

Sample Answer Lab 04

Task 1: Descriptive Statistics and Relation to the Mixture Distributions of Random Variables (2.5 points)

Recreate several mixture distributions shown in Table 1 with 1,000 observations using the R-script `MixtureDistribExercise.r`. The function `makeMixDist()` generates the data and the function `mixHist()` plots the a histograms and two box-plots.

Your *intellectual challenge* here is to select a proper set parameters for the two parent normal distributions, i.e., their means and standard deviations, and the mixture proportion of both parent distributions, so that you obtain the distributions shown below.

[a] Find approximately the set of parameters that generate the similar histograms of the mixture distribution. Also insert their associated group-wise box-plots. Select approximate the same landscape plot frame for your box-plots. Each plot here has a width of 3.5". (1 point)

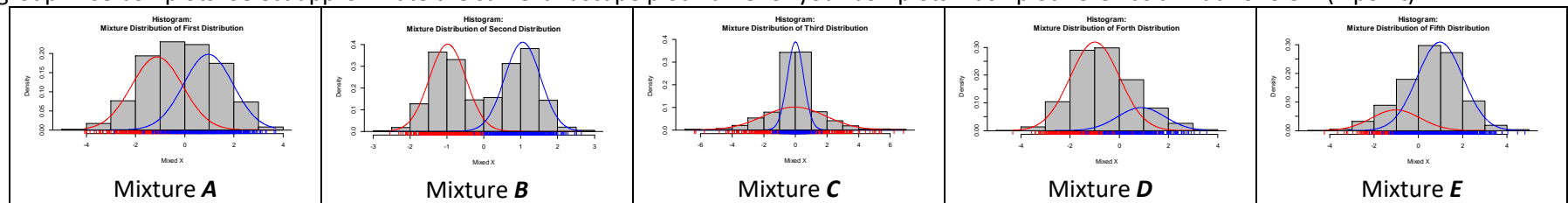
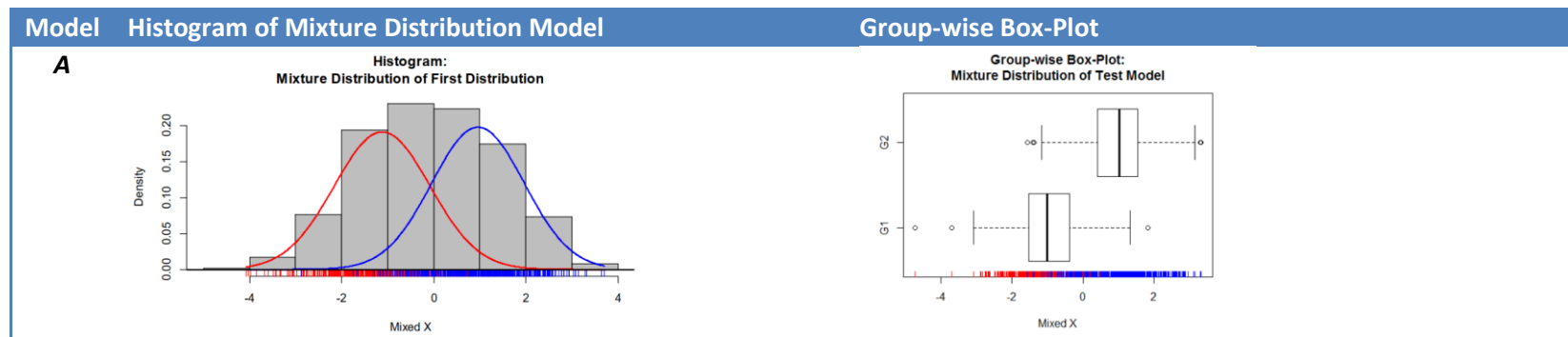
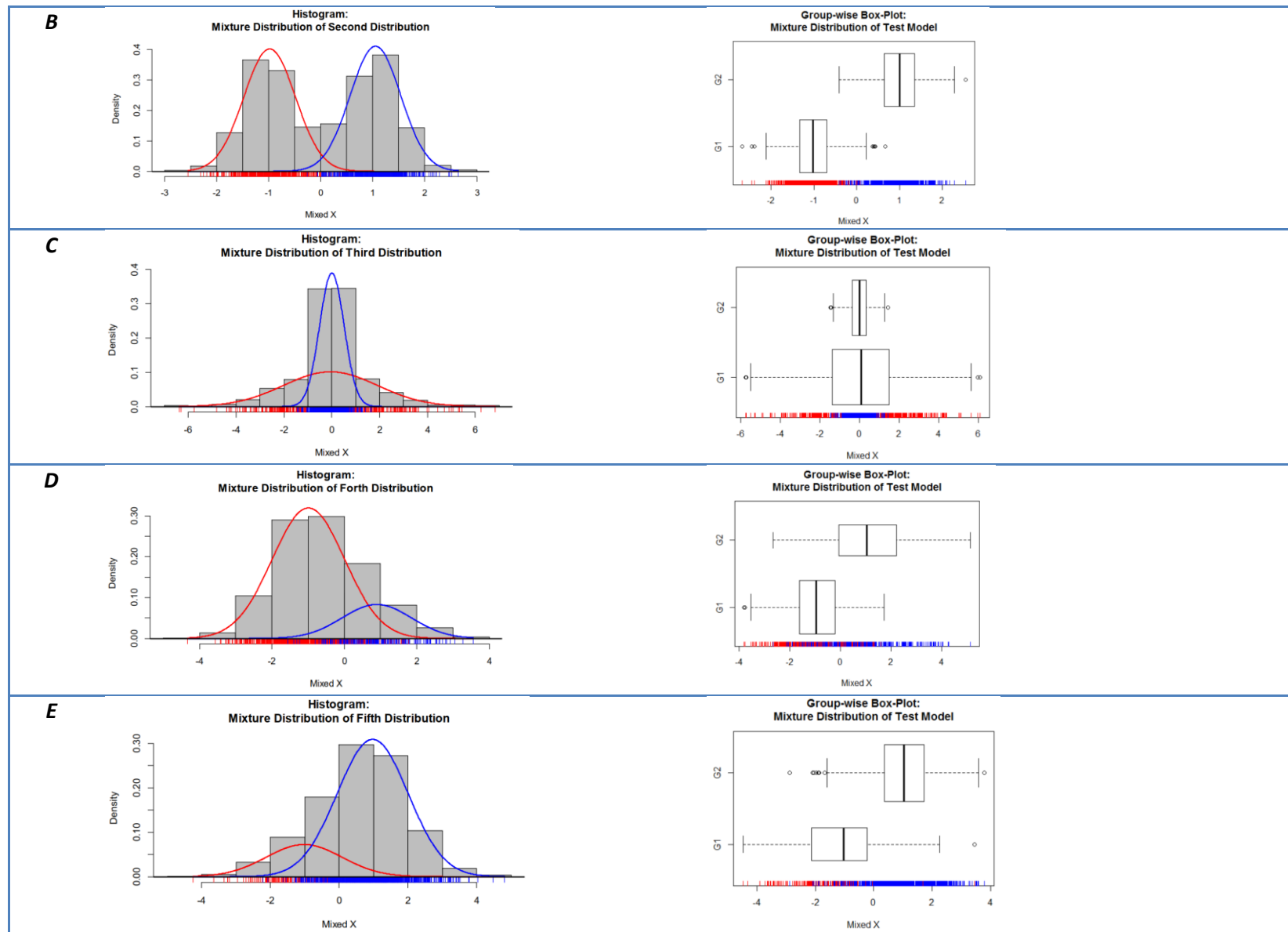


Table 1: Target Mixture Distributions

Note: Record the parameters of each distribution in the tables at task 1 [b] and 1 [c].

Insert the histograms of the mixture distributions and the group-wise box-plots here:





[b] Report your selected parameters of each of the mixture distributions models. (0.5 points)

Model	Mixture Proportion	μ_1	σ_1	μ_2	σ_2
A	0.5	-1	1	1	1
B	0.5	-1	0.6	1	0.6
C	0.5	0	2	0	0.5
D	0.75	-1	1	1	1.1
E	0.25	-1	1	1	1

[c] Report in a table the estimated statistics describing the joint distribution of each of the five mixture distribution models. (0.5 points)

Model	Mean	Winsorized Mean at 10%	Standard Deviation	Skewness	Kurtosis	Bimodality Index
A	-0.047	-0.061	1.389	0.091	-0.503	0.402
B	0.016	0.020	1.163	-0.014	-1.141	0.522
C	0.047	0.035	1.518	0.207	2.676	0.183
D	-0.464	-0.522	1.307	0.401	0.203	0.362
E	0.515	0.549	1.339	-0.257	-0.044	0.360

[d] Explain how the skewness, kurtosis and bimodality index relate to the selected parameters that you have selected for the five mixture distribution models? (0.5 points)

Model A: Unimodal and symmetric with negative kurtosis. Make two means close enough to strengthen the center of the joint distribution.

Model B: Bimodal and symmetric. Same proportion and standard deviation. Separate two means far enough.

Model C: Unimodal and symmetric with positive kurtosis. Make two means the same because the mixture distribution should be unimodal and symmetric. Increase the standard deviation of the first parent distribution to flatten both tails of the joint distribution.

Model D: Unimodal with positive skewness. Make two means close enough. The second distribution has a larger standard deviation.

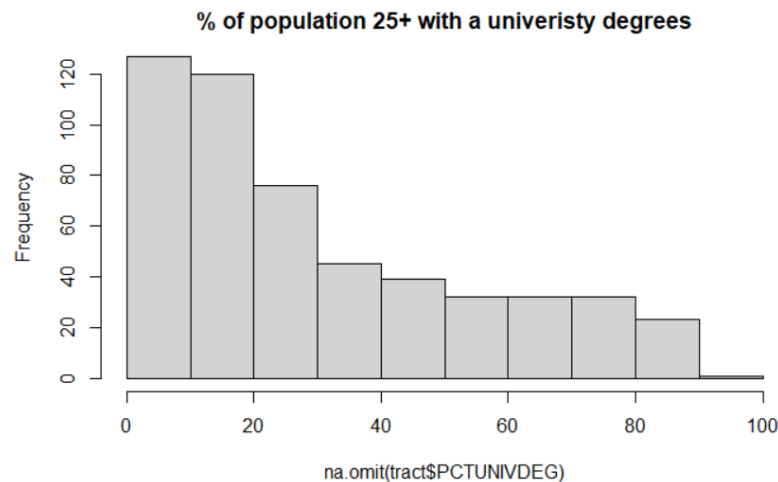
Model E: Unimodal with negative skewness. Make two means close enough. The first distribution has a larger standard deviation.

Task 2: Aggregation – Merger – Weighted Mean (1.5point)

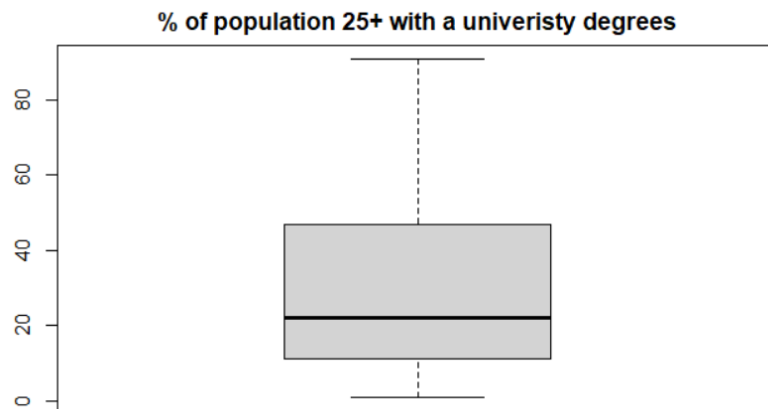
Open the spatial polygon data-frame `tractShp` in the package `TexMix`. Extract the data-frame `tract` `<- as.data.frame(tractShp)` and continue working with `tract`.

[a] Generate a professionally labeled histogram and box-plot of the variable `PCTUNIVDEG` (caution: there are missing values) as well as a side-by-side box-plot for `PCTUNIVDEG` broken down by the factor `CITYPERI`. **Describe** the distribution of the university degree percentage for the sectors of Dallas County and provide *meaningful* summary statistics of the university degree percentages within each sector. (0.3 points)

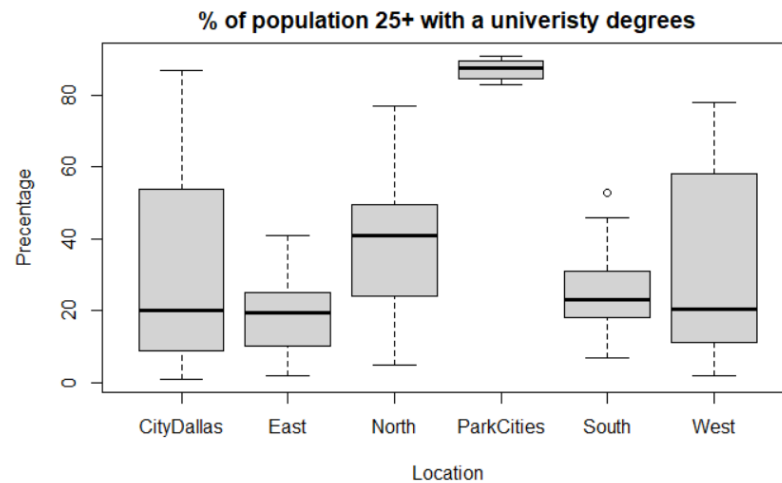
```
hist(na.omit(tract$PCTUNIVDEG), breaks = 12, main="% of population 25+ with a univeristy degrees")
```



```
boxplot(na.omit(tract$PCTUNIVDEG), breaks = 12, main="% of population 25+ with a univeristy degrees")
```



```
boxplot(tract$PCTUNIVDEG~tract$CITYPERI, main="% of population 25+ with a univeristy
degrees",xlab = "Location",ylab="Precentage")
```



```
tapply(tract$PCTUNIVDEG, tract$CITYPERI, summary)
$CityDallas
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
  1.00   9.00   20.00   31.54   54.00   87.00     1
$East
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	10.00	19.50	18.86	25.00	41.00

\$North

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
5.0	24.0	41.0	38.0	49.5	77.0

\$ParkCities

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
83.00	85.25	87.50	87.12	89.25	91.00

\$South

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7	18	23	25	31	53

\$West

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2.00	11.00	20.50	30.36	57.50	78.00	1

Comment: Based on the histogram, boxplot, and summary statistics, we can observe the overall university ratio have a positively skewed distribution. The mean ratio is larger than the median. The boxplot indicates many outliers (Very high ratio) have existed. Among the six neighborhoods, Park City has the highest ratio, whereas, the East and South have the lowest ratios. The City of Dallas the North and West consist of census tracts that vary from low to high ration.

[b] Aggregate the data-frame by the factor **CITYPERI** into a new data-frame with the aggregated statistics **mean**, and **sd** for the variable **PCTUNIVDEG** as well as number of census tracts in each sector (use the **length()** -function). Show your code and the aggregated data-frame with your calculated statistics. Name the variables properly. (0.4 points)

```
DF <- stats::aggregate(PCTUNIVDEG~CITYPERI,data = tract,function(x){ c(mean = mean(x,na.rm = TRUE),sd =sd(x,na.rm = TRUE), Number = length(x)) })
```

	CITYPERI	PCTUNIVDEG.mean	PCTUNIVDEG.sd	PCTUNIVDEG.Number
1	CityDallas	31.537367	26.718363	281.000000
2	East	18.864865	9.743736	74.000000
3	North	38.000000	19.167681	51.000000
4	ParkCities	87.125000	3.090885	8.000000
5	South	25.000000	10.158125	33.000000
6	West	30.362500	24.736712	80.000000

[c] Compare the regular mean of the university degree percentages based on the census tracts with the **weighted mean** based on the aggregated sector means. Use the number of census tracts in each sector as weight. Justify verbally why both means do not differ. (0.4 points)

```
(RegularMean <- mean(tract$PCTUNIVDEG, na.rm = TRUE))
```


```
30.63947
```

```
(WeightMean <- weighted.mean(x=DF$PCTUNIVDEG[,1], w=DF$PCTUNIVDEG[,3]))
```

```
30.63947
```

Both means are identical. For your information this is because:

$$\begin{aligned}
 \bar{x}_w &= \frac{1}{\underbrace{\sum_{j=1}^J n_j}_{=1/n}} \cdot \sum_{j=1}^J n_j \cdot \bar{x}_j \\
 &= \frac{1}{n} \cdot \sum_{j=1}^J n_j \cdot \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j} \\
 &= \frac{1}{n} \cdot \sum_{j=1}^J \sum_{i=1}^{n_j} x_{ij} \\
 \bar{x} &= \frac{1}{n} \cdot \sum_{i=1}^n x_i
 \end{aligned}$$

[d] Merge the aggregated information to your census tract data-frame. Show your code and check that the merger was performed properly by showing the first six records (see the  function `head()`). (0.4 points)

```
DF <- data.frame(DF$PCTUNIVDEG, DF$CITYPERI)
```

```
m <- merge(DF, tract, by.x = "DF.CITYPERI", by.y = "CITYPERI")
```

```
head(m)
```