The University of Hong Kong
School of Public Health

CMED 6020 Advanced Statistical Methods I
TUTORIAL 1 – R / regression in R

---

1. Central limit theorem - the mean of a sufficiently large number of identically distributed independent random variables each with finite mean μ and variance σ^2 will be approximately normally distributed with mean μ and variance σ^2/n irrespective of the shape of the original distribution.

   This can be demonstrated in a simulated experiment:

   (a) Simulate 3 random variables from a uniform distribution U(1,3) and calculate the mean.

   (b) To understand the distribution of the mean, we can simulate 1000 such means (ie, 1000 means, each from 3 random variables)

   [Hint: store all random numbers in a 3 × 1000 matrix]

   (c) Calculate the mean and variance of the 1000 means and plot their distribution.

   (d) The theoretical mean and variance for U(1,3) distribution is 2 and 1/3 respectively. Compare the mean μ and variance σ^2/n with the results in (c).

   (e) Instead of simulating 3 random variables, simulate 30 random variables for each mean. Repeat part (c) for comparison.

   (f) Repeat the above using the following bimodal discrete distribution:

   $P(X=1) = P(X=5)=0.1$
   $P(X=2) = P(X=4)=0.35$
   $P(X=3) = 0.1$

   (g) Perform a statistical test for the normality of the means in (b), (e) and (f).

2. Save the dataset “mvc” in your local computer. (URL: http://web.hku.hk/~ehylau/mvc.csv) Read the dataset into R.

   (a) Define a variable which categorized the male alcoholics into younger (≤ 40y) and older adults (> 40y). Name the variable as “younger”.

   (b) Calculate the mean MVC for the two age groups.

(c) Draw a boxplot of MVC by age categories.

(d) Draw a scatterplot between height and MVC. Add a linear regression line in the figure to show the relation.

(e) Draw a 2x2 panel of scatterplot showing the scatterplot and regression linear as in (d) for male alcoholics aged > 20, 30, 40 and 50y respectively. For comparison purpose, use the same limits for the x and y-axes. Please also label the figures.

(f) Add the linear regression equations in the figures.

(g) It is proposed that a quadratic relation between MVC and height may exist. Fit a linear regression to test this hypothesis.

(h) Compare the model with or without quadratic terms of height using AIC.

(i) Perform stepwise selection for the model with predictors age, height and height^2. [Hint: may also use the function stepAIC in the package "MASS"]

(j) Based on the fitted model with age and height (without the squared term) only as predictors, calculate the predicted MVC for a male alcoholic of age 50y and height of 170cm. [Hint: may use the function "predict"]

(k) Predict the MVC for a male alcoholic with the same age but with height 220cm. Compare the prediction intervals with (j).