Probability and Mathematical Statistics in Data Science

Lecture 27: Section 10.1: Density Section 10.2: Expectation and Variance

Continuous Random Variables

- Recall the definition of pmf for a discrete random variable $P(X=x)$. Can we extend this definition to continuous random variables?

- The probability model for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes.

- The probability model of X is often described by a smooth curve, which is the probability density function (pdf) of X.

Probability Density Function

- The probability density function (pdf) of a continuous rv X is a function *f*(*x*) such that for any two numbers *a* and *b* with *a* ≤ *b*,

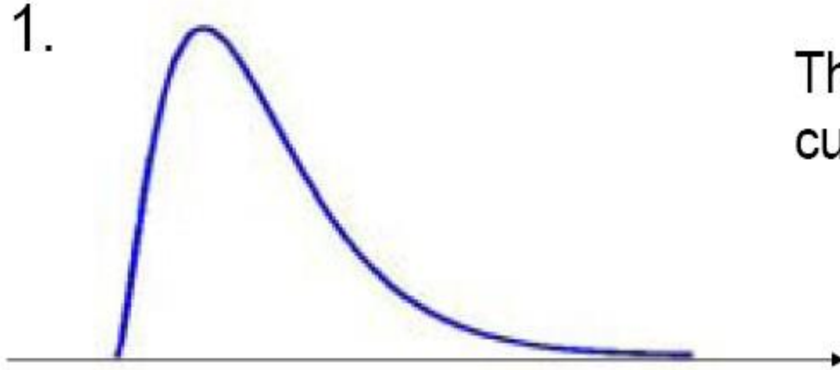
$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

  The graph of *f*(*x*) is often referred to as the *density curve*.

- This means the area under the density curve represents probability!

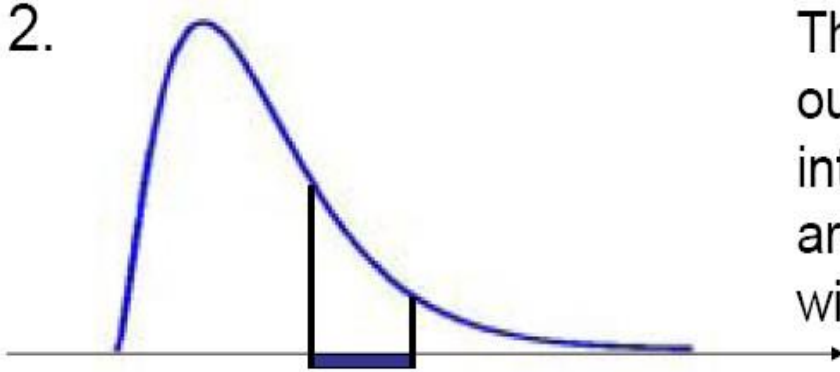
Properties of PDF

1.



The total area under the curve must equal 1.

2.



The probability that the outcome lies in a specific interval is given by the area under the curve within that interval.

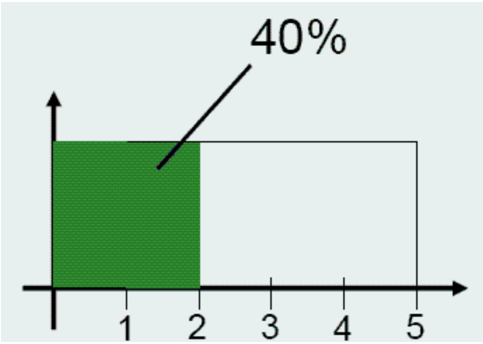
Uniform Distribution

- A continuous random variable X is said to have a uniform distribution on the interval [*A*, *B*] if the pdf of X is

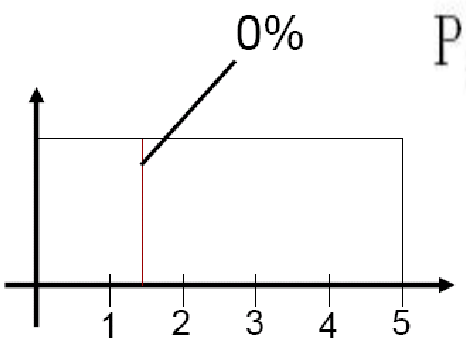
$$f(x; A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

Example

Suppose a bus arrives equally likely at any time between 7:00 – 7:05 AM. What is the probability it arrives sometime between 7:00 – 7:02 AM?



$$P(0 \leq X \leq 2) = \int_0^2 \frac{1}{5} dx = \frac{2}{5}$$



$$P(X = c) = \lim_{\epsilon \to 0} P(c - \epsilon \leq X \leq c + \epsilon) = \lim_{\epsilon \to 0} \int_{c-\epsilon}^{c+\epsilon} \frac{1}{B - A} dx = 0$$

The Cumulative Distribution Function

- The cumulative distribution function (cdf) $F(x)$ for a continuous random variable X is defined for every number x by

$$F(x) = \mathrm{P}(\mathrm{X} \le x) = \int_{-\infty}^{x} f(y)dy$$

- $F(x)$ is in fact the probability that a random variable X is smaller than x.

- It is easy to compute probabilities using $F(x)$.
  - $\mathrm{P}(\mathrm{X} > a) = 1 - F(a)$
  - $\mathrm{P}(a \le \mathrm{X} \le b) = F(b) - F(a)$

Probability Density Function (PDF) from Cumulative Density Function (CDF)

- If X is a continuous random variable with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists,

$$F'(x) = f(x).$$

- $f(x)$ is often a smooth curve, which is the probability density function (pdf) of X.

- The median of a continuous distribution, denoted by $\tilde{\mu}$, is the 50th percentile, so $\tilde{\mu}$ satisfies $.5 = F(\tilde{\mu})$. That is, half the area under the density curve is to the left of $\tilde{\mu}$ and half is to the right of $\tilde{\mu}$.

Expected Values

- Notice that the pdf $f(x)$ of a continuous distribution is actually playing the role of pmf $p(x)$ of a discrete distribution.

- Recall that the expected value of a discrete distribution is calculated by

$$\mu_X = \mathrm{E}(\mathrm{X}) = \sum_{x \in D} x \cdot p(x)$$

- Therefore, similarly we can define the expected value of a continuous distribution by

$$\mu_X = \mathrm{E}(\mathrm{X}) = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

Variance

- With a similar argument as in the discrete case, we can also define the expectation of a function of a continuous random variable as well as the variance of a continuous random variable .

- If X is a continuous random variable with pdf *f*(*x*) and *h*(X) is any function of X, then

$$\mathrm{E}[h(\mathrm{X})] = \int_{-\infty}^{\infty} h(x) \cdot f(x)dx$$

- As a special case of the above is the variance of X defined by

$$\sigma_X^2 = \mathrm{Var}(\mathrm{X}) = \mathrm{E}(\mathrm{X} - \mathrm{E}(X))^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x)dx$$

The standard deviation of X is $\sigma_X = \sqrt{\mathrm{Var}(\mathrm{X})}$

Same Properties as Discrete Random Variables – From Text

Properties of expectation and variance are the same as before. For example,

- Linear functions: $E(aX+b) = aE(X)+b$, $SD(aX+b) = |a|SD(X)$
- Additivity of expectation: $E(X+Y) = E(X)+E(Y)$
- Independence: X and Y are independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all numerical sets A and B.
- Addition rule for variance: If X and Y are independent, then $Var(X+Y) = Var(X)+Var(Y)$

The Central Limit Theorem holds too: If $X_1, X_2, \ldots$ are i.i.d. then for large n the distribution of $S_n = \sum_{i=1}^{n} X_i$ is approximately normal.

Question

Shortcut Formula: $V(X) = E(X^2) - [E(X)]^2$

11. Let X denote the amount of time a book on two-hour reserve is actually checked out, and suppose the cdf is

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x^2}{4} & 0 \le x < 2 \\ 1 & 2 \le x \end{cases}$$

Use the cdf to obtain the following:

a. $P(X \le 1)$

b. $P(.5 \le X \le 1)$

c. $P(X > 1.5)$

d. The median checkout duration $\tilde{\mu}$ [solve $.5 = F(\tilde{\mu})$]

e. $F'(x)$ to obtain the density function $f(x)$

f. $E(X)$

g. $V(X)$ and σ_X

h. If the borrower is charged an amount $h(X) = X^2$ when checkout duration is X, compute the expected charge $E[h(X)]$.