Notes

Week 2

Module 1 Week 2B

Continuous Random Variables and the Normal Distribution

Random Variables

- A **Continuous Random Variable** can take on any value in an interval
- The probability of any single value is *zero*
- **Probability Density Function (PDF)**
  - The probability of a random variable takes on a value between a and b
    - Equal to the area under the *PDF* between a and b (via *Integration*)
- **Cumulative Distribution Function (CDF)**
  - The probability a random variable will take on a value equal to or less than some value.
    - Equal to the area under the CDF to the left of the value of interest.
- **Uniform Distribution (Continuous)**
  - All outcomes are equally likely
  - $X \sim U(a, b)$
  - **PDF**:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in (a, b) \\ 0, & \text{otherwise} \end{cases}$$

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

Normal Distribution

- **Normal Distribution**
  - By far, the most well-known and widely used probability distribution
  - Symmetric
  - **Mean** = **Median** = **Mode**
  - $X \sim N(\mu, \sigma^2)$
  - **PDF**:

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

  - **Mean** $= \mu$
    - Determines the center.
  - **Variance** $= \sigma^2$
    - Determines the spread.
  - Can be fully characterized by just *Mean* and *Variance*
- **Standard Normal Distribution**
  - Normal distribution of standardized values - for easier comparisons when original *Random Variables* are measured in different units.
  - We can use one distribution, the standard normal distribution, to make probability statements about any normally distributed variable.
  - Remember, a variable is standardized by subtracting the mean from each value and dividing by the standard deviation. The units of measurement become **Standard Deviations**. For a *Normal Distribution*, these standardized values are called **Z-Scores**.

- If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, where:
- $Z \sim (0, 1)$
- Standardized version Z be used to calculate probabilities associated with any *Normal Random Variable.*
- **PDF**:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

- **Mean** $= 0$
- **Variance** $=$ *Standard Deviation* $= 1$
- The nice properties of the normal distribution facilitate calculating probabilities facilitates calculating probabilities. E.g.:
- Suppose $X \sim N(10, 4)$
- What is the probability that X is greater than 14?
  - $\mathbb{P}(X > 14)$
  - Convert to standard normal distribution:
    - $X = 14$ corresponds to $Z = \frac{(14-10)}{2} = 2$
      - 14 is 2 *Standard Deviations above* the mean of X
  - Using a statistical table for the *Standard Normal Distribution* (or a computer) we find:
    - The area to the right of $Z = 2$(i.e. to the right of X $= 14) = 0.0228$
- What is the probability that X is less than 8?
  - $\mathbb{P}(X < 8)$
  - Convert to standard normal distribution:
    - $X = 8$ corresponds to $Z = \frac{(8-10)}{2} = -1$
      - 8 is 1 *Standard Deviations below* the mean of X
  - Using a statistical table for the *Standard Normal Distribution* (or a computer) we find:
    - The area to the left of $Z = -1$(i.e. to the left of X $= 8) = 0.1587$
- What is the probability that X is in between 8 and 14?
  - $\mathbb{P}(8 < X < 14)$
    - The area in between Z-scores -1 and $2 = 1\check{}(0.0228 + 0.1587) = 0.8185$

Central limit theorem (CLT)

- The **Central Limit Theorem (CLT)** is one reason why the *Normal Distribution* is so important in statistical methods.
- Roughly, the *CLT* states that for a random variable X with a mean (μ) and finite variance (σ^2) the distribution of the sum of X and the mean of X are normally distributed.
  - $\sum X \sim N(n\mu, n\sigma^2)$
  - $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- Remember, random *Sampling* makes *Statistics Random Variables.*
  - *Random Variables* are described with a *Probability Distribution*
  - The *Probability Distribution* for a *Statistic* is called its **Sampling Distribution**
    - *Sampling Distribution* is what links an unobservable *Population* to the observed *Sample* of data
    - Many *Statistics* involve sums, making the *CLT* an important tool for characterizing the *Sampling Distribution*
- The *CLT* holds (if n is large enough) even if X is *not* normally distributed.