NRES 798 — Lab 5

Distributions, summary statistics

Graph the PDF $f(x)$, CDF $F(x)$ and quantile function $F^{-1}(p)$ for the normal, uniform, log-normal, and exponential distributions (`?Distributions`). Hint: `curve(dnorm(x, 0, 1), from=-3, to=3)`. See what effect the distribution parameters have. Expressions can be conveniently plotted with `curve`; what is the difference between `curve` and `plot`?

Generate a large sample (say 100000 values) from a standard uniform distribution: `x <- runif(100000)`. Examine `x`: do a `summary`, `head(x)`, plot a histogram (`hist(x)`). What is the (theoretical) population mean and the mean of the 100000 values? The population variance is known to be 1/12; what about `var(x)`?

Think now of `x` as a population (a Bayesian would say that any population is a sample, or *realization*, from some superpopulation). Add another standard uniform random value to each of the 100000 that you have: `x <- x + runif(100000)`. Think about it, each x_i is now the sum of 2 (nearly) uniform random variables (independent, if the random number generator is any good). What should the (population) mean of these sums be? (expected value of a sum of 2 RVs). And the variance? Compare with your sample. Plot a histogram.

Repeat `x <- x + runif(100000)`, so that now you have sums of 3 variables. Find the theoretical and observed means and variances. Plot histogram.

Repeat a few more times (hint: up-arrow). What is going on? What does the distribution look like?

Do the same with an exponential instead of the uniform. Then try a Bernouilli for some `p`, `x <- sample(1:0, 100000, replace=T, prob=c(p, 1-p))`, or `x <- rbinom(100000, 1, p)` (why?). Amazing, huh?

Generate your sample of linphid spider tibial spine lengths: `tibia <- rnorm(50, 0.253, 0.0039)` (mm, parameters from Fig. 2.6). Display, summarize, "histogramize". Do `plot(x)`. Draw a box plot (`??box`), compare to the numbers from `summary`.

Compute and compare the arithmetic mean, geometric mean, harmonic mean, and median for your measurements. Do it the long way using the definitions, and using any built-in functions that you might find. For the median you can use `sort`. And indexing (`[]`) if you are too lazy to count.

For the GM, check that averaging logs gives the same as using products. Find the mode for histograms with various numbers of bins (`breaks` parameter).

Compute the two versions (denominator n and $n-1$) of the variance and standard deviation, the long way from the definitions, and with built-in functions. How do they, and the sample mean, compare to the population values (see above, used to generate the sample)? How can you calculate the values with denominator n from those with $n-1$, or vice-versa, after you lost the data? Verify.

Define an *R* function `RMS <- function(x) ...` to compute the *root mean square*, the square root of the mean square (= the standard deviation using denominator n). Test it.

Define a function `Skewness(x)` to compute the skewness. The s in the formula is the s from equation (3.7), not that of (3.9) or (3.10). You can use `mean` and `RMS`. Do the same for the kurtosis (again, s is the RMS). Try it out on `tibia`. The population value for a normal is 0 in both cases.

Generate another tibia lengths sample, same as before. Hint: start typing `tibia <-` and press *Ctrl + up-arrow* (or in *RStudio*, look up the command in the *History* tab, upper-right). Compute the skewness and kurtosis. Repeat a few times to get a feeling of the variability of these statistics.

Let's do this better, run a simulation, e.g., for the kurtosis: First, make a vector to store the results, `k <- rep(x=0, times=10000)`. Now the loop, `for(i in 1:10000) {tibia <- rnorm(50, 0.253, 0.0039); k[i] <- Skewness(tibia)}`. Now `k` has 10000 values of skewness in samples of size 50. Plot a histogram, calculate the standard deviation (standard error), etc.

Do `plot(ecdf(tibia))`, and `quantile(tibia, seq(0, 1, 0.1))`. Look up these functions in the *Help*.

If you still have time, do some more experimenting and exploration around the above.