Bioinformatics 525: Module 2

Introduction to Statistics

Lab Session #1

1. **Introduction to R**

2. **Data entry, calculations and graphics.** Enter the following height and weight data in R.

Height	Weight
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150

a. Derive BMI=Weight(Kg)/(Height(m)*Height(m)) Or BMI=703 x Weight(lb)/(Height(in)*Height(in))

b. Calculate the following:

	Weight N=	Height N=	BMI N=
Mean			
Median			
Variance			
SD			
Q1			
Q3			
IQR			
Min			
Max			

c. Graph the histograms, boxplot, and Q-Q plot for, weight, height, and BMI using multi figure format (par(mfrow=c(3,3))) . Do weight, height, and BMI look normally distributed?

d. Use Shapiro.test() to test if weight, height, and BMI are normally distributed. Which variables are not normally distributed?

e. Calculate the overweight variable: owt=1 if BMI > 25 and 0 otherwise. What is the number and the percent of subjects who are overweight (BMI > 25)?

Note: To use help in R type help(hist) or example(hist) to find out more about hist() functions. Same applies for any function in R.

3. **TRial Of Preventing HYpertention (TROPHY) Study.**

We will use data from TROPHY study to apply some of the methods presented during the class.

- **Brief Introduction**

TROPHY was an investigator-initiated study to examine whether early treatment of prehypertension, might prevent or delay the development of subsequent incident hypertension.

- **Objective**

The primary objective of the study was to determine whether in patients with prehypertension two years of treatment with candesartan (at a dose of 16 mg daily) reduces the incidence of hypertension at the end of 2 year treatment and at two years after the discontinuation of active treatment.

- **Data Set**

The data set is in text format, TROPHY.csv.

a. Read TROPHY.csv data in RStudio, then type attach(TROPHY).

Type dim(TROPHY) to get the number of rows (observations) and the number of columns (variables) for this data set. Type head(TROPHY) to look at variables name. The following subset of variables is part of this data.

Variable Name	Code
Smoke	Smoking status at baseline: 1=yes/2=no
Age	Age in years at baseline
BMI	Body Mass Index at baseline
Insulin	Insulin at baseline
Gluc_fast	Fasting Glucose at Baseline
Ins_gluc	Insulin:Glucose Ratio at baseline
Triglyceride	Triglyceride at baseline
HDL	High Density Lipoprotein Cholesterol
LDL	Low Density Lipoprotein Cholesterol
HDL_LDL	HDL:LDL Ratio at baseline
Cholesterol	Total Cholesterol at baseline
DBP0	Systolic Blood pressure at baseline
SBP0	Systolic Blood pressure at baseline
BMI24	Body Mass Index at 24 months follow-up
DBP24	Systolic Blood pressure at 24 months follow-up
SBP24	Systolic Blood pressure at 24 months follow-up
HT	Hypertension status at 24 months follow-up: 1=yes/0=No
Trt	1=Candesartan/2=Placebo

b. What is the mean, sd, Median, Q1, Q3, IQR, Min and Max for baseline blood pressure (DBP0) for each treatment group?

c. Look at the histograms, boxplot and Q-Q plot to see if DBP0 and SBP0 are normally distributed

d. Use both graphical display (hist, boxplot, q-q plot) and the Shapiro Wilks test to show that HDL is not normally distributed.

e. Does a log transformation, lHDL<-log(HDL), result in normality for lHDL?

4. **Simulations.** Use simulations in R to illustrate that a linear combination of two normally distributed random variables is normally distributed.
That is, if $z_1 \sim N(m_1, s^2_1)$ and $z_2 \sim N(m_2, s^2_1)$ then $z=a*z_1+b*z_2 \sim N(a*m_1+b*m_2, a^2*s^2_1+b^2*s^2_1)$.

   a. Simulate 1000 data points for $z_1 \sim N(m=1,sd=1)$ and $z_2 \sim N(m=2,sd=2)$

   b. Calculate $z=3*z_1+2*z_2$

   c. Use both graphical display tools and the Shapiro Wilks test to test whether z is normal.

   d. Is the mean of z equal to 7(=3*1+2*2); the variance=25(=9*1+4*4);sd=5?

e. Simulate 10000 data points y ~ $N(7,5)$.

f. Use summary() function to compare z and y

g. Use side-by-side boxplot to visually compare the distribution of z and y. Are they the same?

5. **Simulations.** Use simulations in R to illustrate central limit theorem

a. Simulate 100 data points for k=(5,30,100) random variable x1,x2,…xk ~ *Bernoulli(.3).*
The generated data will be a matrix, with k columns (for x1, x2, … ,xk) and 100 rows.

b. Calculate xsum= of x1+x2+…+xk

c. Look at histogram, Q-Q plot of xsum, does it look normal for different values of k?

d. Use Shapiro test to test normality for xsum for different k