**Detecting Expense Fraud**

Using the data file “expenses.xlsx” (either from Resources Session 5 or downloaded from [here](https://docs.google.com/a/hawaii.edu/spreadsheets/d/1cuz7bN1Jz-4VPXDZZ35HQNacA7dunAV-SWVb1i4G5BM/edit?usp=sharing)) compute the distribution of first digits and some descriptive statistics for the first and second digits. In principle Transaction level data such as disbursements or sales are expected to obey Benford’s Law in distribution of the digits for the data values (see <https://en.wikipedia.org/wiki/Benford%27s_law> for more details). The Benford’s Law density function of 1st digits is Prob[1st digit = n] = log10(1+1/n) where n=1,2,...,9. Some properties (these are not parameters because there is only one Benford distribution):

mean 3.440

variance 6.057

skewness 0.796

kurtosis −0.548

The density function exact probabilities for the joint occurrence of the first two digits according to Benford's law is known but a bit more complex than we need to deal with, however it is useful to know the population correlation between the first and second digits is ρ = 0.0561.

* 1. Compare the distributions of the 1st digits for Expenses1 and Expenses2 with the Benford’s Law density. Discuss how well each seems to fit the Benford’s Law density and interpret what this may imply about the expense data.

**Use polygon (line graph) rather than histograms (bar chart) to compare.**

benlaw <- **function**(d) log10(1 + 1 / d)

digits <- 1:9

baseBarplot <- barplot(benlaw(digits), names.arg = digits, xlab = "First Digit", ylim = c(0, .35))

firstDigit <- function(x) substr(gsub('[0.]', '', x), 1, 1)

pctFirstDigit <- function(x) data.frame(table(firstDigit(x)) / length(x))

de1 <- pctFirstDigit(Expenses$amount1)

head(de1)

names(de1)[1] = "Digit"

lines(x = baseBarplot[,1], y = de1$Freq, col = "green", lwd = 4, type = "b", pch = 23, cex = 1.5, bg = "green")

de2 <- pctFirstDigit(Expenses$amount2)

lines(x = baseBarplot[,1], y = de2$Freq, col = "blue", lwd = 4, type = "b", pch = 23, cex = 1.5, bg = "blue")

* 1. Generate X1, X2, X3 ~ U[0,100] for N = length(Expenses$amount1) . Plot the distribution of the 1st digits for X1^2 / X2 \* X3. What do you observe and what can you conclude from this?

N <- length(Expenses$amount1)

df3 <- pctFirstDigit(runif(N, 0, 100)^2 / runif(N, 0, 100) \* runif(N, 0, 100))

lines(x = baseBarplot[,1], y = df3$Freq, col = "blue", lwd = 4,

type = "b", pch = 23, cex = 1.5, bg = "blue")

* 1. Use the descriptive stats and the properties of the Benford distribution to confirm your findings from (a). Give some indication of your confidence and why.

multi.fun **<-** **function(**x**)** **{**

      c**(**mean **=** mean**(**x**)**, var **=** var**(**x**)**, skew **=** skewness**(**x**),** kurt **=** kurtosis**(**x**) )**

**}**

sapply(de1, multi.fun)

* 1. Now load the benford.analysis package. Explore the results of using this analysis

e1 <- benford(Expenses$amount1)

e2 <- benford(Expenses$amount1)

**For computational details see** <http://investexcel.net/benfords-law-excel/>