Central Limit Theorem/ Estimation Summary

1. Any properly formed (and defined) probability distribution function will have a mean and a variance; if such a distribution is **non-normal** in nature, by virtue of the Central Limit Theorem, it can be **approximated to a normal distribution** if the sample size under investigation is sufficiently large (typically >30).

The mathematics is given as follows:

Let X denote a random variable characterised by a non-normal distribution with mean μ and variance σ^2. Then, if n is large, **by CLT**,

(i) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (ii) $\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \dots\dots\dots + X_n \sim N(n\mu,\ n\sigma^2)$.

Note that version (i) given above is much more commonly examined compared to version (ii), and the clues to accurately detecting the requirement for CLT approximation are:

(a) The absence of a proper label for the probability distribution function provided in the problem (ie **not stated explicitly** that things are normally distributed),

(b) Keywords such as “average” and “mean” surfacing within a sentence structure which is phrased in the style of a question rather than stating a fact. Learn to tell the difference between the two examples given below:

On **average**, an upscale piano store sells 3 baby grand pianos in 2 months.

(This is simply a sentence stating a fact.)

Find the probability that the **mean** number of bananas donated away exceeds 20.

(This sentence is phrased as a question.)

2. The following approximation templates are provided for two popular non-normal distributions (Binomial and Poisson variants). Assume that the size of the sample n extracted is sufficiently large to warrant a CLT conversion.

Binomial Random Variable X with n_0 trials and probability of success p

$$X \sim B(n_0, p) \approx \overline{X} \sim N(n_0 \bullet p, \frac{n_0 \bullet p \bullet q}{n})$$

(Note that there is no minimum value criteria for n_0, CLT ONLY acts upon n)

Poisson Random Variable Y with parameter (within a specified context frame) λ

$$Y \sim P_0(\lambda) \approx \overline{Y} \sim N\left(\lambda, \frac{\lambda}{n}\right)$$

(Note that there is no minimum value criteria for λ, CLT ONLY acts upon n)

Fully worked sample problem to reinforce concept of CLT:

A circular card, with a pointer pivoted at the center, is divided into 5 unequal sectors numbered “1”, “2”, “3”, “4”, and “5”. The pointer is spun and the score will be the number at which the pointer stopped at. The probability of scoring a “5” is $1-q$. The pointer is spun 10 times independently and the number of “5”s obtained is denoted by Y.

Given that $Var(Y) = [E(Y)]^2$, show that $q = \frac{10}{11}$.

Suppose there are 50 people invited to spin the pointer 10 times each. Find the probability that the **mean** number of times they obtain a “5” exceeds 1.

SOLUTIONS :

Let the random variable Y denote the number of “5”s obtained in 10 spins of the pointer.

Then $Y \sim B(10, 1-q)$

Based on this distribution, $E(Y) = 10(1-q)$ and $Var(Y) = 10(1-q)(q)$

Since $Var(Y) = [E(Y)]^2$,

$$10(1-q)(q) = [10(1-q)]^2 = 100(1-q)^2$$

$$(1-q)(q) = 10(1-q)^2$$

$$(1-q)[q - 10(1-q)] = 0$$

$$(1-q)[11q - 10)] = 0$$

$\therefore\ q = 1$ (rejected) or $q = \frac{10}{11}$ (shown)

$$E(Y) = 10\left(1 - \frac{10}{11}\right) = \frac{10}{11},\quad Var(Y) = 10\left(1 - \frac{10}{11}\right)\left(\frac{10}{11}\right) = \frac{100}{121}$$

Let the random variable $\overline{Y}$ denote the mean number of "5"s obtained amongst 50 people.

Since sample size $n = 50$ is large, by **Central Limit Theorem**,

Then $\overline{Y} \sim N\left[\frac{10}{11}, \frac{\left(\frac{100}{121}\right)}{50}\right] = \left(\frac{10}{11}, \frac{2}{121}\right)$ approximately

$$P(\overline{Y} > 1) = 1 - P(Y \le 1) = 1 - 0.7603 = 0.2397 \text{ (shown)}$$

3. In reality, the true mean and variance of a population under study are usually impossible to compute due to the sheer number of members involved and ever changing environmental circumstances. Hence, a more practical methodology **would be to conduct sampling** and calculate unbiased estimates of these parameters. The following formulas are relevant:

Unbiased estimate of population mean $= \overline{x} = \frac{\sum x}{n} = \frac{\sum (x-a)}{n} + a$

Unbiased estimate of population variance $= s^2 = \frac{1}{n-1}\left[\sum \left(x - \overline{x}\right)^2\right]$

$$= \frac{1}{n-1}\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)$$

$$= \frac{1}{n-1}\left(\sum (x-a)^2 - \frac{(\sum (x-a))^2}{n}\right)$$

$$= \frac{n}{n-1} \bullet \text{ sample variance}$$