# 36-669 HW 1

## Q0: Survey

### 0.A

I preferred to be called "Dan" although Dan or Daniel is fine.

### 0.B

I have read and understood the entire syllabus.

### 0.C

Statistics Courses I've taken: Since I am a graduate student and have taken these courses elsewhere, I will list the title and what I think is the CMU Statistics Department equivalent next to it:

Statistics I & II (36-200 and 36-202)

Mathematical Theory of Probability (36-225)

Mathematical Theory of Statistics (36-226)

Regression Methods (36-401)

Introduction to Statistical Programming

### 0.D

Course that have taught me coding:

Introduction to Computer Science

Introduction to Statistical Programming

Regression Methods

Econometrics

### 0.E

None

### 0.F

I hope to learn more about how statistics can be applied to genetic data to make inferences and improve my writing and software skills. I am also hoping to use the project to use the research project to investigate the inheritability of diseases that are afflicting my immediate family, specifically Inflammatory Bowel Disease.

## Q1: Basic Analysis in R

### 1.A

Define the following terms:

SNP: SNP stands for single nucleotide polymorphism. It is a single nucleotide change and is responsible for most variation in the genome.

Gene: A gene is a unit of heredity that is transferred from a parent to an offspring and is held to determine characteristics of the offspring.

GWAS: Genome-wide association studies. These are used to study common disease caused by many small genetic effects (can test up to millions of SNPs at a time). The Manhattan Plot can be used to visualize results of GWAS.

Allele: An allele is one of two or more possible versions of a gene. Individuals inherit two alleles for each gene, one from each parent.

Genotype: A genotype is a pair of alleles inherited from the parents by the offspring. It is the specific makeup of the genome and may not be readily apparent by the feature of a subject. Subjects of different genotypes can have the same phenotype.

Phenotype: A phenotype is an appearance or feature of a subject. Knowing the phenotype does not necessarily imply the knowing the genotype.

Recombination: Process by which pieces of DNA are broken and recombined to produce new combinations of alleles. This creates genetic diversity at the level of genes that reflects differences in DNA sequences.

Mutation: Change in the DNA that results in the genetic background of an individual. Mutations can still occur in the DNA of offspring of parents without being present. They can also appear after many generations of offspring when the mutation occurred.

Linkage Disequilibrium: Spatial correlation in chromosomes that is generated over time from recombination.

### 1.B

```
# loading in the data set
famuss <-  read.csv("https://raw.githubusercontent.com/xuranw/469_public/master/data/famuss.csv")
# enumerate names of the columns in the data
colnames(famuss)
```

```
##  [1] "id"                   "actn3_rs540874"       "actn3_rs1815739"
##  [4] "actn3_1671064"        "adrb2_rs1042718"      "akt1_t22932c"
##  [7] "akt1_g15129a"         "akt1_c10744t_c12886t" "akt1_t10726c_t12868c"
## [10] "akt1_t10598a_t12740a" "akt1_c9756a_c11898t"  "akt1_t8407g"
## [13] "akt1_a7699g"          "akt1_c6024t_c8166t"   "akt1_g2347t_g205t"
## [16] "akt1_g2375a_g233a"    "akt1_g4362c"          "akt1_g22187a"
## [19] "akt2_rs892118"        "akt2_2304186"         "akt2_969531"
## [22] "ankrd6_q122e"         "ankrd6_m485l"         "ankrd6_p636l"
## [25] "ankrd6_t233m"         "ankrd6_i128l"         "ankrd6_g197805a"
## [28] "ankrd6_a545t"         "ankrd6_k710x"         "bcl6_4686467"
## [31] "bcl6_3774298"         "bcl6_17797517"        "bcl6_1056932"
## [34] "bmp2_rs15705"         "c8orf68_rs6983944"    "carp_a8470g"
## [37] "carp_c105t"           "cast_rs754615"        "cast_rs7724759"
## [40] "cntf_g6a"             "ctsf_572846"          "ddit_rs1053227"
```

```
## [43] "esr1_rs1801132"       "esr1_rs1042717"       "esr1_rs2228480"
## [46] "esr1_rs2077647"       "fbox32_rs6690663"     "fbox32_rs3739287"
## [49] "fbox32_rs4871385"     "fstl1_rs9631455"      "gnb3_rs5443"
## [52] "igf1_pro"             "igf2_rs680"           "il15ra_3136618"
## [55] "il15_1057972"         "il15_1589241"         "il15_rs2296135"
## [58] "insig2_rs7566605"     "irs1_g972r"           "kchj11_rs5219"
## [61] "mylk_c37885a"         "mylk_g91689t"         "myod1_rs2249104"
## [64] "nos3_rs1799983"       "p2ry2_rs1783596"      "pik3_rs3173908"
## [67] "ppara_1800206"        "ppar_gp12a"           "pparg_c1a_rs8192678"
## [70] "rs302964"             "tcfl72_12255372"      "tcfl72_7903146"
## [73] "tcfl72_rs12255372"    "tcfl72_rs7903146"     "tpd52l1_3778458"
## [76] "tpd52l1_514096"       "tpd52l1_4896782"      "tdp52l1_9321028"
## [79] "tpd52l1_3799736"      "vdr_rs731236"         "Gender"
## [82] "Age"                  "Race"
```

Based on the results above, we can determine the number of SNPs in the database as: SNPs in database = Total Number of Cols - Cols for (ID, Gender, Age, Race) = 83 - 4 = 79

```
# determine which column corresponds to SNP actn3_1671064
which(colnames(famuss) == "actn3_1671064")
```

```
## [1] 4
```

```
# using the table function
table(famuss[,4])
```

```
##
##  AA  GA  GG
## 169 262 107
```

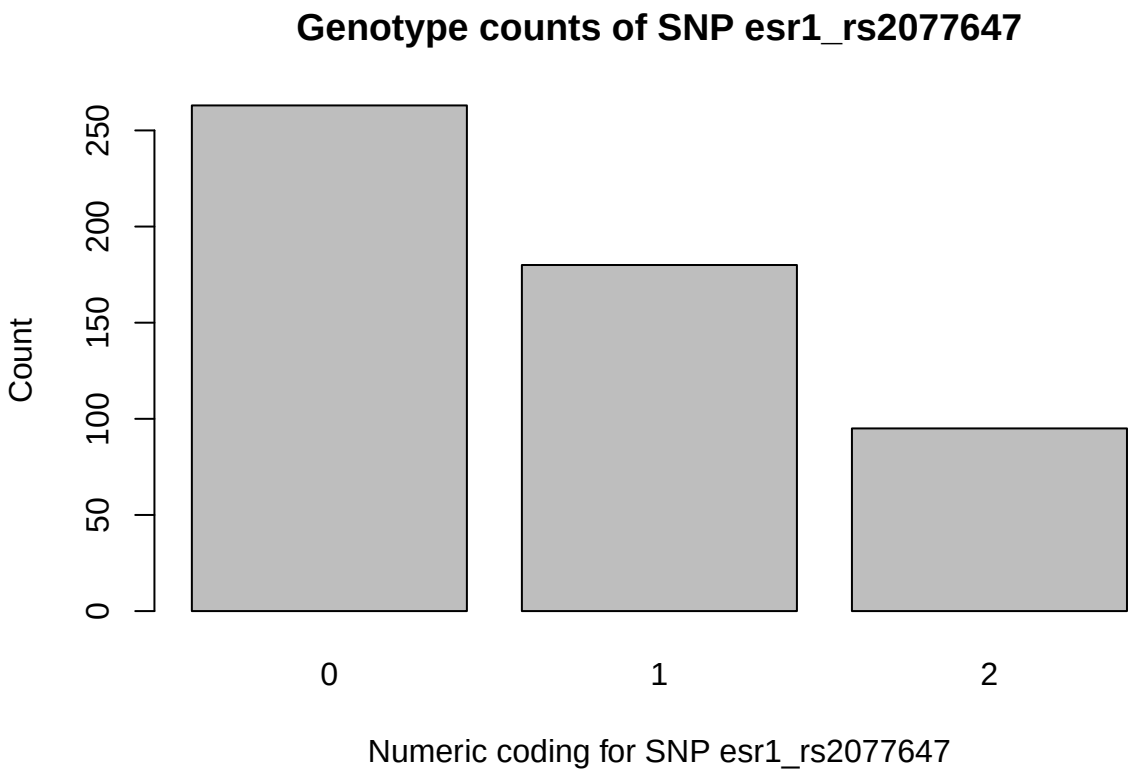
```
# table function stratified by race
table(famuss[,4], famuss$Race)
```

```
##
##      African Am Asian Caucasian Hispanic Other
##   AA         13    15       127        7     7
##   GA          5    28       211        9     9
##   GG          4    13        80        5     5
```

### 1.C

Creating a barplot for SNP esr1_rs2077647

```
x <- as.vector(table(famuss$esr1_rs2077647))
names(x) <- c(1,0,2)
x <- sort(x, decreasing = T)
barplot(x, xlab = "Numeric coding for SNP esr1_rs2077647",
        ylab = "Count",
        main = "Genotype counts of SNP esr1_rs2077647")
```




## Q2: Empirical verification on the Central Limit Theorem

### 2.A

```
# reading the generate_data function into R
source("https://raw.githubusercontent.com/xuranw/469_public/master/hw1/clt.R")
```

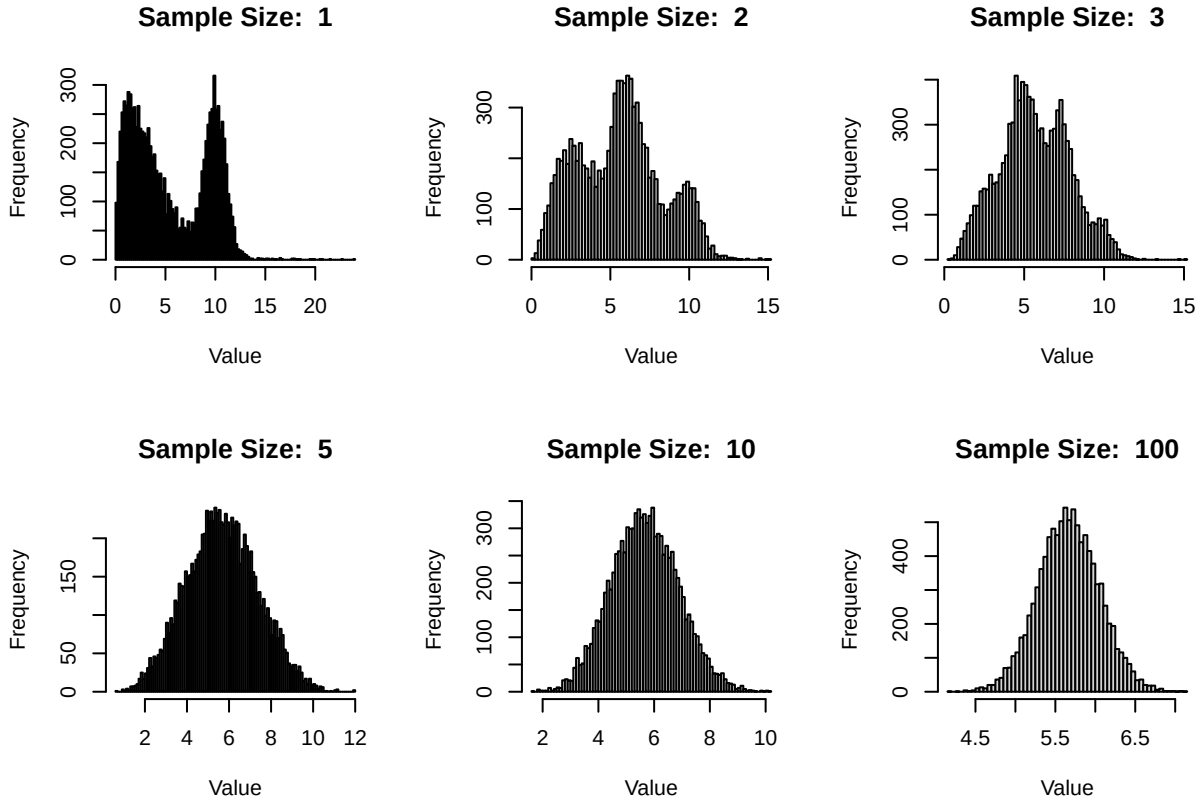
Description of the function: The "generate_data" function takes a positive integer as an input and returns an output of a vector with the length of inputted integer where each entry is a random number. The random numbers are generated from a normal distribution with mean 10 and standard deviation 1, a gamma distribution with shape parameter 2 and scale parameter 2, and a chi-square distribution with 3 degrees of freedom. The frequency with which numbers from these distributions are sampled is determined by the sample function built into R.

### 2.B

```
# creating a data frame and assigning names to col
hist_df <- data.frame(vector("numeric", 10000), vector("numeric", 10000),
                      vector("numeric", 10000), vector("numeric", 10000),
                      vector("numeric", 10000), vector("numeric", 10000))
```

```r
# filling out the data frame
z <- c(1,2,3,5,10,100)
for(i in seq_along(z)){
  hist_df[,i] <- replicate(10000, mean(generate_data(z[i])))
}

# creating histograms for each sample size
par(mfrow = c(2,3))
for(i in seq_along(z)){
  hist(hist_df[,i], breaks = 100, xlab = "Value", main = paste("Sample Size: ", z[i], sep = " "))
}
```




The plots help to verify the Central Limit Theorem because it dictates that the sampling distribution of the sample mean is approximately normally distributed and centered around the population mean as sample size approaches infinity. This is evident by the shape of the plot as sample size increases from which the sample mean is calculated, as it more closely resembles the normal distribution.