EPISODE 2 EXERCISE SHEET

Descriptive Statistics

EXERCISE 1 *

A random sample of 50 personal property insurance policies showed the following number of claims over the past 2 years.

Number of Claims	0	1	2	3	4	5	6
Number of Policies	21	13	5	4	2	3	2

(a) Which type of data is Number of Claims?

(b) Determine the percentage of policies (relative frequency) for each claim level.

(c) Compute the five number summary and the mean of Number of Claims, then assess the symmetry of its sample distribution.

(d) In **R Studio** define a vector called `Claims` containing the raw data (data not summarised in a table) of the variable Number of Claims. The vector should have 50 entries in total. Apply the function `Summary` to the vector to double check your numerical answers to (c). Use the vector `Claims` to produce a barplot and a boxplot of Number of Claims. Double check your symmetry assessment from point (c).

EXERCISE 2 *

A publisher receives a copy of a 500-page textbook from a printer. The page proofs are carefully read and the number of errors on each page is recorded, producing the data in the following table:

Number of Errors	0	1	2	3	4	5
Number of Pages	102	138	140	79	33	8

(a) Which type of data is “Number of Errors”?

(b) Determine the relative frequencies of the number of errors.

Number of Errors	0	1	2	3	4	5
Percentages	20.4%	27.6%	28.0%	15.8%	6.6%	1.6%

(c) In **R Studio** define a vector called `Errors` containing the raw data (data not summarised in a table) of the variable Number of Errors. Using the five number summary and appropriate graphical representations, assess the symmetry of its sample distribution. **[Min=0, Q1=1, Median=2, Mean=1.654, Q3=2, Max=5. From the five number summary and from barplot, boxplot, no clear symmetry emerges]**

EXERCISE 3 *

A financial newspaper report the following distribution table which refers to the salaries of a sample of 200 individuals, which data grouped in salary bands.

Salary Band	[0,10)	[10,20)	[20,35]	[35,50]	[50,100]	[100,200]
Number of Individuals	35	50	81	20	10	4
Relative Frequencies	0.175	0.250	0.405	0.100	0.050	0.020
Frequency Densities	0.0175	0.0250	0.027	0.0067	0.001	0.0002
Mid-Points	5	15	27.5	42.5	75	150

(a) Report an approximation of the five numbers summary of the sample distribution.

(b) Report an approximation of the mean of the sample distribution.

(c) State on which assumption the approximations are based on.

EXERCISE 4 *

(a) Load the dataset DW contained in the file DW.RData. How many variables and observations does the dataset contain?

(b) Plot a pie chart of the categorical variable region to visualise the region composition of individuals in the sample. Don't plot a pie chart again in your life.

(c) Plot a bar chart of the discreet variable education. Specify the parameters col="green" and cex.names=0.5 in the function barplot to change the colour of the bars and the x axis labels size.

(d) In order to provide a summary of the variables Wage and Education, complete the table below.

	Wage	Education
Sample Mean	603.73	13.07
Median	522.32	12
Interquartile Range	474.84	3
Quantile of Order 0.20	268.28	12
95th Percentile	1305.79	18
Sample Variance	205705.2	8.41
Sample Standard Deviation	453.5474	2.90
Sample Correlation	0.301644	

(e) With social policy we want to target the poorest 10% of workers. Based on the dataset, what is the salary threshold under which the worker is targeted by the policy? How many workers in the dataset would be targeted in this case? **[182.1]**

(f) Produce the box plot for the variable wage. Assess the symmetry of the sample distribution of the variable wage. **[right/positve skewed]**

(g) Produce a histogram with 20 classes of equal width for the variable wage. Then define the variable logwage as the logarithm **base 10** of the variable wage. Produce a histogram with 20 classes of equal width for the newly defined variable logwage. Which of the two histograms produced is more informative?

(h) The log wage (in base 10) of person A is X, while the log wage (in base 10) of person B is X+1. How can we interpret a unit difference in log wage between these two individuals?

□ Person B earns 1 dollar more than Person A

□ Person B earns 10% more than Person A

☑ Person B earns 10 times what Person A earns

□ Person B earns 10 dollars more than Person A

(i) The log wage (in base e) of Person A is X, while the log wage (in base e) of Person B is X+2. How can we interpret a unit difference in log wage (in base e) between these two individuals?

□ Person B earns $2 \cdot e$ dollars more than Person A

☑ Person B earns e^2 times what Person A earns

□ Person B earns 2^e dollars more than Person A

□ Person B earns 2e% more than Person A

(j) The dummy variable smsa assumes the value “yes” when the corresponding individual resides in a city (standard metropolitan statistical area) and the value “no” otherwise. Consider the R commands:

```
>tapply(DW$wage,DW$smsa,mean)
>boxplot(logwage~DW$smsa)
```

What does the function tapply do?

What can we say about the distribution of wages of individuals who reside in a city with respect to the distribution of wages of individuals who do not?