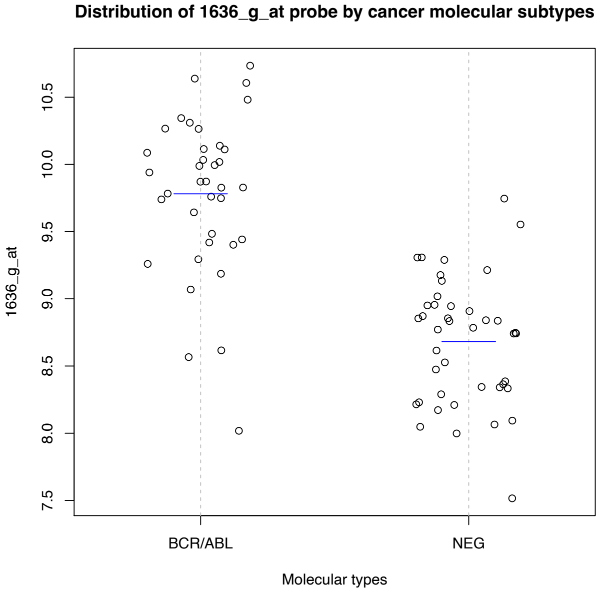
Homework Assignment 3
(Due Thursday, September 28, 2023 at 5PM)

Please email your answer (compiled pdf file from R markdown) and R code to Cenxiao (CENXIAO@email.sc.edu).

1. Departure from Hardy-Weinberg equilibrium is often a sign of potential genotyping problems. Among 1,000 study subjects, we see 649 AA, 300 AB, and 51 BB at one locus, and 640 AA, 360AB and 0 BB at another locus. Use χ2 test to find out whether or not the loci are in Hardy-Weinberg equilibrium. (5 points)
2. Use the FAMuss data to test if there are differences in nondominant arm muscle strength (NDRM.CH) between actn3_r577x genotype groups.
   >fmsURL<-“http://people.stat.sc.edu/hoyen/STAT588/Data/FMS_data.txt”
   > fms<-read.delim(file=fmsURL, header=TRUE, sep="\t")
   (a) Plot the data using NDRM.CH and actn3_r577x. [Hint: use stripchart] (5 points)
   (b) Obtain the analysis of variance table, and comment on the results. (5 points) [Hint: Use aov(y ~ x) where y is NDRM.CH and x is actn3_r577x genotype.]
   (c) Remove two biggest observations in each genotype group, repeat (a) and (b). (5 points)
3. Use the ALL dataset and create the following plot. (10 points) [Hint: Use the R code in http://people.stat.sc.edu/hoyen/STAT588/Lab6.R to access the expression data. stripchart(y~x, method= "jitter", jitter=0.2, vertical=T, ylab=⋯, main=⋯) where y is the expression data and x is the cancer molecular subtypes (mol.biol). Use ylab to label y-axis correctly and main to create a main title. For the blue lines indicating means in each group, use lines(c(x1,x2), c(y1,y2), col=4) where x1, x2, y1, y2 are the locations of the line in x-axis and y-axis respectively.]



4. This exercise is for practicing central limit theorem.
   (a) Draw n=5 samples from uniform distribution and calculate sample means. Repeat this experiment 200 times, plot the distribution of sample means. (5 points) [Hint: To simulate n samples from uniform distribution, use runif(n). Use plot(density(x)), where x is the vector contains the sample means from these 200 experiments.]
   (b) Repeat (a) but use n=100 (2 points).
   (c) Compare the sample distributions obtained in (a) and (b), what do you observe? (3 points)

5. Perform the following steps in R:
   (a) Simulate 30 samples from Normal(mean=0, sd=1) (2 points)
   (b) Randomly assign 15 samples into control and 15 into treatment group (10 points) [Hint: Use sample]
   (c) Perform two sample T-test and report the p value. (2 points)
   (d) Randomly generate 1000 samples from uniform distribution, and plot the histogram of the 1000 samples. [Hint: Use hist(x) to plot a histogram of x.] (2 points)
   (e) Repeat (a) (b) (c) 1000 times, and stored the corresponding 1000 p values in a vector, plot a histogram using these 1000 p values. What is the distribution of p values? (10 points)