STAT 210
Applied Statistics and Data Analysis
Problem List 2
(due on week 3)

Fall 2022

Exercise 1

This is an exercise on the use of `plot` and its arguments.

1. Load library MASS and use `str` to explore the structure of the set `crabs`.
2. Divide the plotting window in two rows using the function `par` with argument `mfrow`. In the first row draw a histogram of `BD` for the orange species and in the second row draw a histogram of the same variable for the blue species. Use appropriate titles for the plots and for the axes. Make sure to use the same scale in both axes for both plots. What do you see in these graphs?
3. Plot a graph of `RW` against `BD`. Include as title 'Data on Crabs' by using `main`. Use a solid dot as plotting symbol and use species (`sp`) to determine the color of the points. Add a legend on the upper left corner. Can you conclude anything from this graph?
4. Plot a graph of `RW` against `BD`. Include as title 'Data on Crabs' in the plot by using `main`. Use a solid dot as plotting symbol. Use `sex` to determine the color of the points. Add a legend on the upper left corner. Can you conclude anything from this graph?
5. Divide the plotting window in two columns. Plot on the left a boxplot of `RW` against `sex` and on the right `RW` against `sp`. Comment.
6. Using `plot` draw a scatterplot matrix with the numerical variables in `crabs` (columns 4 to 8). Add color by sex. Comment.

Exercise 2

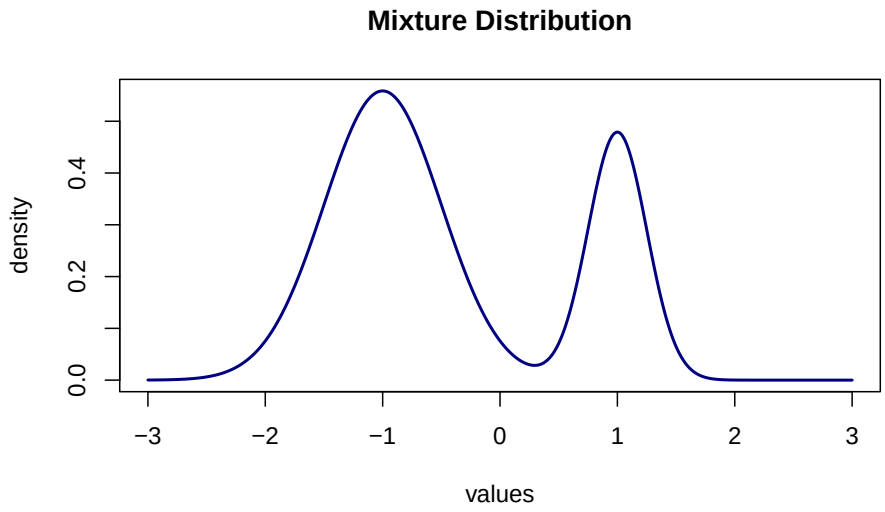
Histograms

For this exercise we are going to use simulated data from a mixture of normal distributions. In this population, 70% of the points come from a normal distribution with mean -1 and standard deviation 0.5, and 30% come from a normal distribution with mean 1 and standard deviation 0.25.

$$0.7 \times N(-1, 0.5^2) + 0.3N(1, 0.25^2)$$

The code below plots the density for this distribution.

```
points.x <- seq(-3,3,length=1000)
points.dens <- 0.7*(dnorm(points.x, mean=-1, sd = 0.5)) +
  0.3*(dnorm(points.x,mean=1, sd = 0.25))
plot(points.x,points.dens,type='l',xlab='values',ylab='density',lwd = 2,
     col = 'navyblue', main = 'Mixture Distribution')
```



The following commands draw a sample of size 500 from this mixture and print the range of values for the simulated data. The sample is stored in the vector `mix.sample`

```
n <- 500
unif.sample <- runif(n) <= 0.7
mix.sample <- unif.sample *rnorm(n,mean=-1, sd = 0.5) +
  (1-unif.sample)*rnorm(n,mean=1, sd = .25)
(rng <- range(mix.sample))
```

```
## [1] -2.476129  1.685508
```

We will use this sample to draw histograms with the function `truehist` in the `MASS` package. Look up the help for `truehist`. It is also a good idea to explore the use of the function `hist` on the base package by repeating this exercise using `hist`.

1. Divide the plotting window into 4 using the function `par` with argument `mfrow`. Select four disjoint subsets of data of length 25 and draw histograms for them. Set the bin width to 0.5 in all plots. Make sure that the scales are the same for all plots. Are these plots similar to the density in the previous slide?

2. Divide the plotting window into 4 using the function `par` with argument `mfrow`. Draw successive histograms of relative frequency for the first 25, 50, 100, and 500 points in `mix.sample`. Set the bin width to 0.5 in all plots. Make sure that the scales are the same for all plots. Are these plots similar to the density in the previous slide?

3. Using again the function `par` with argument `mfrow`, set the graphical window to a single graph. Draw a histogram of relative frequency using all the points in `mix.sample`. Choose the number of bins (`nbins`) using the `Scott` rule.

4. Using the function `lines` with argument `density(sample.mix)`, add an estimate of the density for this sample. Color the line in blue. Add also a graph of the theoretical density in red (look back to the previous page to see how this density was plotted before and make the necessary changes). Comment on what you observe.

Exercise 3

In this exercise we look at quantile plots. In all cases we will consider samples simulated from the normal distribution. We explore the effect of size, mean, and variance, and also use `qqplot` to compare samples.

1. Divide the graphical window into four regions using `par` and `mfrow`. Generate four samples from the standard normal distribution of size 10 and draw normal quantile plots. Add lines with `qqline`.

Comment on what you observe.

2. Repeat for sample sizes 20, 50, and 100. Comment on what you observe.
3. Draw samples of size 50 from normal distributions with means -6, -2, 2, and 6, all with variance 1 and draw the corresponding quantile plots. To be able to compare the four graphs, find a suitable common scale for the axes for all plots. Comment on the similarities and differences between the plots.
4. Draw samples of size 50 from normal distributions with mean 1 and standard deviations 0.5, 2, 4, and 6, and draw the corresponding quantile plots. To be able to compare the four graphs, find a suitable common scale for the axes for all plots. Comment on the similarities and differences between the plots.
5. Draw two samples of size 10 from the standard normal distribution and compare them using `qqplot`. Repeat a total of four times. Plot the four graphs on the same window. Comment on what you see.

Exercise 4

In this exercise we look again at quantile plots, but now we explore the effect of comparing samples from other distributions with the normal.

1. **Distributions with heavy tails**. Draw samples of size 50 from the following distributions: t with 2 degrees of freedom, t with 5 degrees of freedom, t with 10 degrees of freedom, and Cauchy with standard parameters. For each of these samples draw normal quantile plots along with the reference line. Use a single graphic window divided in four. Comment on what you observe.
2. **Distributions bounded below**. Draw samples of size 50 from the following distributions: χ^2 with 2 degrees of freedom, F with 5 and 10 degrees of freedom, lognormal with standard parameters, and pareto with location parameter equal to 1. For the pareto distribution you need to install and load the package `EnvStat` and use the function `rpareto`. For each of these samples draw normal quantile plots along with the reference line. Use a single graphic window divided in four. Comment on what you observe.
3. **Bounded distributions**. Draw samples of size 50 from the following distributions: Uniform in $[-1, 1]$, Beta with both parameters equal to 0.5, Beta with both parameters equal to 2, and Beta with `shape1 =`1, `shape2 =` 3. For each of these samples draw normal quantile plots along with the reference line. Use a single graphic window divided in four. Comment on what you observe.
4. **Mixtures of normal distributions**. Modifying the commands in question 2 of this list, draw samples of size 50 from the following distributions:
   (a) A mixture of normal distributions with 50% of the population coming from a normal distribution with mean -1 and standard deviation 0.5, and 50% coming from a normal with mean 1 and standard deviation 0.5.
   (b) Same as above but change the proportions to 80% and 20%.
   (c) Same as (a) but change the standard deviations to 0.25.
   (d) Same as (b) but change the standard deviations to 0.25.