Definition (Population)

A *population*, denoted by U, is a set of elements of interest.

Definition (Study variable)

A *study variable* is an attribute of interest associated to each element in the population.

ID	**Contact**	$\mathbf{x_1}$	$\mathbf{x_2}$	$\cdots$	$\mathbf{x_J}$
ID_1	$Contact_1$	x_{11}	x_{21}	$\cdots$	x_{J1}
ID_2	$Contact_2$	x_{12}	x_{22}	$\cdots$	x_{J2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ID_N	$Contact_N$	x_{1N}	x_{2N}	$\cdots$	x_{JN}

Table: Array collecting the information in a statistical study

Definition (Parameter)

A *parameter* is a characteristic of interest from the population.

Definition (Mean)

The *average* or *arithmetic mean* or simply, the mean, of a numerical variable x in the population U is defined as

$$\bar{x}_U \equiv \frac{1}{N} \sum_U x_i$$

i.e. the sum of all values of x in the population divided by the size of the population.
In particular, the mean of a dummy variable x is known as a *proportion* and is denoted by P_x.

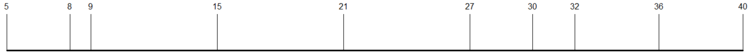
Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

Find the mean of x and the mean of y.


5
8
9
15
21
27
30
32
36
40

Definition (Median)

Let x be a variable that is at least ordinal. The *median* of x, $\check{x}_U$, is the value that divides the population in two halves, in such a way that (at least) half of the x-values are smaller or equal than $\check{x}_U$ and (at least) half of the x-values are larger or equal than $\check{x}_U$.

$$\check{x}_U \equiv \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even} \end{cases}$$

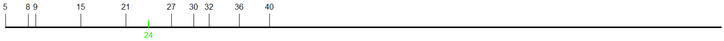
Example

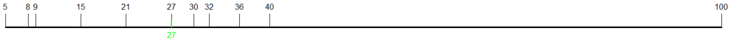
Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

Find the median of x and the median of y.


5
8
9
15
21
27
30
32
36
40
24


5
8
9
15
21
27
30
32
36
40
100
27

Definition (Mode)

The *mode* of a variable x in the population U, $\dot{x}_U$, is defined as the most frequently occurring value.

Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

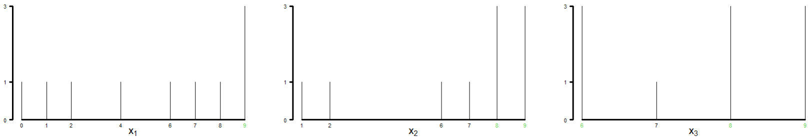
Find the mode of x and the mode of y.

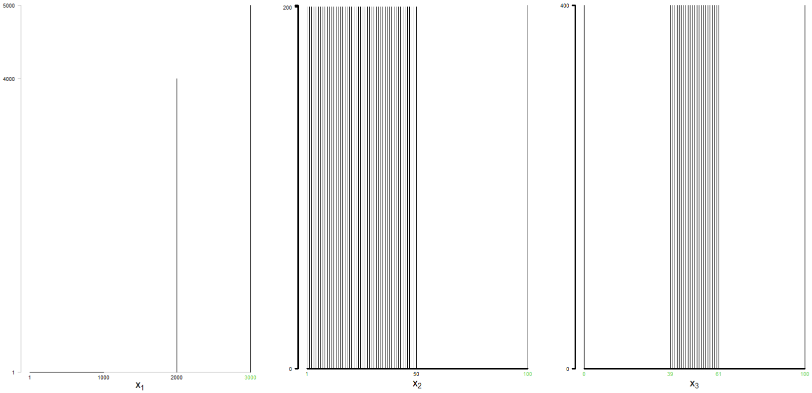
Example

Consider the following variables:

$$\begin{aligned} x_1 &= \{0, 1, 2, 4, 6, 7, 8, 9, 9, 9\} \\ x_2 &= \{1, 2, 6, 7, 8, 8, 8, 9, 9, 9\} \\ x_3 &= \{6, 6, 6, 7, 8, 8, 8, 9, 9, 9\} \end{aligned}$$

Find the mode of x_1, x_2 and x_3.


3
1
0
0 1 2 4 6 7 8 9
x1
3
1
0
1 2 6 7 8 9
x2
3
1
0
6 7 8 9
x3


5000
4000
1
1
1000
2000
3000
x1
200
0
1
50
100
x2
400
0
0
39
61
100
x3

Definition (Percentiles)

Let x be a variable that is at least ordinal. For p in the interval $(0\,,1)$, the 100pth *percentile* of x, $\check{x}_{p,U}$, is the value that divides the population in two parts, in such a way that (at least) $100p\%$ of the x-values are smaller or equal than $\check{x}_{p,U}$ and (at least) $100(1-p)\%$ of the x-values are larger or equal than $\check{x}_{p,U}$.

More formally, let $c = (N-1)p + 1$, a be the integer part of c and b be the decimal part of c, the 100pth percentile is

$$\check{x}_{p,U} \equiv (1-b)x_{(a)} + bx_{(a+1)},$$

where $x_{(a)}$ and $x_{(a+1)}$ are, respectively, the ath and $(a+1)$th observations in the x-ordered population.

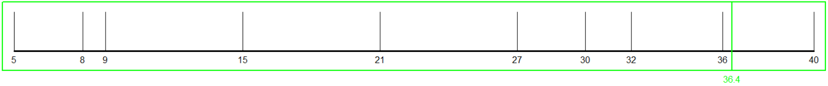
Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

Find the 90th percentile of x and y.


5
8
9
15
21
27
30
32
36
40
36.4

Definition (Quartiles)

The *quartiles* are the percentiles that divide the population into four quarters, so the first quartile is the 25th percentile, $\check{x}_{25,U}$; the second quartile is the 50th percentile, $\check{x}_{50,U}$; and the third quartile is the 75th percentile, $\check{x}_{75,U}$.

Note that the second quartile coincides with the median, $\check{x}_U$.

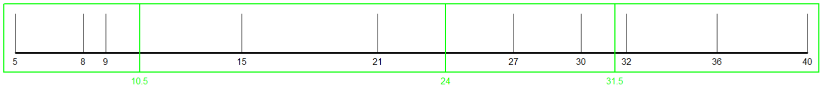
Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

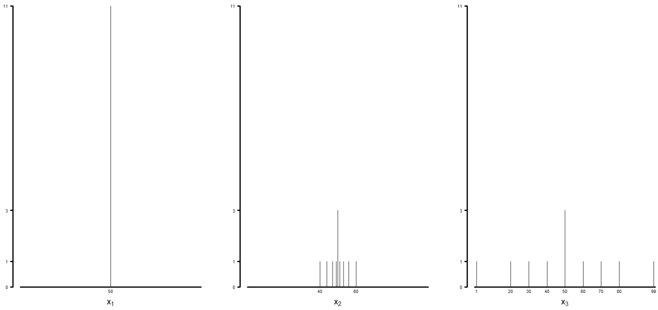
i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

Find the first quartile of x and y.


5
8
9
15
21
27
30
32
36
40
10.5
24
31.5

x_1	x_2	x_3
50	40	01
50	44	20
50	47	30
50	49	40
50	50	50
50	50	50
50	50	50
50	51	60
50	53	70
50	56	80
50	60	99



Definition (Range)

Let x be a variable that is at least ordinal, the *range* is the difference between the maximum and the minimum of x, i.e.

$$\text{range}_{x,U} = x_{(N)} - x_{(1)}.$$

Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam . The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27

Table: Points of ten students in an exam in Statistics

Find the range of x.

Definition (Interquartile range)

Let x be a variable that is at least ordinal, the *interquartile range* of x in the population U, $\mathrm{IQR}_{x,U}$, is the difference between the third and the first quartiles of x, i.e.

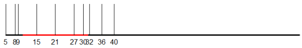
$$\mathrm{IQR}_{x,U} = \breve{x}_{0.75,U} - \breve{x}_{0.25,U}.$$

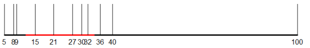
Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam . The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27

Table: Points of ten students in an exam in Statistics





Find the interquartile range of x.

Definition (Variance)

There are two slightly different definitions of the *variance*. The first one (that is more intuitive) is

$$S'^2_{x,U} \equiv \frac{1}{N} \sum_U (x_i - \bar{x}_U)^2 ,$$

which is simply the mean of the square distances from each observation to the mean. The second definition uses $N-1$ instead of N in the denominator, i.e.

$$S^2_{x,U} \equiv \frac{1}{N-1} \sum_U (x_i - \bar{x}_U)^2 .$$

Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table: Points of ten students in an exam in Statistics

Find the variance of x, $S^2_{x,U}$, and the variance of y, $S^2_{y,U}$.

Definition (Standard deviation)

The *standard deviation* is the positive square root of the variance. As we have two different definitions of the variance, we also have two different definitions of the standard deviation:

$$S'_{x,U} \equiv \sqrt{S'^2_{x,U}} = \sqrt{\frac{1}{N}\sum_U (x_i - \bar{x}_U)^2}$$

and

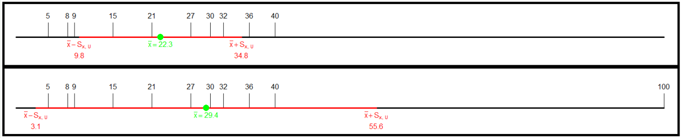
$$S_{x,U} \equiv \sqrt{S^2_{x,U}} = \sqrt{\frac{1}{N-1}\sum_U (x_i - \bar{x}_U)^2}.$$

Example

Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam . The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27

Table: Points of ten students in an exam in Statistics



Find the standard deviation of x, $S_{x,U}$.

Definition (Skewness)

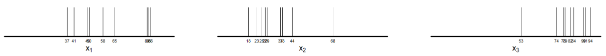
The *skewness* of a numerical variable x in the population U is defined as

$$Sk_{x,U} \equiv \frac{\frac{1}{N}\sum_U (x_i - \bar{x}_U)^3}{S_{x,U}^3}$$

where $\bar{x}_U$ is the mean and $S_{x,U}$ is the standard deviation.

Example

x_1	x_2	x_3
37	18	53
41	23	74
49	26	78
50	28	79
58	29	82
65	37	84
84	38	90
85	44	91
86	68	94



Find the skewness of x_1, x_2 and x_3.

Definition (Outliers)

The ith element of the population U is an outlier with respect to the variable x if its value x_i satisfies

$$x_i < \check{x}_{0.25,U} - 1.5IQR_{x,U} \qquad \text{or} \qquad x_i > \check{x}_{0.75,U} + 1.5IQR_{x,U}$$

where $\check{x}_{25,U}$, $\check{x}_{75,U}$ and $IQR_{x,U}$ are, respectively, the first quartile, the third quartile and the interquartile range.

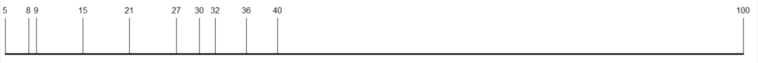
> *One of the authors asked the prominent statistician John W. Tukey [...] why the outlier nomination rule cut at 1.5 IQRs beyond each quartile. He answered that the reason was that 1 IQR would be too small and 2 IQRs would be too large. That works for us.*

Example

Let U be the population of $N = 11$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam . The following table shows the observed values:

i	1	2	3	4	5	6	7	8	9	10	11
x_i	8	15	5	36	40	30	9	21	32	27	100

Table: Points of eleven students in an exam in Statistics



Identify any potential outlier in the population of students.

5	4	7	6	5	4	5	5	5	5
8	6	2	1	4	7	6	6	3	6
8	8	8	5	2	4	4	1	4	6
2	1	6	3	4	7	5	3	9	11
6	4	7	6	4	7	10	7	2	3
7	9	9	4	7	3	7	2	1	3
2	2	4	5	7	4	2	2	3	3
6	2	7	4	10	7	3	4	5	7
4	2	6	7	4	8	6	6	4	9
4	4	2	3	5	6	4			

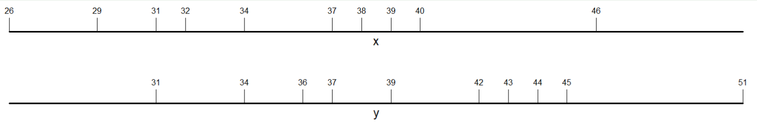
1	1	1	1	2	2	2	2	2	2
2	2	2	2	2	2	3	3	3	3
3	3	3	3	3	3	4	4	4	4
4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5
5	5	5	5	5	5	5	6	6	6
6	6	6	6	6	6	6	6	6	6
6	7	7	7	7	7	7	7	7	7
7	7	7	7	7	8	8	8	8	8
9	9	9	9	10	10	11			

Value x_k	Absolute frequency f_k	Relative frequency w_k	Cumulative absolute F_k	Cumulative relative W_k
1	4	0.0412	4	0.0412
2	12	0.1237	16	0.1649
3	10	0.1031	26	0.2680
4	20	0.2062	46	0.4742
5	11	0.1134	57	0.5876
6	14	0.1443	71	0.7320
7	14	0.1443	85	0.8763
8	5	0.0515	90	0.9278
9	4	0.0412	94	0.9691
10	2	0.0206	96	0.9897
11	1	0.0103	97	1.0000
Total	97	1		

Example

Let us consider the age (in years) of ten individuals as of December 31, 2018 ($= x$), and the age (in years) of the same individuals as of December 31, 2023 ($= y$):

x	26	29	31	32	34	37	38	39	40	46
y	31	34	36	37	39	42	43	44	45	51



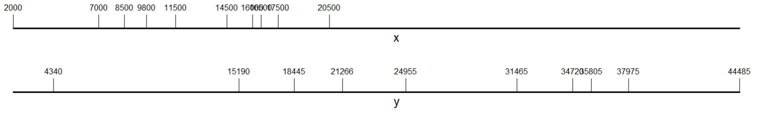
Example

Type	Parameter	x	y
Location	First quartile	31.25	36.25
	Mean	35.2	40.2
	Median	35.5	40.5
	Third quartile	38.75	43.75
Variability	Range	20	20
	IQR	7.5	7.5
	Variance	35.29	35.29
	Standard deviation	5.94	5.94
Shape	Skewness	0.16	0.16

Example

Let us consider the price of ten cell phones in a particular store in Swedish Krona SEK ($= x$) and in Czech Koruna CZK ($= y$):

x	2000	7000	8500	9800	11500	14500	16000	16500	17500	20500
y	4340	15190	18445	21266	24955	31465	34720	35805	37975	44485



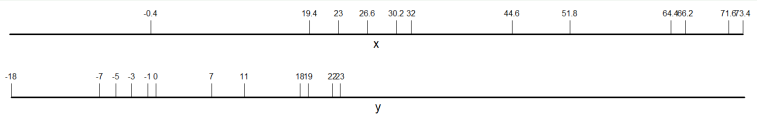
Example

Type	Parameter	x	y
Location	First quartile	8825	19150
	Mean	12380	26860
	Median	13000	28210
	Third quartile	16375	35530
Variability	Range	18500	40150
	IQR	7550	16380
	Variance	31 770 000	149 600 000
	Standard deviation	5636	12230
Shape	Skewness	-0.3094	-0.3094

Example

Let us consider the temperatures in a weather station in Sweden measured at twelve different time points over a year in Fahrenheit ($= x$) and Celsius (y):

x	-0.4	19.4	26.6	32.0	44.6	64.4	73.4	71.6	66.2	51.8	30.2	23.0
y	-18	-7	-3	0	7	18	23	22	19	11	-1	-5



Example

Type	Parameter	x	y
Location	First quartile	25.70	-3.50
	Mean	41.90	5.50
	Median	38.30	3.50
	Third quartile	64.85	18.25
Variability	Range	73.80	41.00
	IQR	39.15	21.75
	Variance	563.5	173.9
	Standard deviation	23.74	13.19
Shape	Skewness	-0.0984	-0.0984

Result

Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U, let a and b be two constants and let

$$y_i = b(x_i + a) \qquad \text{for all } i = 1, 2, \cdots, N.$$

We have

$$\begin{aligned} \bar{y}_U &= b(\bar{x}_U + a) & \dot{y}_U &= b(\dot{x}_U + a) \\ \check{y}_{p,U} &= b(\check{x}_{p,U} + a) & Sk_{y,U} &= Sk_{x,U} \\ \text{range}_{y,U} &= b\,\text{range}_{y,U} & \text{IQR}_{y,U} &= b\,\text{IQR}_{y,U} \\ S_{y,U} &= b\,S_{y,U} & S^2_{y,U} &= b^2\,S^2_{y,U} \end{aligned}$$

Result

Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U and let

$$z_i = \frac{x_i - \bar{x}_U}{S_{x,U}} \qquad \text{for all } i = 1, 2, \cdots, N.$$

then

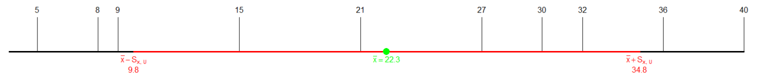
$$\bar{z}_U = 0 \qquad \text{and} \qquad S_{z,U} = 1.$$

The variable z is called the *standard form* of x and the process of substracting the mean to a variable and then dividing by its standard deviation is called *standardization*. The resulting z-values are called *standardized values* or simply the z-scores.

Example

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
z_i	-1.14	-0.58	-1.38	1.09	1.41	0.61	-1.06	-0.10	0.77	0.38

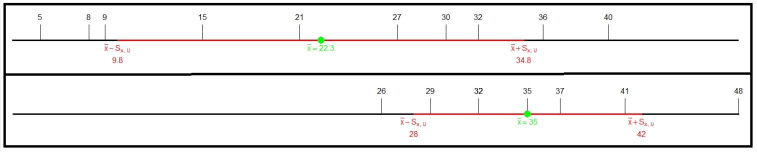
Table: Points of ten students in an exam in Statistics and their z-scores



Example

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
z_i	-1.14	-0.58	-1.38	1.09	1.41	0.61	-1.06	-0.10	0.77	0.38
x_i	48	29	26	32	41	37	35	32		
z_i	1.85	-0.86	-1.28	-0.43	0.86	0.29	0.00	-0.43		

Table: Points of ten students in an exam in Statistics and their z-scores



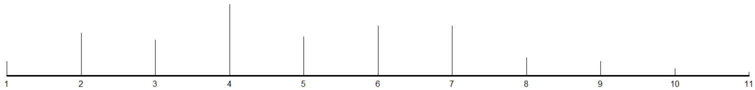


Figure: Dotplot of the number of employees of 97 startups

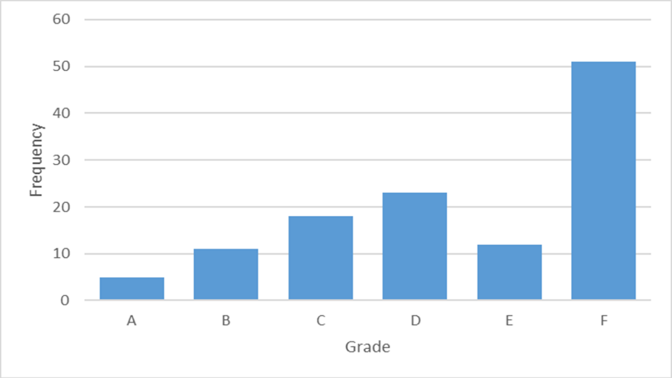


Figure: Bar chart of the grades of 120 students in an exam.

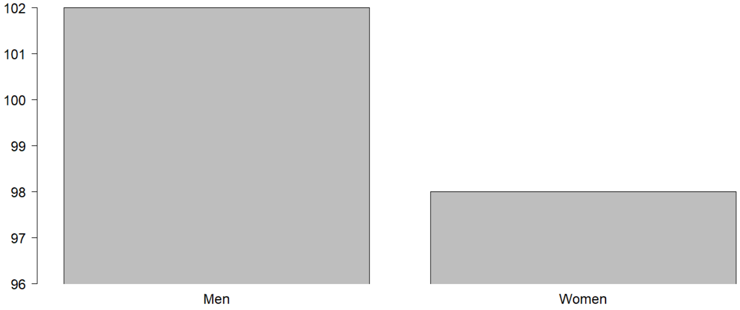


Figure: Bar chart of the the number of men and women in a company.

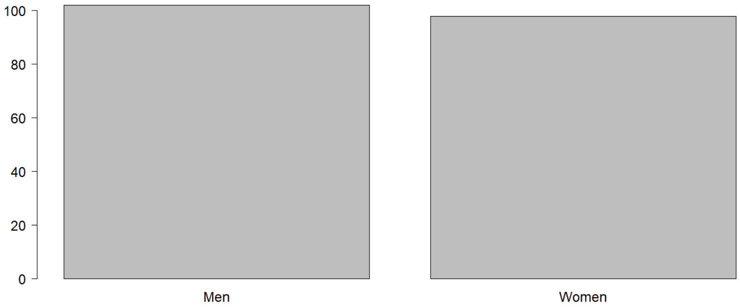


Figure: Bar chart of the the number of men and women in a company.

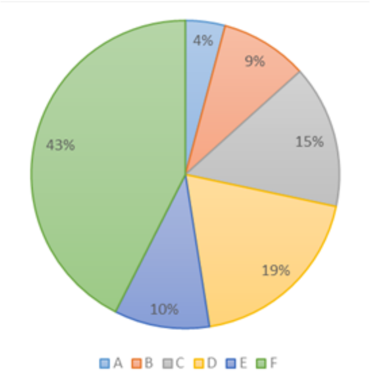


Figure: Pie chart of the grades of 120 students in an exam and an assignment.

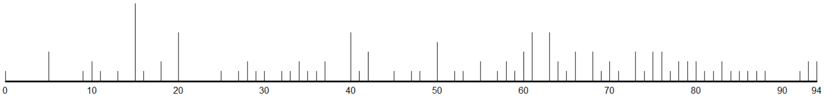


Figure: Dotplot of the number of points of 120 students in an exam.

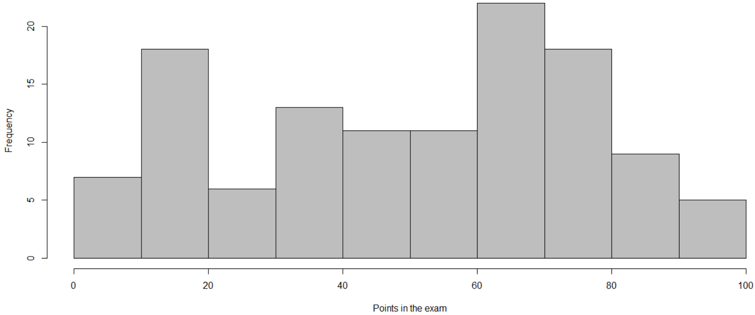


Figure: Histogram of the number of points of 120 students in an exam.

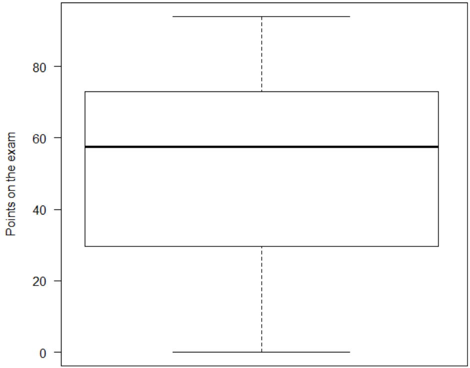


Figure: Box-and-whisker plot of the number of points of 120 students in an exam.