Statistical Corner

Daniel Yee-Dak FONG 方以德[1], **Chun-Fan LEE** 李俊帆[1], **Wilfred Hing-Sang WONG** 黃慶生[2]
[1]Department of Nursing Studies; [2]Department of Paediatrics & Adolescent Medicine, The University of Hong Kong, Hong Kong

Question 1:
What is a Normal distribution?

Answer :
Normal distribution is used to model the distribution of a continuous measure. It is completely characterized by the mean and the variance, as depicted in Figure 1. A Normal curve is uni-modal and symmetric about the mean. Note however, a symmetric distribution does not necessarily be Normal.

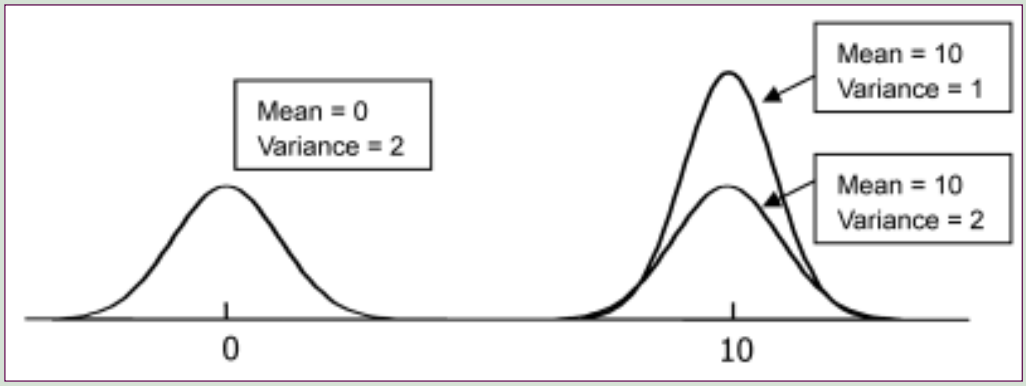

Figure 1. Normal distributions with different means and variances

Question 2:
Is Normal distribution the first distribution known in history?

Answer:
Normal distribution is the most well-known statistical term to not only statisticians but also others who already have a first touch in statistics. However, it was not the first distribution developed in history. Binomial distribution is the one before and indeed the mathematical formulation of the Normal distribution was developed by Abraham De Moivre in 1733 as an approximation to the binomial distribution. In 1809 it was extensively used by Johann CF Gauss in analyzing astronomical data and thus has also been known as Gaussian distribution.

Question 3:
Why do we bother about Normal distribution?

Answer:
Normal distribution has been an underlying assumption in many statistical procedures. For example, the student t-test and ordinary regression analysis are commonly used methods which bear the Normal distribution assumption. Conclusions derived from these statistical procedures without checking for the Normality assumption can be very mis-leading. Therefore, statistical reviewers often ask for justification of statistical procedures adopted and Normal distribution is a common assumption to check.

Question 4:
How can we check if Normal distribution is satisfied?

Answer:
Normal distribution may be checked by either performing a statistical test (e.g. Kolmogorov-Smirnov test and Shaprio-Wilk test) or examining a PP (or QQ) plot. When performing a statistical test (SPSS: Analyze>Descriptive Statistics>Explore, click "Options" button and select "Normality plots with tests"), an insignificant p-value (i.e. >0.05, say) indicates no evidence of departure from Normality. When constructing a PP plot (SPSS: Graphs> P-P), points aligning on the diagonal line (line of perfect Normality) indicates good Normality (Figure 2). Although statistical tests may provide objective conclusions regarding the validity of Normality, they are practically affected by sample size. Specifically, a huge sample will almost always give a highly significant p-value but yet points on the PP plot indicate good Normality. Therefore, it is often more practical to use PP plot for checking Normality although experience in judging departure from Normality is needed.

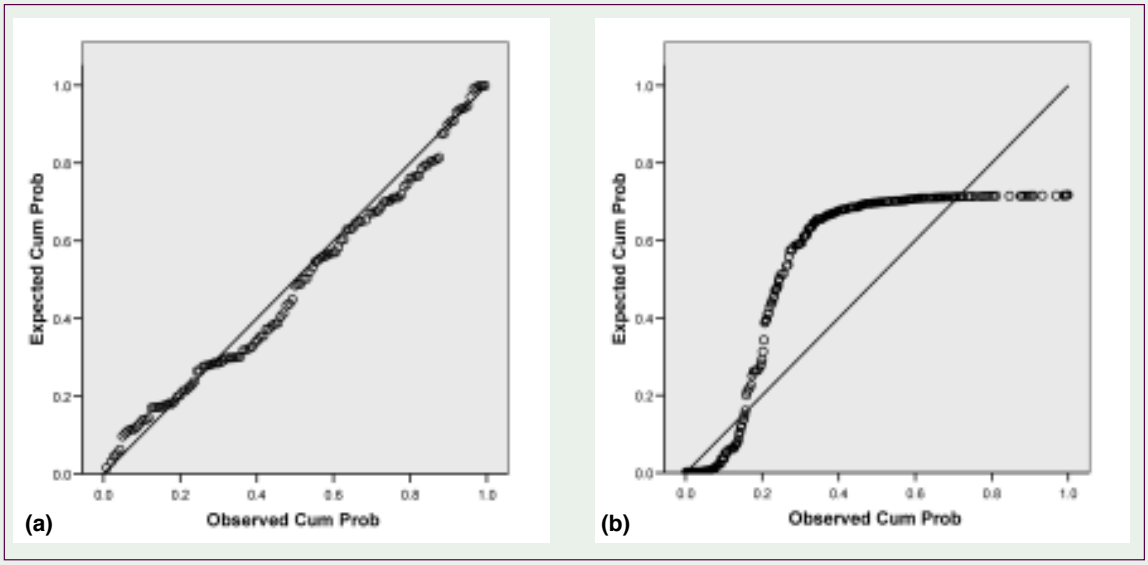

Figure 2. PP plots showing (a) no discernable departure and (b) sever departure, from Normality.

Question 5:
What if Normality is not satisfied?

Answer:
The way to handle non-Normality varies with different statistical procedures being used. Generally, it is preferable to look for aberrant values and determine if any of them are erroneous. If none or errors are rectified but the problem persists, transforming the variable of concern may be attempted. The logarithmic transformation is perhaps the most easily accepted one but there is no guarantee that it works. Other transformations such as the box-cox set of transformations may help but interpretation of results may become difficult. Therefore, if all the interpretable transformations failed, one may need to seek for other statistical methods that do not require the Normal distribution assumption.