CSC591: Foundations of Data Science
HW1: Exploratory Analysis, Basic Probability, Random Variables and Probability Distributions

Released: 9/7/15
Due: 9/21/15 (23:55pm). (One day late: -25%; -100% after that).

Student Name: Parth Satra
Student ID: 200062999 (pasatra)

Notes

- Filename: Lastname_StudentID.pdf (only pdf).
- You can also submit scanned hand written solution (should be legible, TA's interpretation is final).
- This h/w is worth 5% of total grade.
- You can discuss with your friends, but solution should be yours.
- Any kind of copying will result in 0 grade (minimum penalty), serious cases will be referred to appropriate authority.
- All submission must be through Moodle (you can email to TA with cc to Instructor – only if these is a problem – if not received on time, then standard late submission rules apply)
- No makeups or bonus; for regarding policies, refer to syllabus and 1st day lecture slides.

Q#	Max Points
1	10
2	8
3	12
4	15
5	10
6	5
7	10
8	10
9	10
10	10

Q1. Describe concisely various forms of data preparation (or preprocessing). (**10** points)

Answer

Data preprocessing is a broad aspect. It mainly involves selecting data objects and attributes for analysis or modifying attributes to improve the time, cost and quality of the analysis. Following are the various forms of data preparation.

- **Aggregation**: Aggregation helps in combining multiple data objects. Aggregation thus helps in reducing the size of a very large data set. Thus data reduction now requires less memory and processing time and enables usage of more expensive data mining algorithms. Aggregation also provides a different higher level perspective of the data.
- **Sampling**: Sampling is a technique of selecting a subset from the actual dataset to be analyzed. Motivation behind sampling is to enable analyzes which otherwise would be too expensive and highly time consuming. Sampling can give the same analytical results as the original dataset if the sample is completely representative. The sample is completely representative if the sample has same properties as the representative data set. The simplest form of sampling is random sampling but random sampling might not guarantee the same properties.
- **Dimensionality Reduction**: Dimensionality reduction is yet another technique of data preprocessing. A data set typically can have very large number of features. Dimensionality reduction aims at reducing the features in such a way that these fewer features are also successful in representing the properties of the dataset. This mainly helps in removing the noise from the dataset and many data mining algorithms work better at lower dimensions. Data with higher dimensions also suffer from the curse of dimensionality. Curse of dimensionality is a phenomenon where the data becomes increasingly sparse in the space it holds. This makes analytics increasingly difficult with increase in the dimensionality.
- **Feature Subset Selection**: This is another preprocessing step at reducing the dimensionality of the dataset. It aims at only selecting a subset of the features. This step is especially useful when there are a lot of irrelevant and duplicate features. Though this preprocessing step may as well result in information loss if the wrong subset is selected. To get a better feature subset each subset is iteratively evaluated to see if the subset matches a particular criterion. This step also provides the same benefits as the previous step with lesser dimensionality.
- **Discretization and binarization**: Some data mining algorithms require data to have categorical attributes especially for classification algorithms. This step mainly aims at converting continuous attributes into discrete categories or binary categories and converting discrete attributes into binary categories for better efficiency.
- **Variable transformation**: Variable transformation aims at transforming values of a particular attribute for all the data objects. These are simple transformations like Normalization and standardization. Normalization helps in getting values of the within range like 0-1. Similarly, standardization would be to get percentage values of the marks across all systems to be able to compare them.

Q2. A company wants to know the customer satisfaction and conducts survey over 100 customers. The survey form includes the following questions. (**8** points)

Survey Form:

a) Are you: Male Female (Circle one of the choice)
b) How old are you? ______ (in years)
c) How much do you spend on groceries? _____ (in $$$$.$$)
d) How much do you spend on soft drinks? ___ (in $$$$.$$)
e) Which soft beverage do you prefer? _____ (Coke, Pepsi, Dr. Pepper, …)
f) How satisfied are you with diet beverages? _____ (Very satisfied, Satisfied, Not Satisfied)
g) How likely are you to buy 6-pack diet coke?___ (Very likely, Likely, Not Likely, Very unlikely)

Assume that all surveyed customers returned survey forms with correct answers.

Answer the following questions:

Your objective is to:

1. Design the database (one table) and enter the data. Show (draw) table with few data entries.
2. For each resulting column (attribute), list the type of attribute (in terms of Nominal, Ordinal, Interval, Ratio)
3. For each attribute, what kind summary (statics) make sense?
4. For each attribute, what kind of graphical representation makes most sense? (e.g., pie chart, bar chart, …)

Answer

Following is the database table that contains a few data entries.

Gender	Age	Groceries Expenditure	Soft Drinks Expenditure	Preferred Soft Drink	Satisfaction of Diet beverages	Likelihood to buy diet coke
Male	**20**	**50**	**10**	**Coke**	**Not Satisfied**	**Not Likely**
Female	**28**	**60**	**10**	**Pepsi**	**Very Satisfied**	**Very Likely**
Male	**50**	**30**	**5**	**Dr. Pepper**	**Satisfied**	**Likely**
Male	**35**	**50**	**15**	**Pepsi**	**Satisfied**	**Not Likely**
Female	**18**	**30**	**12**	**Coke**	**Not Satisfied**	**Very unlikely**

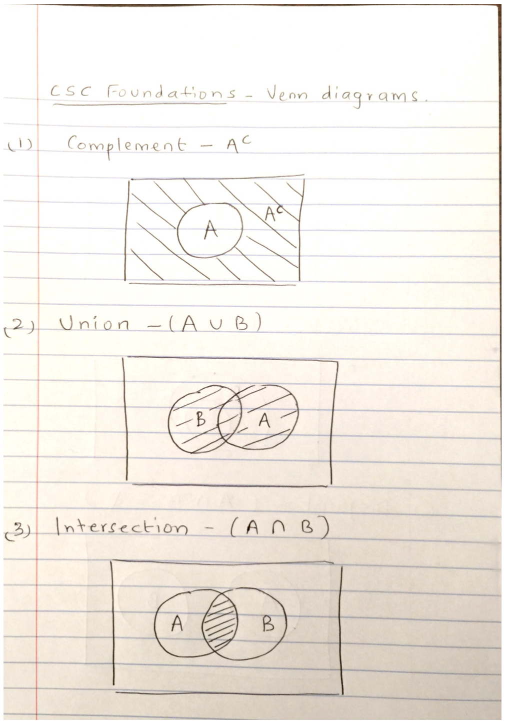
All the attribute related information is present in the below table

Attributes	Attribute Type	Summary Statistics	Graphical Representation
Gender	Nominal	mode, entropy, contingency, correlation.	Pie charts, histograms. These can be used since the attribute has low cardinality
Age	Ratio	Geometric mean, harmonic mean, percent variation	Scatter Plots. Since these attributes have high cardinality
Grocery Expenditure	Ratio	Geometric mean, harmonic mean, percent variation	Scatter Plots. Since these attributes have high cardinality
Soft Drink Expenditure	Ratio	Geometric mean, harmonic mean, percent variation	Scatter Plots. Since these attributes have high cardinality
Preferred Soft Drink	Nominal	mode, entropy, contingency, correlation.	Histograms, Pie charts. These can be used since the attribute has low cardinality
Satisfaction of diet drinks	Ordinal	Median, percentile, rank correlation, run test	Pie charts, Histograms. These can be used since the attribute has low cardinality
Likelihood of getting diet drink	Ordinal	Median, percentile, rank correlation, run	Pie charts, Histograms. These can be used since the attribute has low cardinality

Q3. Draw Venn diagrams for basic set operations (defined on slide 10, 8/27/15) (**12** points)

Answer

Following Venn Diagrams show all the set operations specified in the slide.

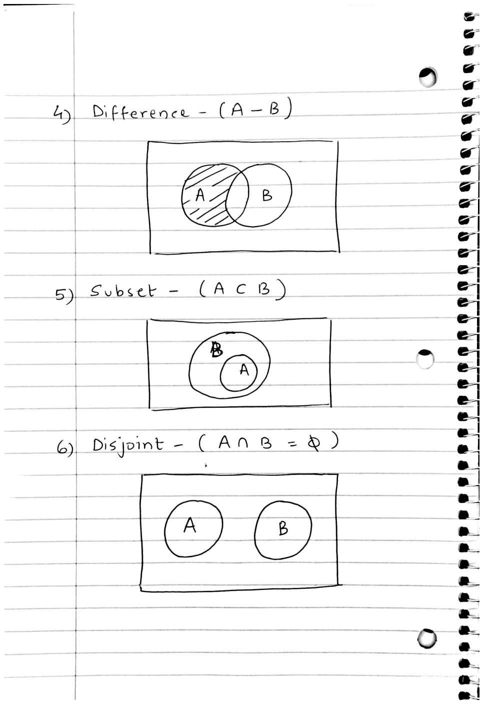


CSC Foundations - Venn diagrams.

(1) Complement — A^c

(2) Union —$(A \cup B)$

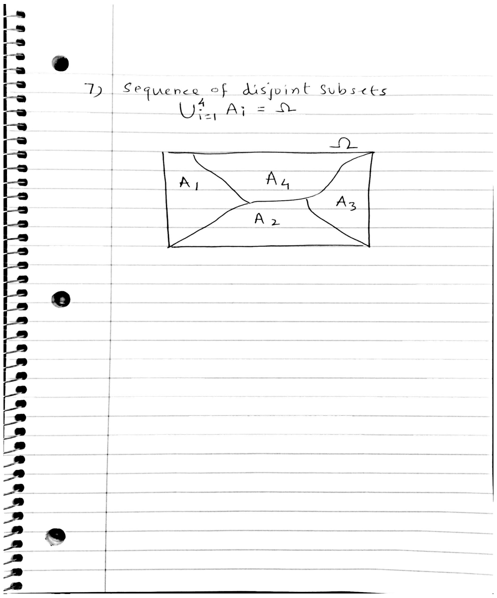
(3) Intersection - $(A \cap B)$



4) Difference - $(A - B)$

5) Subset - $(A \subset B)$

6) Disjoint - $(A \cap B = \phi)$



7) Sequence of disjoint subsets

$\bigcup_{i=1}^{4} A_i = \Omega$

Q4. (**15** points) The probability of a union: For any two events A and B, Prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Answer

The Union of two sets A and B can be defined as

$$A \cup B = A \cup (A^c \cap B)$$

Thus we have defined $A \cup B$ as two disjoint sets. Thus probability of $A \cup B$ will be

$$P(A \cup B) = P(A \cup (A^c \cap B))$$

Now since A and $(A^c \cap B)$ are disjoint sets we get

$$P(A \cup B) = P(A) + P(A^c \cap B) \qquad - \qquad (I)$$

Now Probability of B can be defined as

$$P(B) = P(A^c \cap B) + P(A \cap B)$$

Thus

$$P(A^c \cap B) = P(B) - P(A \cap B) \qquad - \qquad (II)$$

Now substituting II in I we get

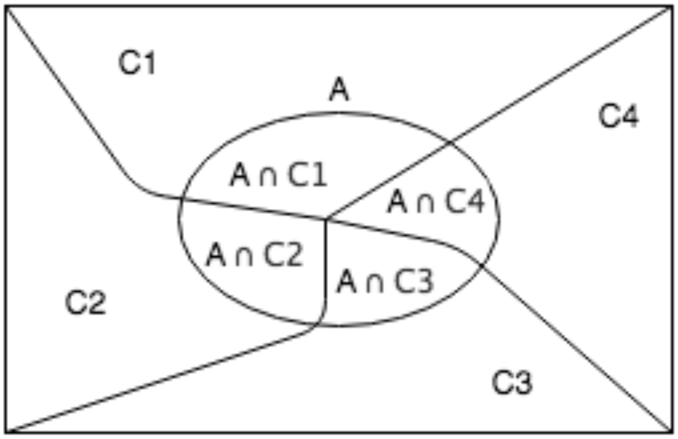
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This proves that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Q5. Law of total probability: Suppose $C_1, C_2, \ldots, C_m$ are disjoint events such that C_1 U C_2 U … U $C_m = \Omega$. The probability of an arbitrary event A can be expressed as: $P(A) = P(A| C_1)P(C_1) + P(A| C_2)P(C_2) + \ldots + P(A| C_m)P(C_m)$. (**10** points)

- Illustrate this law using Venn diagram (for m=4) and derive P(A) using this Venn diagram

Answer



The above is the Venn diagram showing the disjoint events C1, C2, C3 and C4 along with an arbitrary event A. The probability of the event A can be expressed as

$$P(A) = P((A \cap C1) \cup (A \cap C2) \cup (A \cap C3) \cup (A \cap C4))$$

Since C1, C2, C3 and C4 are disjoint and thus each of the above 4 components is disjoint, we can write the probability as

$$P(A) = P(A \cap C1) + P(A \cap C2) + P(A \cap C3) + P(A \cap C4) \qquad - \qquad (I)$$

Now using multiple rule each of these components can be defined as

$$P(A \cap C1) = P(A | C1) \cdot P(C1) \qquad - \qquad (II)$$

$$P(A \cap C2) = P(A | C2) \cdot P(C2) \qquad - \qquad (III)$$

$$P(A \cap C3) = P(A | C3) \cdot P(C3) \qquad - \qquad (IV)$$

$$P(A \cap C4) = P(A | C4) \cdot P(C4) \qquad - \qquad (V)$$

From I, II, III, IV, V we get

$$P(A) = P(A | C1) \cdot P(C1) + \ldots + P(A | C4) \cdot P(C4)$$

Thus the law of total probability can be verified for m = 4 using the given Venn Diagram.

Q6. Let $\Omega = \{a,b,c\}$ be a sample space. Let P(a) = ½, P(b) = 1/3, and P(c) = 1/6. Find probabilities for all subsets of Ω. (**5** points)

Answer

We know that probability of Ω is always 1. Thus

$$P(\Omega) = 1$$

Let's find the sum of P(a), P(b) and P(c). So,

$$P(a) + P(b) + P(c) = 1/2 + 1/3 + 1/6 = 1$$

This means that all the sets a, b, and c are disjoint.

Thus the subsets of Ω can be calculated as follows

$$P(a) = ½$$

$$P(b) = 1/3$$

$$P(c) = 1/6$$

$$P(a \cap b) = 0$$

$$P(a \cap c) = 0$$

$$P(b \cap c) = 0$$

$$P(a \cap b \cap c) = 0$$

$$P(a \cup b) = ½ + 1/3 = 5/6$$

$P(a \cup c) = ½ + 1/6 = 2/3$

$P(b \cup c) = 1/3 + 1/6 = ½$

$P(a \cup b \cup c) = ½ + 1/3 + 1/6 = 1$

Q7. Janak and Madhu are taking FDS course. The course has only three grades: A, B, and C. The probability that Janak gets a B is .3. The probability that Madhu gets a B is .4. The probability that neither gets an A but at least one gets a B is .1. What is the probability that at least one gets a B but neither gets a C? (**10** points)

Answer

Let J_A, J_B, J_C represent events such that Janak gets A, B and C respectively.

Similarly let M_A, M_B and M_C represent events such that Madhu gets A, B and C respectively.

There are totally following outcomes possible for grades of Janak and Madhu. With this reference table we can formulate the given information as below.

$P(J_A \wedge M_A)$	$P(J_B \wedge M_A)$	$P(J_C \wedge M_A)$
$P(J_A \wedge M_B)$	$P(J_B \wedge M_B)$	$P(J_C \wedge M_B)$
$P(J_A \wedge M_C)$	$P(J_B \wedge M_C)$	$P(J_C \wedge M_C)$

Given probability of Janak getting B

$P(J_B) = 0.3$

$P(J_B) = P(J_B \wedge M_A) + P(J_B \wedge M_B) + P(J_B \wedge M_C)$

$P(J_B \wedge M_A) + P(J_B \wedge M_B) + P(J_B \wedge M_C) = 0.3$ - I

Similarly given probability of Madhu getting B

$P(M_B) = P(J_A \wedge M_B) + P(J_B \wedge M_B) + P(J_C \wedge M_B) = 0.4$ - II

Given probability of at least one getting B and neither getting A

$P(J_B \wedge M_B) + P(J_C \wedge M_B) + P(J_B \wedge M_C) = 0.1$ - III

Now subtracting III from I we get

$P(J_B \wedge M_A) - P(J_C \wedge M_B) = 03 – 0.1 = 0.2$ - IV

Finally adding II and IV we get

$P(J_B \wedge M_A) + P(J_A \wedge M_B) + P(J_B \wedge M_B) = 0.2 + 0.4 = 0.6$

Thus the probability that at least one getting B and neither getting C = **0.6**

Q8. Let L be the event that a student was "born in a long month," (i.e., month with 31 days) and R be "born in a month (full name) with the letter r." Then estimate the following two quantities. (**10** points)

(1) We know that a randomly chosen student in class was born in a long month, then what is the probability that he was also born in a month with letter r.

(2) If we have same information as above (that is a person is born in long month), then what is the probability that the person was also born in a month without letter "r".

Answer

1. We know that randomly chosen student in the class was born in long month. We need to find the probability of him being born in a month with letter r. Thus we need to find P(R | L).
   Now we know that there are 7 months in a year which have long months, namely January, March, May, July, August, October, December.
   Thus, P(L) = 7/12
   Now there are 4 months in a year which have both long months and have r in those months, namely January, March, October and December.
   Thus $P(R \cap L) = 4/12$.
   Now $P(R | L) = P(R \cap L) / P(L)$ by definition of conditional probability.
   Thus P(R | L) = (4/12) / (7/12) = **4/7**.

2. We know that as in above case the student is born in long month. But this time we need to find the probability of him being in month without r. Thus we need to find $P(R^c | L)$.
   This is a complement of the above computed probability. Thus the probability of can be computed as
   $P(R^c | L) = 1 - P(R | L)$
   = 1 - 4/7 = **3/7**.

Q9. Let y1 = $[1\ 1\ 1]^T$, y2 = $[0\ 1\ -1]^T$, and y3 = $[1\ 4\ -2]^T$. Determine if the vectors y1, y2, and y3 are linearly dependent (or independent)? (**10** points)

Answer

According to the definition of linear independence the above vectors are linearly independent only if a = b = c = 0 is the only possible solution for the equation

a * y1 + b * y2 + c * y3 = 0 - (I)

From I we get following 3 equations

a + c = 0 - (II)

a + b + 4c = 0 - (III)

a – b – 2c = 0 - (IV)

Now let us assume a = 1. Thus from equation II we have

c = -1

Now substituting the values of a and c in IV we get

1 – b + 2 = 0

Thus b = 3

Now substituting a =1, b = 3 and c = -1 satisfies the equation III.

1 + 3 – 4 = 0

Thus we have a non-zero model (a = 1, b = 3, c = -1) that satisfies these equations with non-zero values and hence the system of vectors is **linearly dependent**.

Q10. (R Project): For all questions, also include your r code in the solution. (**10** points)

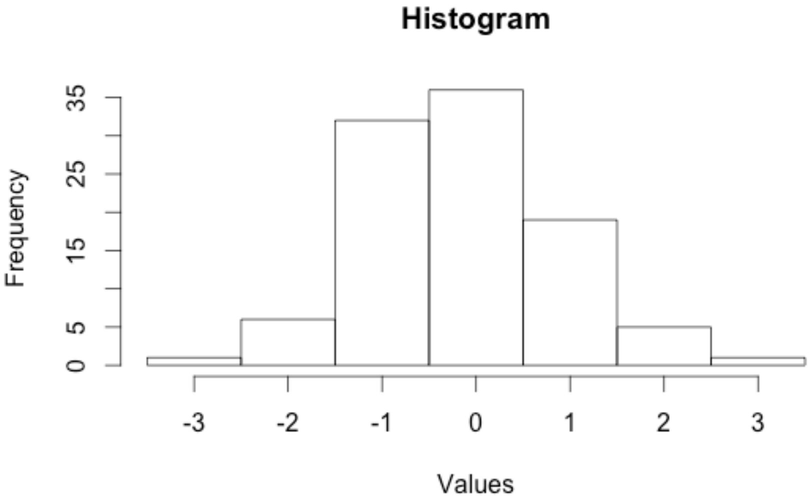
(1) Generate 100 random samples from a univariate normal distribution (default parameters), and compute mean, sd, variance, median, and standard quantiles. Report the quantities.

Output:

```
> samples = rnorm(100)
> mean(samples)
[1] -0.1801087
> sd(samples)
[1] 0.9965239
> var(samples)
[1] 0.9930599
> median(samples)
[1] -0.2150791
> quantile(samples)
        0%        25%        50%        75%       100% 
-2.9258143 -0.9205981 -0.2150791  0.4711117  2.6317879 
```

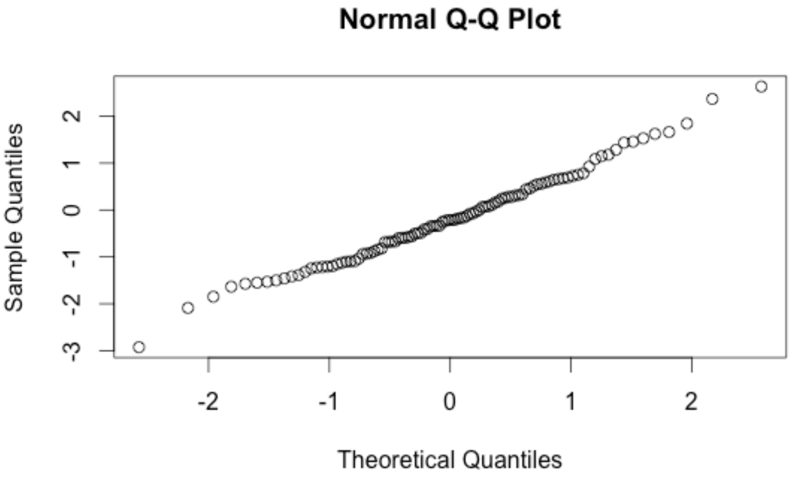
(2) Plot histogram (experiment with breaks parameter), include only one best plot.

Output:



(3) As the data is generated from a normal distribution, one would expect that the data follow a normal distribution, but how to test it. Try “qqnorm” function (understand it first), include your plot. Comment on if the data follows a normal distribution (why or why not)?

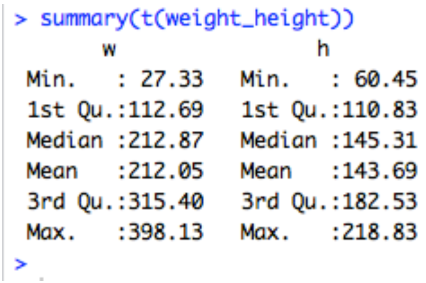
Output:



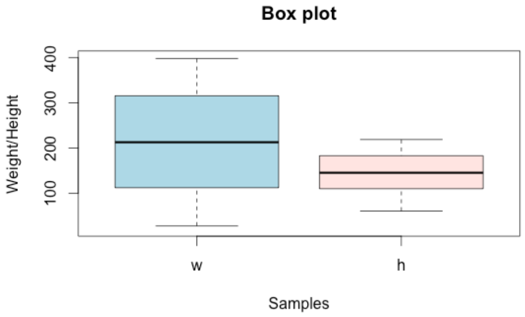
The above diagram shows the “qqnorm” plot of the sample data and the theoretical data with normal distribution. Since the theoretical and sample data match along the line x = y, indicating that the two distributions are same and hence the data follows a normal distribution.

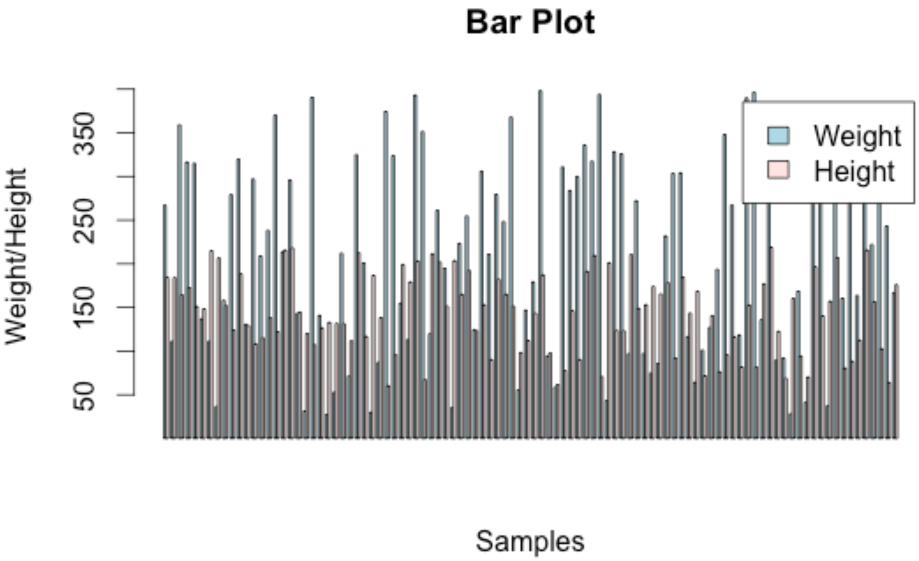
(4) Generate data (100 samples) for 2 random variable (height and weight; that is 2d data) using reasonable parameters. Using this 2d data to, (a) generate summary statistics, (b) generate a bar plots, (c) generate box plots, and (d) generate a scatter plot

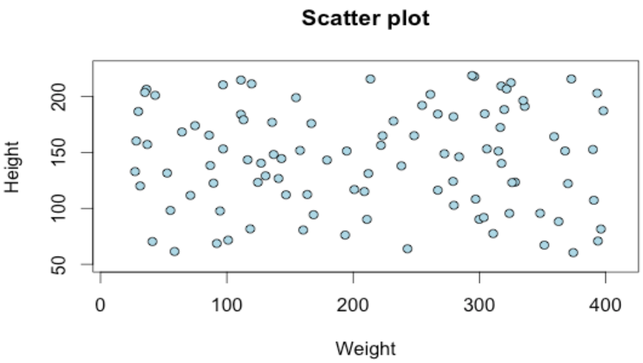
Output:



```
> summary(t(weight_height))
       w                h         
 Min.   : 27.33   Min.   : 60.45  
 1st Qu.:112.69   1st Qu.:110.83  
 Median :212.87   Median :145.31  
 Mean   :212.05   Mean   :143.69  
 3rd Qu.:315.40   3rd Qu.:182.53  
 Max.   :398.13   Max.   :218.83  
>
```

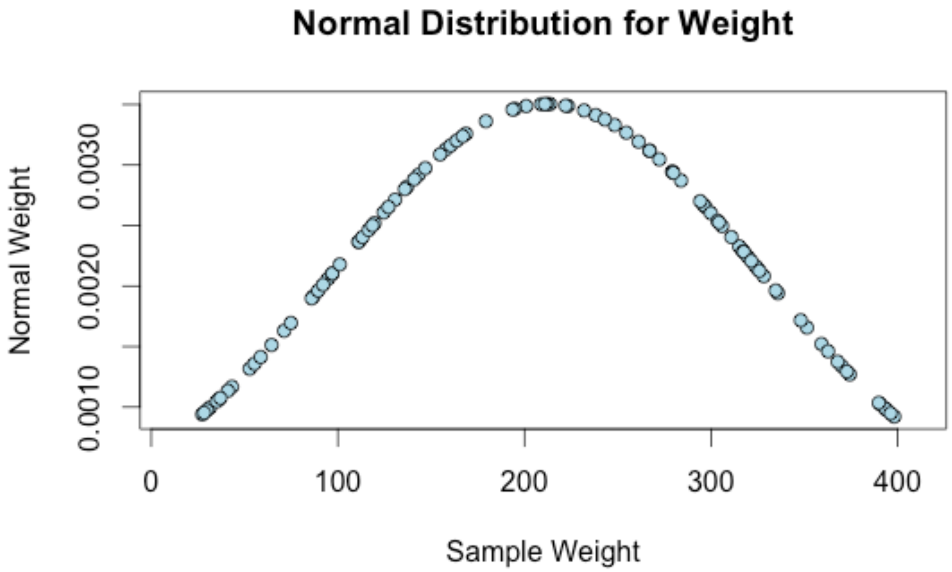


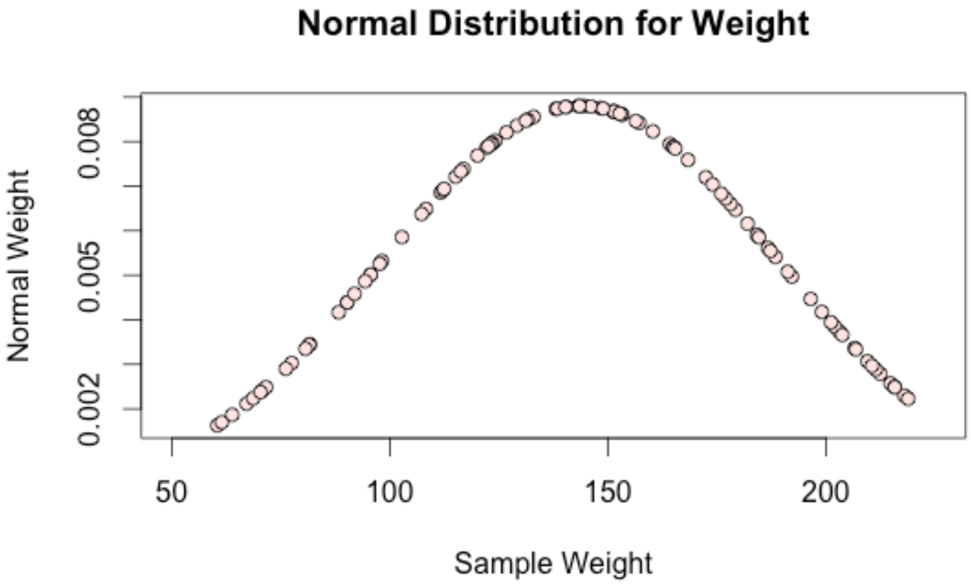




(5) Plot normal distribution for each variable (height and weight) from the 2d data generated above (10.4)

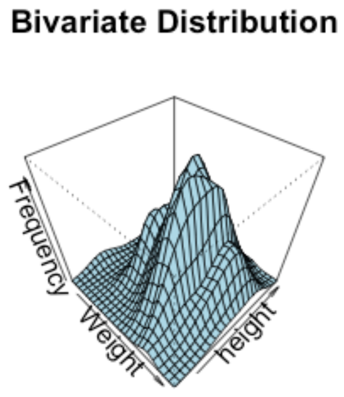
Output:

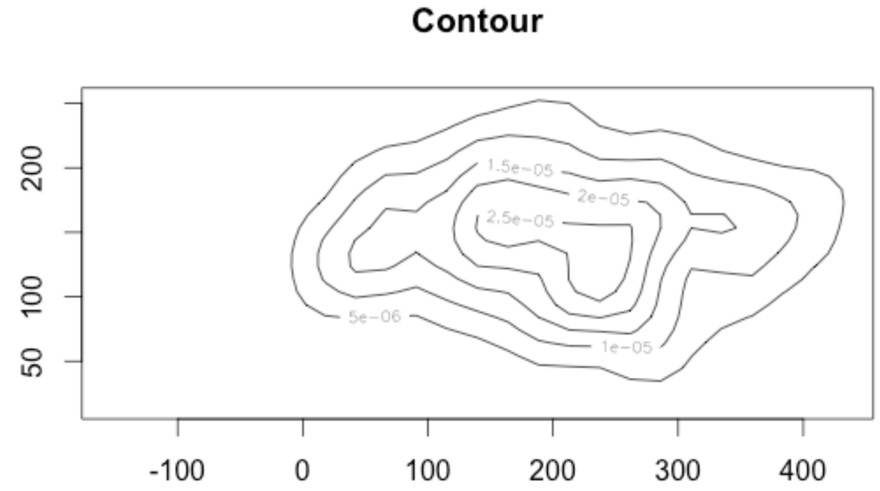




(6) Plot bivariate normal distribution for the above data (data generated in (10.4)).

Output:





The above diagrams the bivariate normal distribution of the data and the contour shows the 2D representation of the data. As the contour is in the form of an eclipse the distribution is normal and the shape will improve and become more evident with increase in the number of samples.

R Code

```
# clean the environment
rm(list = ls())

# include library
library(MASS)

# Question 10.1
# Generating samples with normal distribution with mean = 0 and standard deviation
= 1.
samples = rnorm(100)
# mean
mean(samples)
# Standard Deviation
sd(samples)
# Variance
var(samples)
# Median
median(samples)
# Standard Quantiles
quantile(samples)

# Question 10.2
b = c(-3.5,-2.5,-1.5,-0.5,0.5,1.5,2.5,3.5)
# Generating Histogram
histogram = hist(samples, breaks = b)
plot(histogram, main="Histogram", xlab = "Values", ylab = "Frequency")

# Question 10.3
qqnorm(samples)

# Question 10.4
# (a) Generate weight in lbs (20lbs to 400 lbs) and height in cm (2ft to 7ft)
w = runif(100, 20, 400)
h = runif(100, 60, 220)
weight_height = rbind(w, h)
# Generate Summary
summary(t(weight_height))

# (b) Generate bar plots
barplot(weight_height, beside = TRUE, col = c("lightblue","mistyrose"),
      legend=c("Weight", "Height"), ylim=c(20, 400),
      ylab = "Weight/Height", xlab = "Samples", main = "Bar Plot",)
```

Code Continue

```
# (c) Generate box plots
boxplot(t(weight_height), col = c("lightblue", "mistyrose"),
       legend=c("Weight", "Height"), ylim = c(20, 400),
       ylab = "Weight/Height", xlab = "Samples", main = "Box plot")

# (d) Generate Scatter Plots
plot(weight_height[1,], weight_height[2,], bg = c("lightblue"),
    pch = 21, ylim = c(50, 225), xlim = c(10, 410),
    ylab = "Height", xlab = "Weight", main = "Scatter plot")

# Question 10.5
# Plotting the normal distribution for weight
normal_weight = dnorm(weight_height[1,], m = mean(weight_height[1,]),
                   sd = sd(weight_height[1,]))
plot(weight_height[1,], normal_weight, bg = c("lightblue"),
    pch = 21, xlim = c(10,410), xlab = "Sample Weight",
    ylab = "Normal Weight", main = "Normal Distribution for Weight")

# Plotting the normal distribution for height
normal_height = dnorm(weight_height[2,], m = mean(weight_height[2,]),
                   sd = sd(weight_height[2,]))
plot(weight_height[2,], normal_height, bg = c("mistyrose"),
    pch = 21, xlim = c(50,225), xlab = "Sample Weight",
    ylab = "Normal Weight", main = "Normal Distribution for Weight")

# Question 10.6
# Plot bivariate normal distribution
mu = c(mean(weight_height[1,]), mean(weight_height[2,]))
sigma = matrix(cov(t(weight_height)),2,2)
bivariate_distribution = mvrnorm(100, mu, sigma)
bivariate_distribution.kde = kde2d(bivariate_distribution[,1],
bivariate_distribution[,2], 100)
persp(bivariate_distribution.kde,
     phi = 45,
     theta = 45,
     col = "lightblue",
     xlab = "Weight",
     ylab = "height",
     zlab = "Frequency",
     main = "Bivariate Distribution")
contour(bivariate_distribution.kde, main = "Contour")
```