A Test for Normality

Suppose you randomly sample n observations $\{x_1, x_2, \ldots, x_n\}$ from a population and want to verify that the underlying population follows a certain distribution with CDF F. For example, suppose $\{x_1, x_2, \ldots, x_n\}$ represent the weights of n randomly selected newborn babies at a hospital, and that we are interested in determining whether this sample gives evidence that the weight of newborns, in general, is approximately Normally distributed (with mean weight 3.5kg and standard deviation 0.4 kg).

1. One way to assess the distribution is to create a **histogram** of the sample, which divides the range of the variable into equal length intervals, counts the number of sampled values in each interval, and plots the results using rectangles with bases on each interval and with heights corresponding to the counts in each interval. A histogram of a random vector `v` can be created in R using the `hist(v)` function. If the population is approximately Normal, then the outline of the histogram should look approximately like the Normal density.

   (a) Simulate a sample of 99 observations from a Normal distribution with mean of 3.5 and standard deviation of 0.4 and create a histogram of the results.

   (b) Create a plot of the PDF of an $N(3.5, 0.4^2)$ variable on the interval corresponding to the observed values of your sample. Compare the shape of the Normal PDF curve to the outline of the histogram. How similar do they look? Based on the shape of the histogram *alone*, would you be confident in saying the underlying population is Normally distributed?

   (c) Repeat parts (a) and (b), but this time with a sample of 9 observations from $N(3.5, 0.4^2)$. Based on the shape of the histogram *alone*, would you be confident in saying the underlying population is Normally distributed?

2. An alternative method to assess distributions is to use a QQ-plot, which plots the ordered values of a sample on one axis against the quantile function of a particular distribution on the other axis.

   In particular, suppose $\{x_1, x_2, \ldots, x_n\}$ are values sampled from a population. Sort the x_i's in ascending order, and let a_1 be the smallest, a_2 be the second smallest, and so on. We want to determine whether the x_i's were sampled from a population whose distribution has CDF F.

   The QQ-plot is the plot of the points

$$\left(a_1, F^{-1}\left(\frac{1}{n+1}\right)\right), \left(a_2, F^{-1}\left(\frac{2}{n+1}\right)\right), \left(a_3, F^{-1}\left(\frac{3}{n+1}\right)\right), \ldots, \left(a_m, F^{-1}\left(\frac{m}{n+1}\right)\right),$$

   (a) If the $\{x_1, \ldots, x_n\}$ were sampled from a population with CDF F, what is the expected value of $F^{-1}\left(\frac{i}{n+1}\right)$? (Refer back to Problem 2 from CA 10-10).

   (b) Based on part (a), if the $\{x_1, \ldots, x_n\}$ were sampled from a population with CDF F, what should the shape of the QQ plot look like?

   (c) To verify your conjecture in the previous part:

      i. Generate a sample of 99 points from a $N(3.5, 0.4^2)$ distribution.

      ii. Use the `sort` function in R to create a vector `a` that puts this sample in ascending order.

      iii. Write a function called `my_quantile` in R which takes a probability `p` and outputs the value of the quantile function at `p` for $N(3.5, 0.4^2)$.

      iv. Create a vector `u` of the values $\frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}$ using the `seq` function in R.

      v. Create a QQ plot of `a` against the the function `my_quantile` evaluated at the vector `u`.

   (d) Repeat part (c) 2 or 3 more times to see how much the shape of QQ-plot varies just due to random sampling.

   (e) Repeat part (c), but with 9 points instead of 99 (you will also need to change the vector `u`).

   (f) Now, we'll see what happens when the $\{x_1, \ldots, x_n\}$ were sampled from a population whose CDF is **not** F:

      i. Generate a sample of 99 points from the Uniform $[0, 1]$ distribution using `runif`.

      ii. Use the `sort` function in R to create a vector `a` that puts this sample in ascending order.

      iii. Write a function called `my_quantile` in R which takes a probability `p` and outputs the value of the quantile function at `p` for $N(3.5, 0.4^2)$.

      iv. Create a vector `u` of the values $\frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}$ using the `seq` function in R.

      v. Create a QQ plot of `a` against the the function `my_quantile` evaluated at the vector `u`.