Telecommunications Engineering
Universidad Carlos III de Madrid
Statistics

Assignment 1: Introduction to MATLAB and Descriptive Statistics

Group	Students	Signatures
	• Alonso Herreros Copete	

IMPORTANT: *The teachers of this course apply a 'zero tolerance' policy regarding academic dishonesty. Students that sign up this document agree to deliver an original work. The breach of this commitment will result in academic punishment.*

Observations:

Solve the exercises in the **Assignment1.pdf** file. **Note:** It is advisable to consult the manual for basic operation of MATLAB / Octave available on the website of the course.

---

*All code for each of the 3 problems is in a separate file called Section<n>.pdf

1. Analysis of a data set

1. Calculate the frequency table of variable Area. The table must include the absolute, relative, cumulative absolute and cumulative relative frequencies. In which of the three types of areas most students are concentrated?

The following is the requested frequency table. There is a screenshot of the generated variable, too.

Area	Abs. frequency	Rel. frequency (%)	Abs. cum. freq.	Rel. cum. freq (%)
1	178	37.238	178	37.2385
2	151	31.59	329	68.8285
3	149	31.1715	478	100

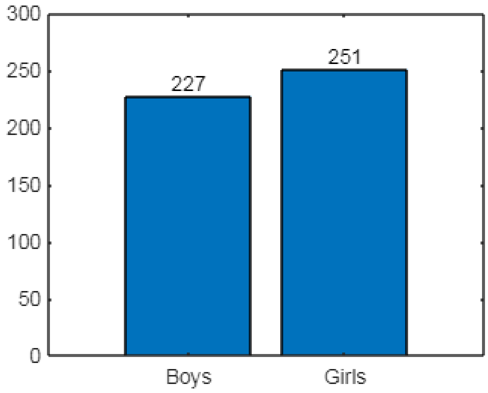
area_ftable = 3×5 table

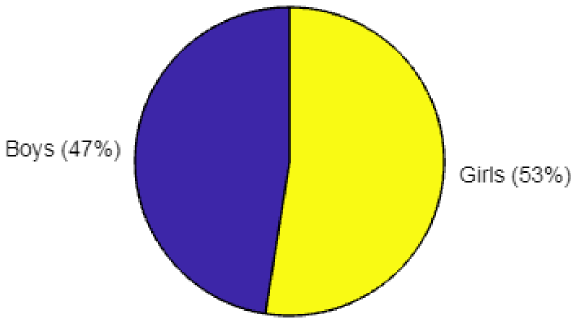
	Area	Abs. freq.	Rel. freq. (%)	Abs. cum. freq.	Rel. cum. freq. (%)
1	1	178	37.2385	178	37.2385
2	2	151	31.5900	329	68.8285
3	3	149	31.1715	478	100

Most students are concentrated in area 1 (urban).

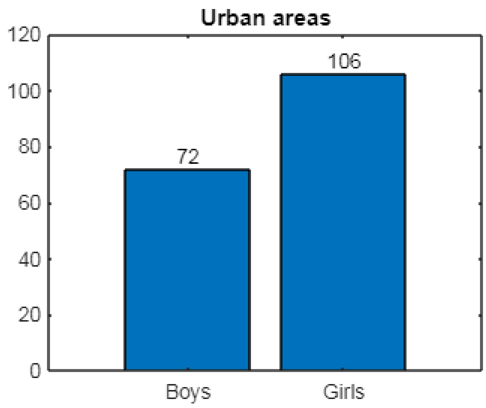
2. What is the proportion of boys and girls? Represent graphically that proportion with a bar and a pie chart. What is the proportion of boys and girls whose schools are established in urban areas?

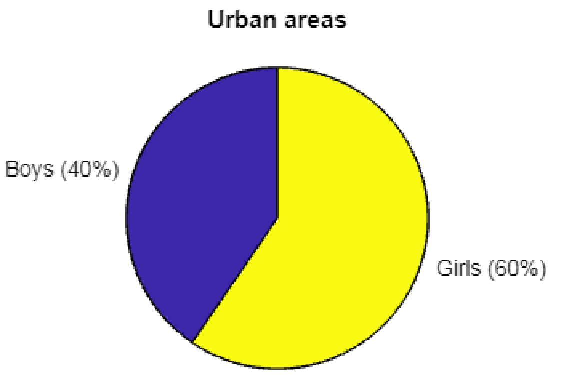
There are 227 boys and 251 girls. That's 47.5% boys and 52.5% girls.





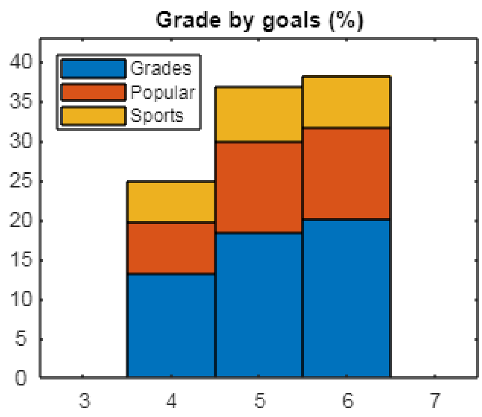
If we focus on urban areas (Area = 1), we can find 72 boys and 106 girls, corresponding to 60% and 40% of students in urban areas, respectively.

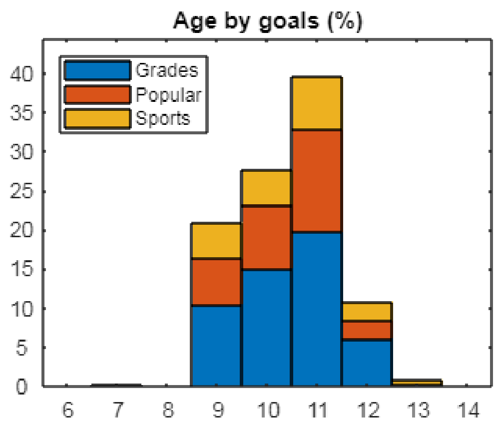




3. Do histograms of the variables Grade and Age by the variable Goals. Calculate the mean and the standard deviation of the variables Grades and Sports by Age groups.

The following are screenshots of the histograms created.





The mean and standard deviation of Grades and Sports by Age groups can be seen in the following table:

Age	Grades mean	Grades std. dev.	Sports mean	Sports std. dev.
7	1.00	0.0000	2.00	0.0000
9	2.38	1.1172	2.34	0.9972
10	2.35	0.9884	2.04	0.9993
11	2.86	1.0747	0.06	0.9519
12	2.90	0.9754	1.81	0.9505
13	2.75	1.5000	2.25	1.2583

This is a screenshot of the variable used to obtain that data

stats_by_age = 6×6 table

		Age	GroupCount	mean_Grades	std_Grades	mean_Sports	std_Sports
1	7	7	1	1	0	2	0
2	9	9	100	2.3800	1.1172	2.3400	0.9972
3	10	10	132	2.3485	0.9884	2.0379	0.9993
4	11	11	189	2.8571	1.0747	2.0582	0.9519
5	12	12	52	2.9038	0.9754	1.8077	0.9505
6	13	13	4	2.7500	1.5000	2.2500	1.2583

4. Analyse the variables Gender and Goals in a double entry table. Calculate the absolute frequency table with its marginal distributions and the relative frequency table with its marginal distributions.

The following are the generated tables, along with screenshots of the variables used. The “Any” rows and columns represent the marginal distributions.

Gender and Goals, absolute frequency

	Get good grades	Be popular	Be good at sports	Any goal
Boys	117	50	60	227
Girls	130	91	30	251
Any gender	247	141	90	478

gender_goals_abs = 3×4 table

		Get good grades	Be popular	Be good at sports	Any goal
1	Boys	117	50	60	227
2	Girls	130	91	30	251
3	Any gender	247	141	90	478

Gender and Goals, relative frequency (%)

	Get good grades	Be popular	Be good at sports	Any goal
Boys	24.48	10.46	12.55	47.49
Girls	27.20	19.04	6.28	52.51
Any gender	51.67	29.50	18.83	100

gender_goals_rel = 3×4 table

		Get good grades	Be popular	Be good at sports	Any goal
1	Boys	24.4770	10.4603	12.5523	47.4895
2	Girls	27.1967	19.0377	6.2762	52.5105
3	Any gender	51.6736	29.4979	18.8285	100

2. Linear Transformations

1. Change of units. Consider the matrix internet in the file internet.mat, and consider the variable MB ("downloaded Mb"). Define a new variable, KB, as the nº of downloaded Kb, recall "1Mb = 1024Kb". The new variable is the result of a linear transformation of the form $y = a + bx$. From this transformation, check with MATLAB/Octave the next theoretical relations:

a) $\bar{y} = a + b\bar{x}$.

b) $y_{med} = a + bx_{med}$, where med is the median.

c) $s_y^2 = b^2 \cdot s_x^2$, where s^2 is the sample quasi-variance.

d) $s_y = |b| \cdot s_x$, where s is the sample quasi-standard deviation.

All relations were checked using MATLAB, and they turned out to be true.

The following is a screenshot of the code used to check these relations, and its output.

```
x = MB;
a = 0;
b = 1024;
y = a + b*x;
```

```
qvar_factor = n/(n-1);
qvar_x = var(x)*qvar_factor
qvar_y = var(y)*qvar_factor
qstd_x = sqrt(qvar_x)
qstd_y = sqrt(qvar_y)
```

```
checks1 = [mean(y) == a+b*mean(x), median(y) == a+b*median(x), ...
    qvar_y == power(b,2)*qvar_x, qstd_y == abs(b)*qstd_x]
```

```
expressions1 = ["\bar{y} = a+b \bar{x}", "y_{med} = a+b x_{med}", ...
    "s_{y}^{2} = b^{2} \cdot s_{x}^{2}", "s_{y} = \left| b \right| \cd
expressions1 = arrayfun(@(expr, val) sprintf("$%s : \\mathrm{%s}$", ex
    expressions1, string(checks1), UniformOutput=false);

clf;
axis off;
for i = 1:numel(expressions1)
    text(0, 1-(i-1)/6, expressions1(i), Interpreter="latex", FontSize=
        Units="normalized", HorizontalAlignment="left", VerticalAlignm
end
```

```
qvar_x = 68.4047
qvar_y = 7.1728e+07
qstd_x = 8.2707
qstd_y = 8.4692e+03
checks1 = 1×4 logical array
    1   1   1   1
```

$\bar{y} = a + b\bar{x}$: true

$y_{med} = a + bx_{med}$: true

$s_y^2 = b^2 \cdot s_x^2$: true

$s_y = |b| \cdot s_x$: true

2. Standardization of variables. Consider the variable MB, and denote it as x. Define a new variable y as the result of the standardization of x. The standardization consists of applying a linear transformation such that subtracts the mean value and divides by its standard deviation. The resulting variable has zero mean, and standard deviation and variance equal to one.

a) Determine the values of a and b of the corresponding linear transformation $y = a + bx$.
b) Obtain the new standardized variable y and check in MATLAB/Octave, the next results: $\bar{y} = 0$, $s_y^2 = 0$, and $s_y = 1$.

a) The normalization can be defined by the expression $y = \frac{x-\bar{x}}{s}$, s being the standard deviation. We can use this expression to identify a and b such that $y = a + bx$ in the following way:

$$\left.\begin{array}{l} y = \frac{x-\bar{x}}{s} = \frac{x}{s} - \frac{\bar{x}}{s} = -\frac{\bar{x}}{s} + \frac{1}{s}x \\ y = a + bx \end{array}\right\} \Rightarrow a = -\frac{\bar{x}}{s}, \; b = \frac{1}{s}$$

b) The expressions were checked using MATLAB, and they turned out to be true, although some adjustments had to be made to account for precision errors.

This is a screenshot of the code used to check the expressions, and its output. The values can be checked directly by looking at them. The last blocks of code perform the numerical comparisons, with a small error margin to account for precision errors, and show the results on screen.

```
x = MB;
y = (x-mean(x))/std(x);
```

```
fprintf("mean = %.4f", mean(y))
fprintf("s^2 = %.4f", var(y))
fprintf("s = %.4f", std(y))
```

```
mean = -0.0000
s^2 = 1.0000
s = 1.0000
```

```
e = 1e-15; % Error to use when comparing floating point numbers
checks2 = [abs(mean(y)-0) < e, abs(var(y)-0) < e, abs(std(y)-0) < e];
```

```
expressions2 = ["\bar{x} = 0", "s^{2} = 1", "s = 1"];
expressions2 = arrayfun(@(expr, val) sprintf("$%s : \\mathrm{%s}$", ex
    expressions2, string(checks2), UniformOutput=false);

clf;
axis off;
for i = 1:numel(expressions2)
    text(0, 1-(i-1)/6, expressions2(i), Interpreter="latex", FontSize=
        Units="normalized", HorizontalAlignment="left", VerticalAlignm
end
```

$\bar{x} = 0 : \mathrm{false}$

$s^2 = 1 : \mathrm{false}$

$s = 1 : \mathrm{false}$

3. Correlation Between Linearly Transformed Variables

Change of units. Consider the matrix internet, and variables MB (x =“downloaded Mb”) and connection (v =“connection time in hours”). From them, create two new variables: y =“nº of downloaded KB” and u =“connection time in seconds”. Note that you are applying a linear transformation of the type: $y = a + bx$ and $u = c + dv$, or simply a change of units: $a = c = 0$ and $b, d > 0$.

Check in MATLAB/Octave, the next result:

$$\rho_{y,u} = \frac{bd}{|b||d|} \cdot \rho_{x,y} = \frac{bd}{bd} \cdot \rho_{x,y} = \rho_{x,y}$$

which indicates that the correlation coefficient between two variables does not change if a change of units is applied.

The expression was checked in MATLAB, and it turned out to be true.

This is a screenshot of the code used to check the expression, as well as its output. Here, the check can be done visually by comparing the values of R_x_v and R_y_u. In the last block of code, the values are actually compared, and the result of the comparison is shown on screen.

```
x = MB;
v = connection;
b = 1024; % MB -> KB conversion factor
d = 3600; % h -> s conversion factor
y = 0 + b*x;
u = 0 + d*v;
```

```
R_x_v = corrcoef(x,v)
R_y_u = corrcoef(y,u)
```

```
check = all([R_y_u == (b*d)/abs(b*d)*R_x_v, ...
    R_y_u == (b*d)/(b*d) * R_x_v, ...
    R_y_u == R_x_v], "all")
```

```
expression = "\rho_{y,u} = \frac{b d}{\left| b \right| \left| d \r
    "\cdot \rho_{x,v} = \frac{b d}{b d} \cdot \rho_{x,v} = \rho_{x
expression = sprintf("$%s : \\mathrm{%s}$", expression, string(exp

clf;
axis off;
text(0, 1, expression, Interpreter="latex", FontSize=14, Units="no
    HorizontalAlignment="left", VerticalAlignment="top")
```

```
connection = 95×1
       1.1000
       3.5000
       3.0000
       2.0000
       1.1000
       5.2000
       2.2000
       1.7000
       1.9000
       2.0000
         ⋮

R_x_v = 2×2
       1.0000    0.7686
       0.7686    1.0000

R_y_u = 2×2
       1.0000    0.7686
       0.7686    1.0000

check = logical
   1
```

$$\rho_{y,u} = \frac{bd}{|b||d|} \cdot \rho_{x,v} = \frac{bd}{bd} \cdot \rho_{x,v} = \rho_{x,v} : \mathrm{true}$$