# How to Characterize a New Distribution

## – Continuous Distributions

### Part I: Concepts from Probability

1. **Concepts Related to A Distribution**
   - Definition of random variables
   - Why study the distribution of random variables? – characterize the random behaviors of random variables.

     **Example**. The average starting salary of undergraduate at a university is fixed non-random number (possible unknown before you have a complete list of starting salaries). However, when taking a sample (subset) of all graduating students from undergraduate programs, the average of sample starting salary is random. We can not say that the average starting salary is equal to the sample mean due to the randomness of the sample. We need to use mathematical methods to characterize this uncertainty.

   - How to characterize a random variable – probability distribution!
   - The mathematical definition of a <u>single</u> continues random variable is through its univariate *density function* (**pdf**) $f(x)$ if use capital $X$ to denote (name) the corresponding random variable.
   - When a function can be the density function of a random variable? – Two basic conditions:

     (1). $f(x) \geq 0$ for all possible values that the random variable can take.
     (2). $\int_A^{\square} f(x)\,dx = 1$, R is the set of all possible values (domain) of random variable $X$. For convenience, we a general notation for the domain $R = (-\infty, \infty)$ for the random variable. The actual domain could a subset of $R = (-\infty, \infty)$.

     Several comments on random variables:
     - A random variable can take both positive and negative values (e.g, temperatures)
     - Some variables can only be positive (e.g., lifetime of light bulbs, salaries, body temperature of human beings)

- Some variables can only be negative – this can be easily converted to possible random variables.

*Caution*: a density function is always non-negative regardless of the sign of the underlying random variable.

- Definition of probability associated with an event defined based on the continuous random variable.

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

Where $X$ is the name of the random variable, $x$ is the value (also called the realization) of random variable $X$. $f(x)$ is the density function to characterize the probability distribution of random variable $X$. Assume $a \leq b$ and both $a$ and $b$ are in the domain of $X$ (or density function $f(x)$).

- Cumulative (Probability) Distribution Function (**CDF**) – usually denoted by $F(x)$ which is defined by

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\,dt$$

Properties of $F(x)$:
(1). $F(x)$ is a continuously increasing function of x.
(2). $0 \leq F(x) \leq 1$
(3). $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$

- The relationship between pdf and CDF

$$\frac{dF(x)}{dx} = f(x)$$

**2. Numerical Measures - Moments**

- Regression is all about the mean of ransom variable(s) – The mean of a single random variable $X$ with density function $f(x)$ is the first order moment which defined to be

$$E[X] = \int_{-\infty}^{\infty} xf(x)\,dx$$

$E(x)$ is usually denoted by $\mu$.

From the Riemann sum, we can consider the above definition of expectation as a weighted average (think about why this is an approximated average?)

- The next important numerical measure of is variance which is the defined to be weighted average of squared deviations from the mean, to be more specific

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx$$

- Definition of k-th moment is given in the following.

$$E[X^k] = \int_{-\infty}^{\infty} x^k f(x)\, dx$$

Based on the above definition, the 2nd moment of $X$ is

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$

As a simple exercise, one can express the variance in terms of the first and the second moments (prove this!).

You can think about deriving the mean and variance of the general normal distribution $X$ with density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\pi)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

- Moment Generating Function (MGF) – since a distribution can be completely determined by moments. It is easier to have a function that can '*generate*' k-th order moment, $E[X^k]$, of a distribution with density function $f(x)$.

To be more specifically, the moment generating function of $X$ is explicitly defined by

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

Apparently, $M_X(t)$ is a function of $t$. As good exercise, you can derive MGF of normal distribution with density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\pi)^2}{2\sigma^2}}, \quad -\infty < x < \infty.$$

and the exponential distribution with density function

$$f(x) = \lambda e^{\lambda x}, \quad 0 < x < \infty$$

This is one of the simplest lifetime distributions (positive random variable).

## 3. Concepts of Survival and Hazard Analysis for Lifetime

One a lifetime distribution is given (i.e., the density function of the random variable that characterize the ageing process is defined), we are interested in several measures used in survival analysis and reliability engineering (that is we lay the foundations and develop new methods for researchers / practitioners in other fields with modeling tools.

You can use the exponential distribution as an example to derive the following measure.

$$f(x) = \lambda e^{\lambda x}, \quad 0 < x < \infty$$

$\lambda$ is the unknown parameter (i.e., a constant).

- Survival Function
  A survival function is a probability function that gives the probability that the lifetime ($X$) goes beyond a time point $x$. The explicit definition is given by

  $$S(x) = P(X > x) = \int_x^{\infty} f(t)\, dt$$

  Where $f(.)$ is the density function of $X$.

  Clearly, $S(x) = 1 - F(x)$. This also implies that $S'(x) = -f(x)$. This survival analysis and reliability analysis, the survival function is the fundamental concept.

  You study the properties of $S(x)$ similarly in terms of monotonicity.

- If we think $S(x)$ plays a similar role like $F(x)$ in non-survival and reliability analysis, what function plays a role to the density function $f(x)$? The function is called hazard function, denoted by $h(x)$, which is used in demography and actuarial science to measure the risks. The definition is given by

  $$h(x) = \frac{f(x)}{S(x)}$$

  In characterizing a new lifetime distribution, we need to provide the explicit form of the hazard function and study the mathematical properties that have practical application or practically meaningful explanations.

This hazard function can also be explained as instantaneous hazard, we might want to know the cumulative hazard.

- Cumulative Hazard Function $H(x)$ is defined based on the survival probability function

$$H(x) = -\log S(x).$$

This is equivalent to

$$S(x) = e^{-H(x)}$$

Since

$$H'(x) = -\frac{S'(x)}{S(x)} = \frac{f(x)}{S(x)}.$$

Therefore,

$$H(x) = \int_0^\infty \frac{f(x)}{S(x)} dx$$

- Mean Survival Time

The mean survival time is just the expectation of the lifetime random variable

$$E(X) = \int_0^\infty x f(x) dx = \int_0^\infty S(x) dx$$

(prove this result!!)

- Mean Residual Lifetime (Expected Future Lifetime) Beyond Time Point $x_0$

You can find how to derive this formula if you like. But this is not important in this project. You only need to know this is an important measure in clinical and reliability analysis and applications.

$$R(x_0) = \int_{x_0}^\infty \frac{S(x)}{S(x_0)} dx$$

If $x_0 = 0$, $R(x_0)$ is simply the mean lifetime.

## 4. The Geometry of Distributions

The shape of density and hazard functions have different applications. We usually provide some properties for the shape of both functions when characterizing the associated distributions such as increasing and decreasing intervals and, particularly the detection of the turning points (change points) – numerical methods for finding the coordinates of the turning points.