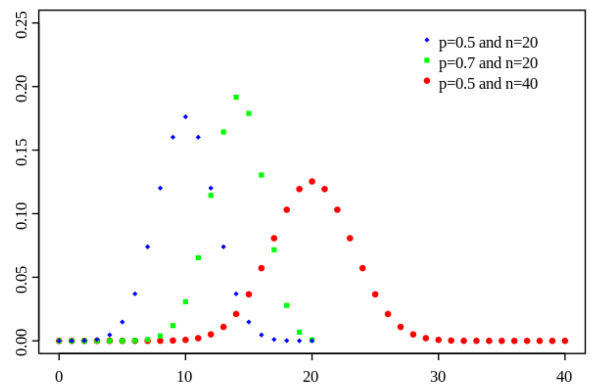
## Review of Probability – Examples of Discrete Random Variables

- Uniform (parameter: $n$ number of elements in sample space)
  - All outcomes equally likely: Roll of a die, Tossing a coin
- Bernoulli (parameters: $p$)
  - Used to describe an experiment that has only two outcomes: *success* or *failure,* represented as 0 and 1, respectively
  - Success with probability $p$ (or failure with probability 1-$p$).
- Binomial (parameters: $n$, $p$)
  - Used to describe the number of success in an experiment that involves multiple Bernoulli trials; takes values 0, 1, …, $n$.
  - Success in each trail has probability p.



8

## Probability Mass Function – Binomial

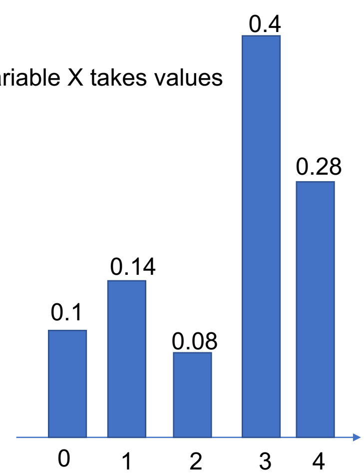





9

## Discrete R.V.

A random variable X takes values 0, 1, 2, 3, 4




1. What is the probability $X \geq 2$?
2. What is the probability $3 \leq X \leq 4$?
3. What is the probability $X \leq 1$ OR $X \geq 4$?
4. What is the probability $X \geq 2$ AND $X \leq 3$?



10

## Expected Value (Discrete Random Variable)

- Definition: the *expected value* of a random variable X, which takes values $x_i$ with probability $p_i$, where $i = 1,2, \dots, n$, is given by the weighted sum of the outcomes:

$$E[X] = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$$ (mathematical definition)

- If the experiment is repeated long enough, the average of all outcomes approaches expected value
  - Suppose $X_i$ is the outcome observed in the $i$-th experiment. Then

$$E[X] \cong \frac{X_1 + X_2 + \cdots + X_T}{T}$$ (data-driven)





## Continuous Distributions

- Uniform (parameters: $l$, $u$)
  - Takes any value in the range [$l$, $u$] with equal probability
- Normal (parameters: $\mu$, $\sigma$)
  - Can take any value in $(-\infty, +\infty)$
  - Mean is $\mu$, and standard-deviation is $\sigma$
- Exponential (parameter: $\lambda$)
  - Can take any value in $[0, +\infty)$
  - Mean is $\lambda$

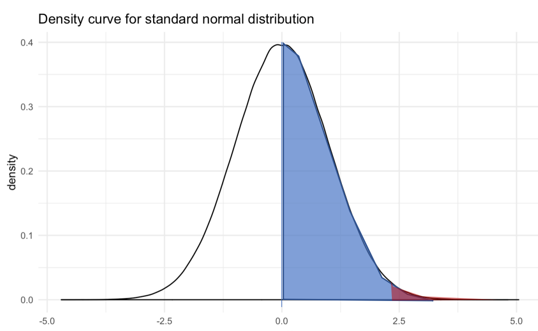
## Continuous Distributions

- *(Cumulative) Distribution Function* (CDF) gives the probability that a random variable is less than or equal to a number
  - CDF denoted by $F(x) = \text{Prob}(X \leq x)$.
  - What is the probability X is between [2, 3]?
    = Probability(X ≤ 3) – Probability(X ≤ 2)
    = F(3) – F(2)
- *Probability density function* (PDF) gives the probability that a random variable belongs to a tiny interval around any point.
  - Probability that r.v. X lies in the interval $[x, x + \Delta]$, where $\Delta > 0$ is a very small number.



13

## Probability Density Function – Standard Normal




1. What is the probability X ≥ 0?
2. What is the probability
   2 ≤ X ≤ 5?
3. What is the probability X ≤ 0?

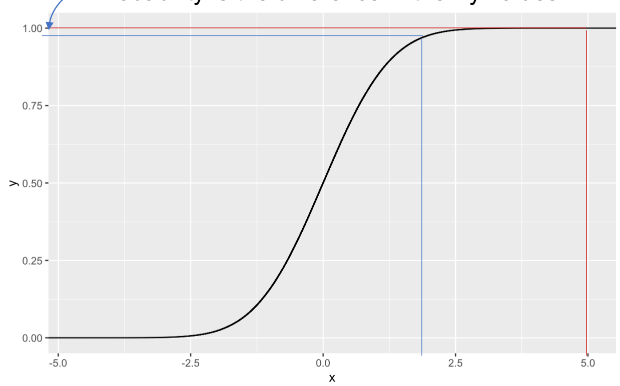
Probability is the area underneath the curve



14

# Cumulative Distribution Function – Standard Normal

Probability is the difference in the ``y-values"




What is the probability 2 ≤ X ≤ 5?

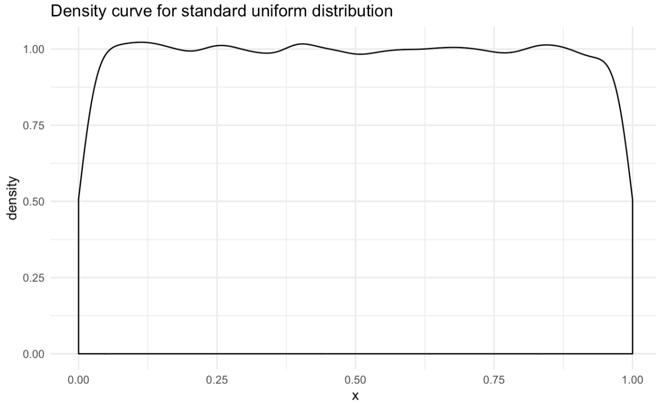
```
> pnorm(5)-pnorm(2)
[1] 0.02274985
```

```
> pnorm(2,lower.tail = FALSE)-pnorm(5,lower.tail = FALSE)
[1] 0.02274985
```



15

# Probability Density Function – Standard Uniform




1. What is the probability X ≤ .5?
2. What is the probability .2 ≤ X ≤ .6?
3. What is the probability X = .8?



16

## Expected Value (Continuous Random Variable)

- Definition: the *expected value* of a random variable X with pdf $f(x)$ is given by the (weighted sum) integral over the outcomes:

$$E[X] = \int x f(x)\, \mathrm{dx}$$ (mathematical definition of *area under curve f(x)*)

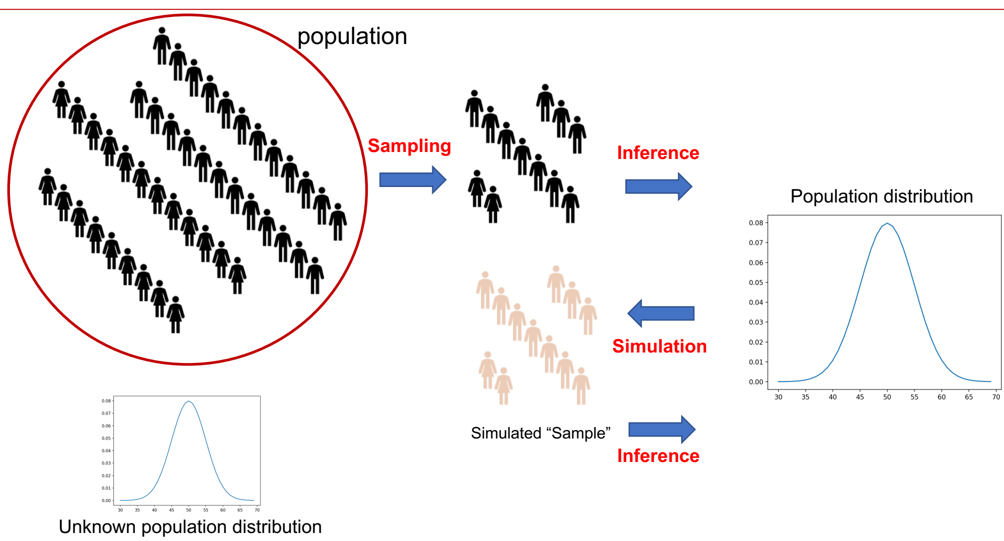
- If the experiment is repeated long enough, the average of all outcomes approaches expected value
  - Suppose $X_i$ is the outcome observed in the $i$-th experiment. Then

$$E[X] \cong \frac{X_1 + X_2 + \cdots + X_T}{T}$$ (data-driven)



17

## Simulation as a Statistical Tool






18

## Simulation

- **Simulation** refers to (imaginary) random sampling from a known population distribution
  - Imaginary because we do not physically "sample" from the population
  - Draw random samples which have the necessary statistical properties
  - Evaluate the quantity of interest (typically sample mean, standard deviation etc.)

- Suppose we know the population distribution of a (random) variable of interest (e.g., customer expenditure). But
  - We do not know it's mean, median, standard deviation, etc.
  - Need to evaluate another random variable (say, firm's profit which depends on customer expenditure) whose definition depends on the original r.v. in a complex fashion.

- The approach is useful even in estimation/inference
  - Bootstrapping
  - Markov Chain Monte Carlo technique





## Restaurant Example

Read the *Restaurant* case uploaded on canvas.