Tutorial 10

1. This question with steps below will help you to use simulation to illustrate how Central Limit Theorem (CLT) works.

   Consider the income of a population which follow an exponential distribution with mean λ. A random sample of size n is collected from this population to estimate λ. Let $\bar{x}$ denote the sample mean. Theoretically, CLT says: for this study, when sample size n is large enough then $\bar{x}$ will approximately follow a normal distribution with mean λ and standard deviation (sd) $\lambda/\sqrt{n}$.

   Steps below will help you to illustrate the CLT by simulation.

   Let's assume $\lambda = 5000$. Write Python code for each question below:

   (a) Generate N samples, each sample has size n where $N = 100$, $n = 30$. Derive $\bar{x}$ for each sample. Derive the mean and sd of the $\bar{x}$'s from these samples.Is the mean close to $\lambda = 5000$ and the sd close to $\sqrt{\lambda^2/n} = 5000/\sqrt{30}$?

   Hint: to generate a set of n values that follows an exponential distribution with mean λ, we use command: `rexp(n, rate = 1/λ)`.

   (b) Plot histogram of these $\bar{x}$. Does histogram have a bell curve ressembling a normal distribution? You can check the shape and also can use the rule of thumb (about 95% of points lie within 2 sd from the mean) to check.

   (c) Repeat 1b with $N = 1000$ but $n = 100$. Does the histogram ressemble a normal distribution (compare the histogram with the previous one in 1b).

   (d) Repeat 1b with same $N = 1000$ but sample size is dropped, $n = 7$. Does the histogram ressemble a normal distribution? Give your comment about the effect of sample size n to the approximation of $\bar{x}$ distribution to a normal distribution.

   (e) Repeat the above question with $N = 50$ and $n = 100$. Compared to part (c), what do you observe about the distribution of $\bar{x}$ when $N = 50, n = 100$ and when $N = 1000, n = 100$? Conclude about the role of N in the approximation of $\bar{x}$ distribution to a normal distribution.

2. For any significance test, the type II error of that test is defined as "do not reject H_0 when it is wrong" (meaning: the test produce large p-value when H_0 is wrong). The probability of a test to commit type II error is denoted as β. From that, the power of the test is defined as $(1 - \beta)$.

   Suppose that $x_1, ..., x_{25}$ is a random sample from a population of $N(\mu; 4)$. We want to test

   $$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0.$$

   The test statistic $T = \bar{x}/(s/\sqrt{25})$ where $\bar{x}$ and s are respectively the sample mean and sample standard deviation, is used. H_0 will be rejected if $|T| > t_{24}(0.025)$. We call this test as Test 1.

   In R, use simulation to find the power of Test 1 when the true value of μ is given:

   (a)$\mu = 0.5$. (b)$\mu = 1$. (c)$\mu = -0.5$. (d)$\mu = -1$.

   (e)Comment on the results of parts (a) - (d).

   Hint: power = Prob (Test 1 has small p-value | Ho is wrong)

   (1) Each sample has $n = 25$ observations. We get $\bar{x}$ and s from this sample to caculate test statistic $T = \bar{x}/(s/\sqrt{n})$..

   (2) If $T > t_24(0.025)$ then the p-value of Test 1 is smaller than 0.05.

   (3) we will conduct N = 1000 times of Test 1, and count how many of them produce p-value < 0.05. The proportion of Test 1 producing p-value less than 0.05 in N tests is the power of Test 1 derived by simulation.

   (4) The process from (1) - (3) is repeated for different values of $\mu = -1, -0.5, 0.5, 1$.