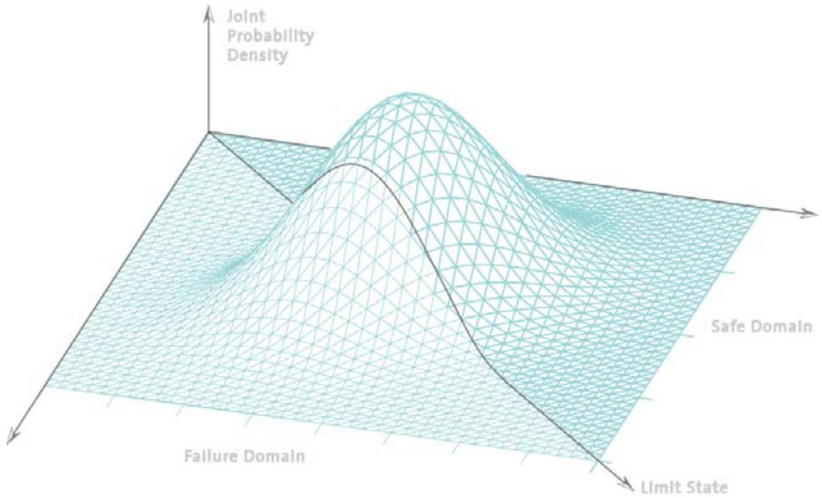


Exercises Tutorial 4

Statistics and Probability Theory

Prof. Dr. Michael Havbro Faber
Swiss Federal Institute of Technology Zurich
ETHZ

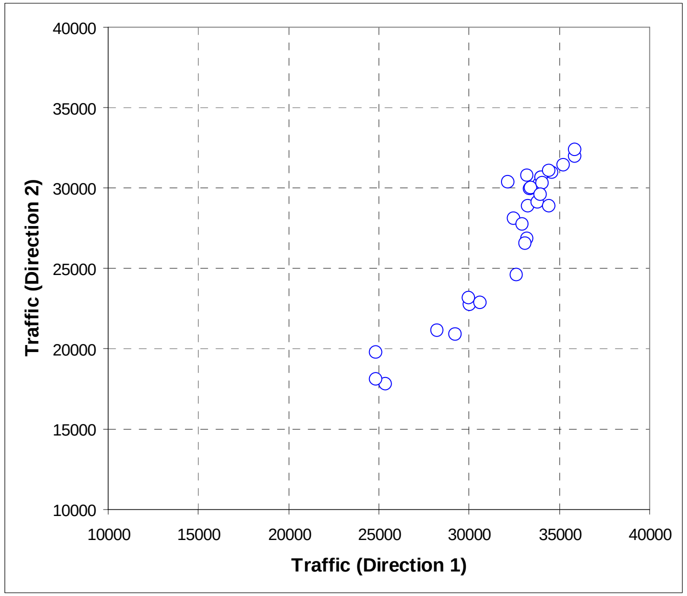
- General
- <u>Correlation plots:</u>
  Plot the UNORDERED observations
- <u>Quantile estimation:</u>
  Order the available data, calculate then the corresponding quantiles

What do we want to know?

Which is the best way to know it? – plot, histogram, statistics etc.

For example,
if you are interested in:
the relation between the traffic of direction 1 and that of direction 2,
but you are not interested in the time element

The graph is correct! Check the unordered pairs of the data!



Correlated!

Quantiles

A quantile is related to a given percentage α, for which α% of all observations in the data set have smaller values.

e.g. the 0.74 quantile of a given data set of observations corresponds to the observation for which 74% of all observations in the data set have smaller values

Direction 1	Direction 2
24846	17805
24862	18123
25365	19735
28252	20903
29224	21145
29976	22762
30035	22828
30613	23141
32158	24609
32472	26525
32618	26846
32962	27746
33091	28117
33197	28858
33198	28877
33245	29080
33380	29586
33406	29965
33788	29994
33888	30263
33937	30313
34007	30366
34013	30629
34076	30680
34425	30788
34455	30958
34576	31074
35237	31405
35843	31994
35852	32384

Q=0.74

74% of the observations
Have a smaller value!

Correct slide in last week's ppt – the unordered data were shown.

Exercise 3.4 (Group Exercise)

Resistivity measurements help to predict the possible corrosion of bridge structures. During a general bridge inspection the data shown in Table 3.4.1 were obtained from resistivity measurements along the two bridge lanes (direction 1 and 2):

a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.

b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.

c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.

Exercise 3.4 (Group Exercise)

Resistivity measurements help to predict the possible corrosion of bridge structures. During a general bridge inspection the data shown in Table 3.4.1 were obtained from resistivity measurements along the two bridge lanes (direction 1 and 2):

a. Draw two box plots for the data provided in Table 3.4.1 (direction 1 and direction 2). Show the main features of the box plots and write their values next to the corresponding points on the diagrams. Plot also the outside values, if any.

b. Tukey box plot is a helpful tool for assessing the symmetry of data sets. Discuss the symmetry/skewness of the resistivity data for both lanes.

c. Choose a suitable number of intervals and plot the histogram for the resistivity data of direction 1.

Steps

1. calculate the median
2. calculate the 75%- and 25%- quantile
3. calculate the adjacent values
4. check for outside values
5. draw the Tukey box plot

Steps
1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 1 (calculate the median)

50%-quantile

$$\nu = nQ_\nu + Q_\nu$$

Median is the value at location:

Steps
1. calculate the median
2. calculate the 75%- and 25%- quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 2 (calculate the 75%- and 25%- quantile)

$$\nu = nQ_v + Q_v$$

Upper quartile (75% quantile):

Lower quartile (25% quantile):

Steps

1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 3 (calculate the adjacent values)

Upper adjacent value: largest observation $\leq$ $(75\%\ quantile) + 1.5r$

In this case, largest value less than

If the largest observation is less than that value,
take the largest observation as the upper adjacent value.

Upper adjacent value =

Steps
1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

Step 3 (calculate the adjacent values)

Lower adjacent value: smallest observation $\geq (25\%\ quantile) - 1.5r$

In this case, lowest value larger than

If the lowest observation is more than that value,
take the lowest observation as the lower adjacent value.

lower adjacent value :

Try the same steps for Direction 2!

Steps
1. calculate the median
2. calculate the 75% and 25% quantile.
3. calculate the adjacent values.
4. check for outside values
5. draw the Tukey box plot

3.4.c

Steps

1. Define number of intervals
2. Count no. of observations within each interval
3. Plot histogram.

Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.
- What do you observe in regard to the traffic flows in directions 1 and 2?
- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.
- What do you observe in regard to the traffic flows in directions 1 and 2?

Direction 2	Direction 1
17805	24846
18123	24862
19735	25365
20903	28252
21145	29224
22762	29976
22828	30035
23141	30613
24609	32158
26525	32472
26846	32618
27746	32962
28117	33091
28858	33197
28877	33198
29080	33245
29586	33380
29965	33406
29994	33788
30263	33888
30313	33937
30366	34007
30629	34013
30680	34076
30788	34425
30958	34455
31074	34576
31405	35237
31994	35843
32384	35852

Steps

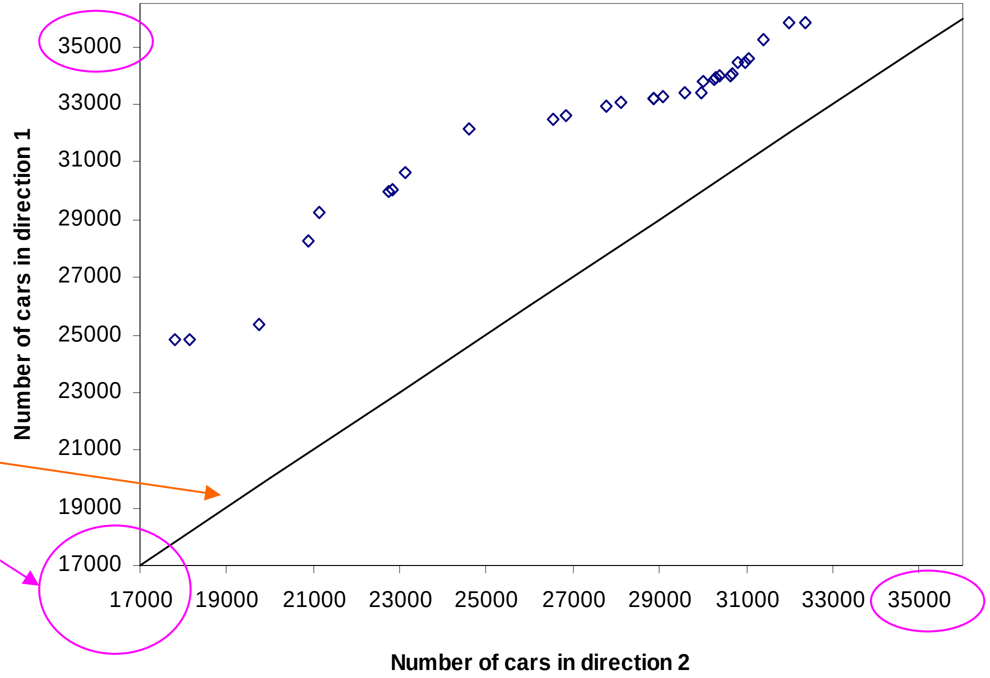
1. sort the data (if not sorted)
2. If $n_x = n_y$ plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line x=y
4. Compare the two data sets

Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.

Steps

1. sort the data (if not sorted)
2. If $n_x = n_y$ plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line x=y (symmetry line)
4. Compare the two data sets

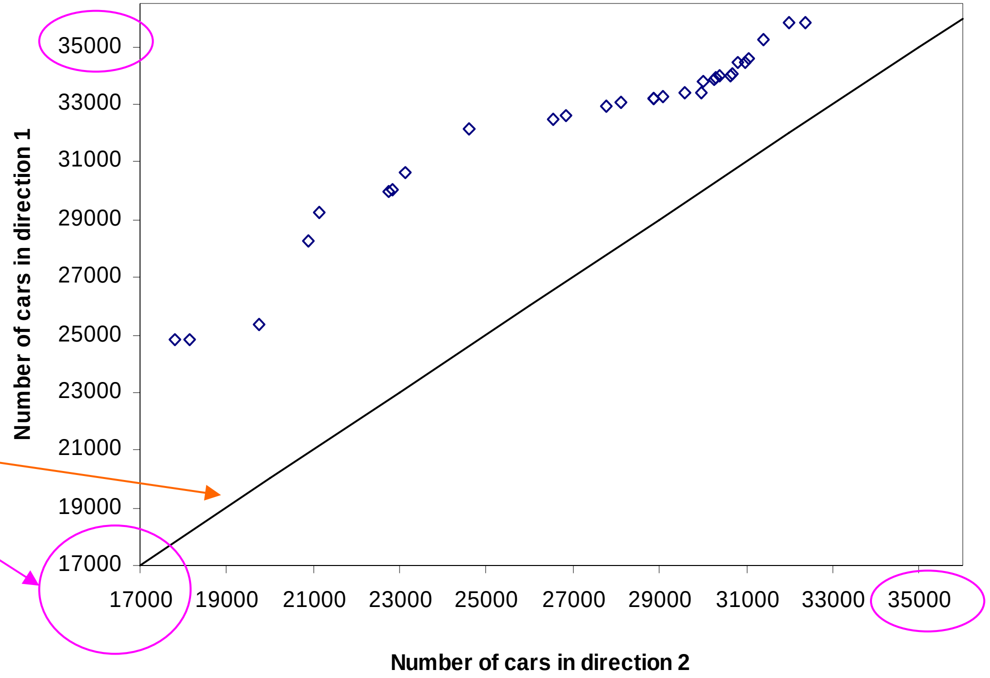


Exercise 3.3

- Make a Q-Q plot (Quantile-Quantile plot) to compare the two data sets of Table 3.1.1.

Steps

1. sort the data (if not sorted)
2. If $n_x = n_y$ plot the data in an x-y system using the same scale and origin for x and y
3. Draw the line x=y (symmetry line)
4. Compare the two data sets



The data lie far from the symmetry line
Concentrated on the side of direction 1- higher traffic flow in direction 1

Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Date	Direction 1	Direction 2
01.04.2001	32618	24609
02.04.2001	33380	29965
03.04.2001	34007	30629
04.04.2001	33888	30263
05.04.2001	35237	31405
06.04.2001	35843	31994
07.04.2001	33197	26846
08.04.2001	30035	22762
09.04.2001	32158	30366
10.04.2001	33406	29994
11.04.2001	34576	30958
12.04.2001	34013	30680
13.04.2001	24846	19735
14.04.2001	28252	21145
15.04.2001	25365	17805
16.04.2001	24862	18123
17.04.2001	32472	28117
18.04.2001	33245	28858
19.04.2001	33788	29080
20.04.2001	34076	30313
21.04.2001	29976	23141
22.04.2001	29224	20903
23.04.2001	32962	27746
24.04.2001	33937	29586
25.04.2001	33198	30788
26.04.2001	34455	31074
27.04.2001	35852	32384
28.04.2001	33091	26525
29.04.2001	30613	22828
30.04.2001	34425	28877

Steps

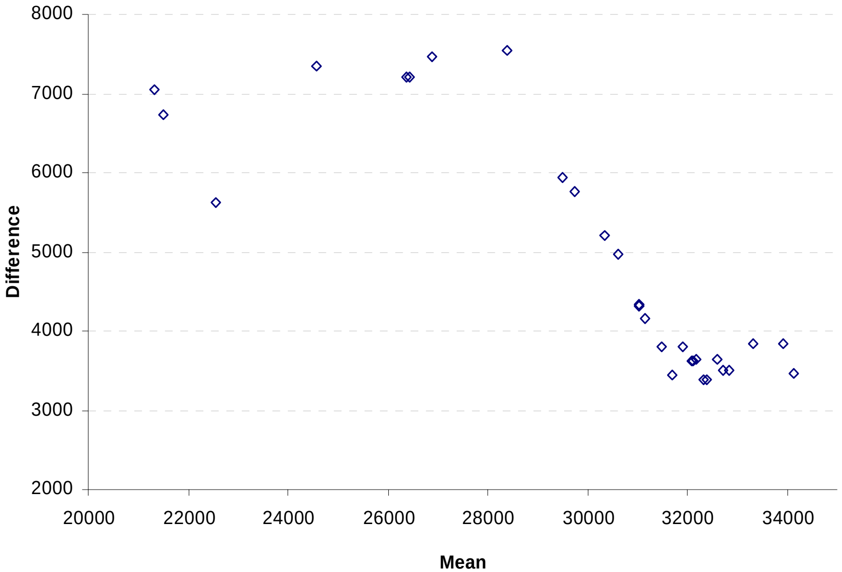
1. sort the data (if not sorted)
2. Calculate $y_i - x_i$ and plot it on the y-axis
3. Calculate $(y_i + x_i)/2$ and plot it on the x-axis
4. Discuss…

Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Steps

1. sort the data (if not sorted)
2. Calculate $y_i - x_i$ and plot it on the y-axis
3. Calculate $(y_i + x_i)/2$ and plot it on the x-axis



x_i	y_i	$y_i - x_i$	$(y_i + x_i)/2$
Direction 2	**Direction 1**	**y_i-x_i**	**$(y_i+x_i)/2$**
17805	24846	7041	21325.5
18123	24862	6739	21492.5
19735	25365	5630	22550.0
20903	28252	7349	24577.5
21145	29224	8079	25184.5
22762	29976	7214	26369.0
22828	30035	7207	26431.5
23141	30613	7472	26877.0
24609	32158	7549	28383.5
26525	32472	5947	29498.5
26846	32618	5772	29732.0
27746	32962	5216	30354.0
28117	33091	4974	30604.0
28858	33197	4339	31027.5
28877	33198	4321	31037.5
29080	33245	4165	31162.5
29586	33380	3794	31483.0
29965	33406	3441	31685.5
29994	33788	3794	31891.0
30263	33888	3625	32075.5
30313	33937	3624	32125.0
30366	34007	3641	32186.5
30629	34013	3384	32321.0
30680	34076	3396	32378.0
30788	34425	3637	32606.5
30958	34455	3497	32706.5
31074	34576	3502	32825.0
31405	35237	3832	33321.0
31994	35843	3849	33918.5
32384	35852	3468	34118.0

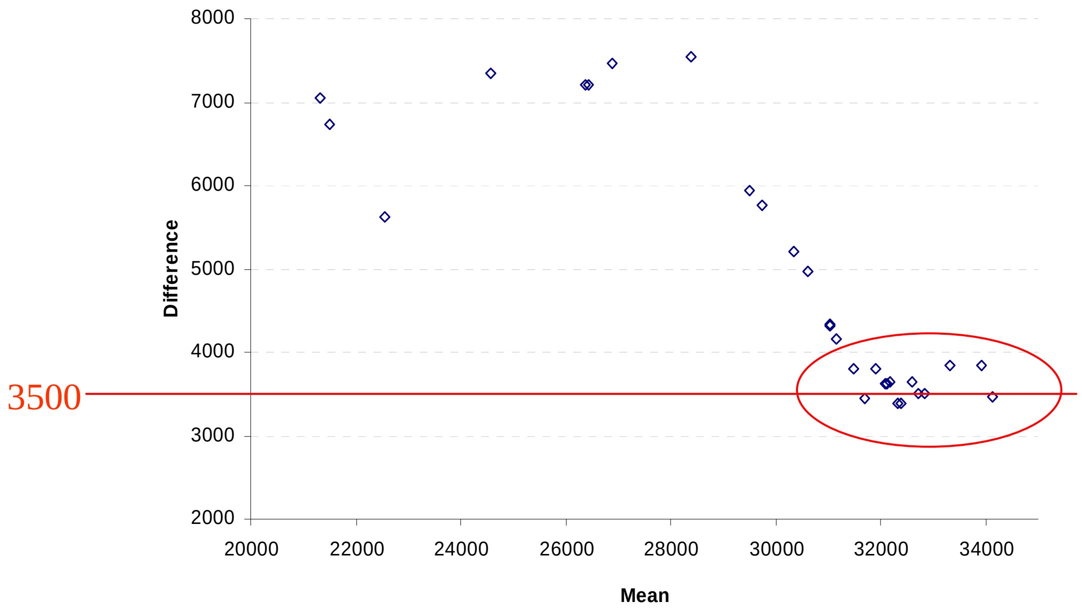
Exercise 3.3

- Provide an approximate value of the difference in the daily traffic flow between the two directions using a Tukey mean-difference plot.

Steps

4. Discuss

for a large part of the data sets the traffic flow in direction 1 is about 3500 cars per day higher than in direction 2



Exercise 4.1

The monthly expense [CHF] for water consumption including sewage fee for a 2-persons household may be considered as a random variable with the following density function:

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

Change to

$$f_X(x) = \begin{cases} c \cdot x \cdot (15 - \frac{x}{4}) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

a. Which value of c should be chosen?

b. Describe the cumulative distribution function $F_X(x)$ of the random variable X.

c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50.00 CHF and 60.00 CHF does not exceed the 90%-quantile of the monthly expense?

d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?

Solution 4.1 a. Which value of c should be chosen?

Probability density function

$$f_X(x) \geq 0$$ ← Non-negative

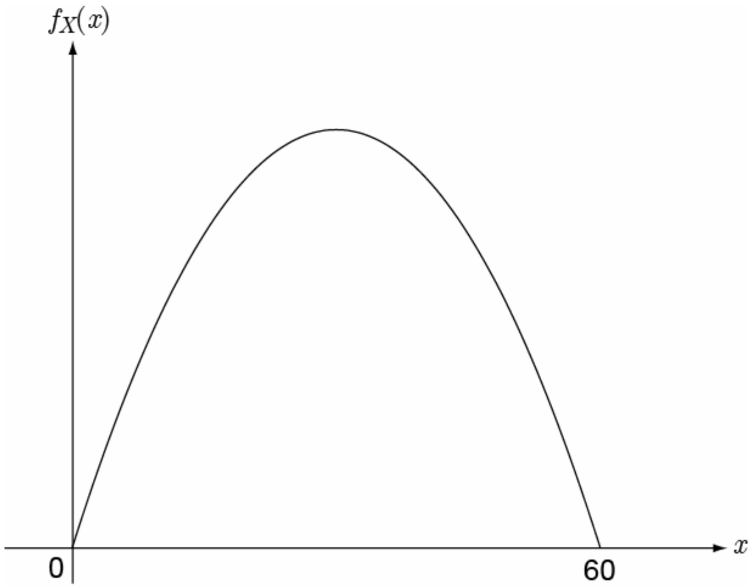
$$\int_\Omega f_X(x)dx = 1$$ ← Area = 1

Solution 4.1 a. Which value of c should be chosen?

Probability density function

$f_X(x) \geq 0$ ← Non-negative

$\int_\Omega f_X(x)dx = 1$ ← Area = 1



$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_0^{60} c \cdot x \cdot (60 - x)\, dx = 1 \Rightarrow c = \frac{1}{36000}$$

Solution 4.1 b. Describe the cumulative distribution function $F_X(x)$ of the random variable *X*.

Cumulative distribution function

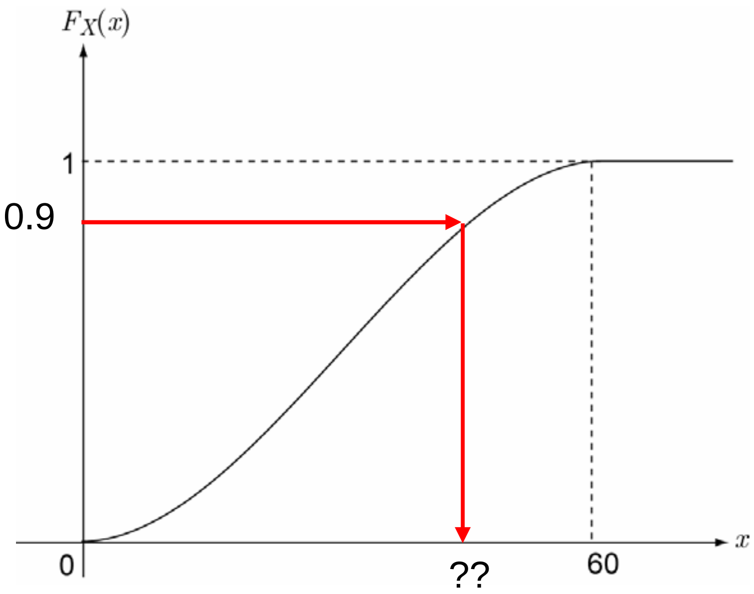
$$F_X(x) = \int_{\Omega} f_X(x) dx$$

$$f_X(x) = \begin{cases} c \cdot x \cdot (60 - x) & \text{for } 0 \leq x \leq 60 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{36000} \cdot \left(\frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \leq x \leq 60 \\ 1 & 60 < x \end{cases}$$

Solution 4.1 c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50 CHF and 60 CHF does not exceed the 90%-quantile of the monthly expense?

First we need to find the value corresponding to the 90% quantile



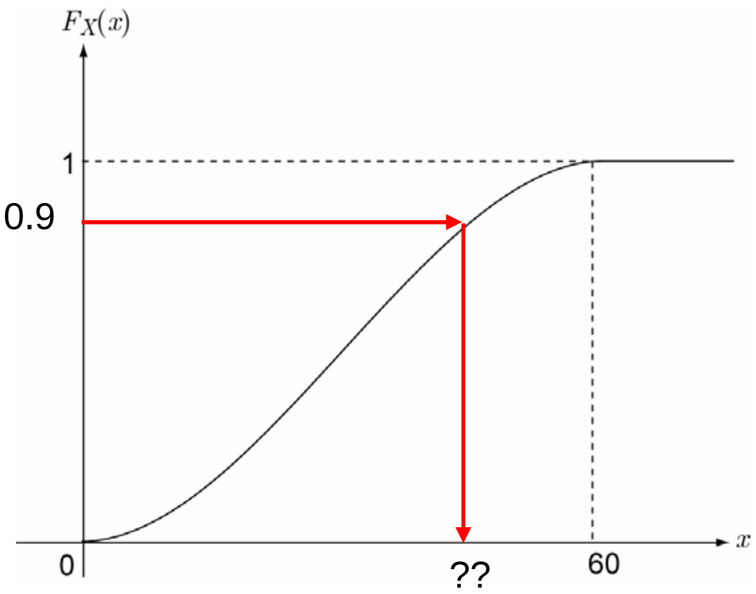
$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{36000} \cdot \left(\frac{60}{2} \cdot x^2 - \frac{1}{3} \cdot x^3 \right) & 0 \le x \le 60 \\ 1 & 60 < x \end{cases}$$

Solution 4.1

c. Which of the following four values, 30.00 CHF, 40.00 CHF, 50 CHF and 60 CHF does not exceed the 90%-quantile of the monthly expense?

First we need to find the value corresponding to the 90% quantile

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \dfrac{1}{36000} \cdot \left(\dfrac{60}{2} \cdot x^2 - \dfrac{1}{3} \cdot x^3 \right) & 0 \le x \le 60 \\ 1 & 60 < x \end{cases}$$

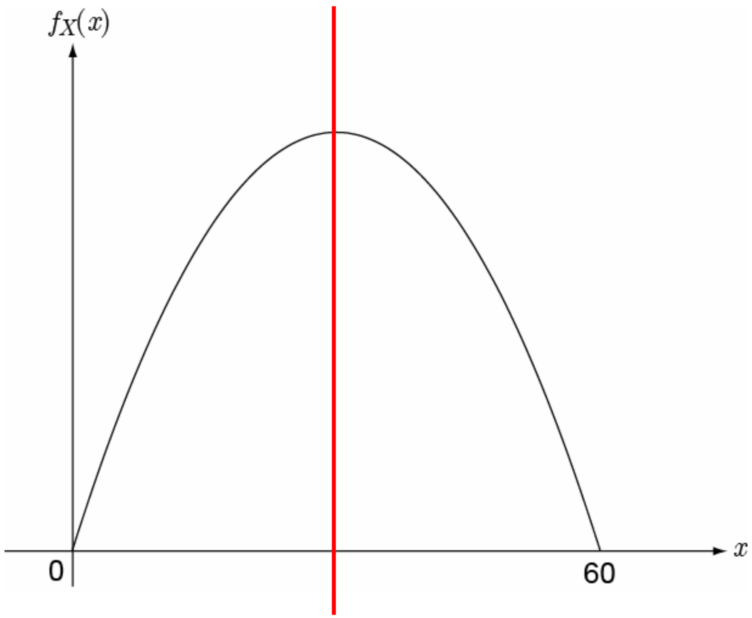


$$P(X \le a) = F_X(x) = 0.9$$

$$P(X \le a) = \frac{1}{36000} \cdot \int_0^{\alpha} x(60\text{-x})\, dx \Rightarrow \alpha = ...$$

Solution 4.1

d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?



Mean = 30

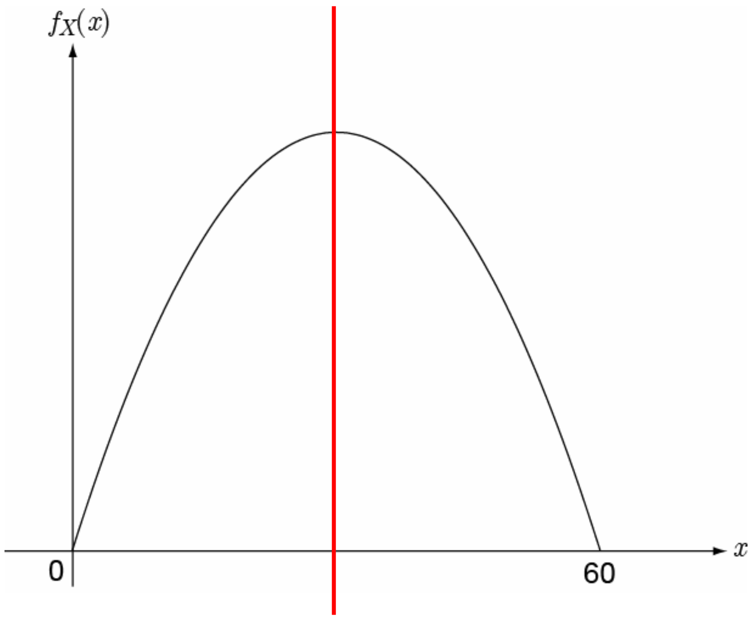
Mean = 30

We can say this directly by looking at the Probability density function. WHY???

Solution 4.1 d. How large is the mean monthly expense for water consumption including sewage fee for a 2-persons household?

Mean---First moment

$$\mu = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$



Mean = 30

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx = \frac{1}{36000} \cdot \int_{0}^{60} x^2 \cdot (60 - x)\, dx$$

Exercise 4.2

The probability function of a basic variable is shown in the following figure.

a. determine analytically the PDF and the CDF.

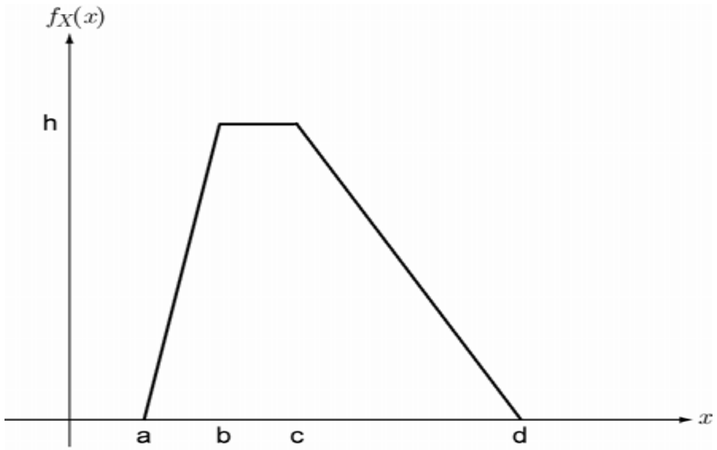
Let a=1, b=2, c=3, d=6. (Change in the exercise)

b. Define the mode and the parameter h.

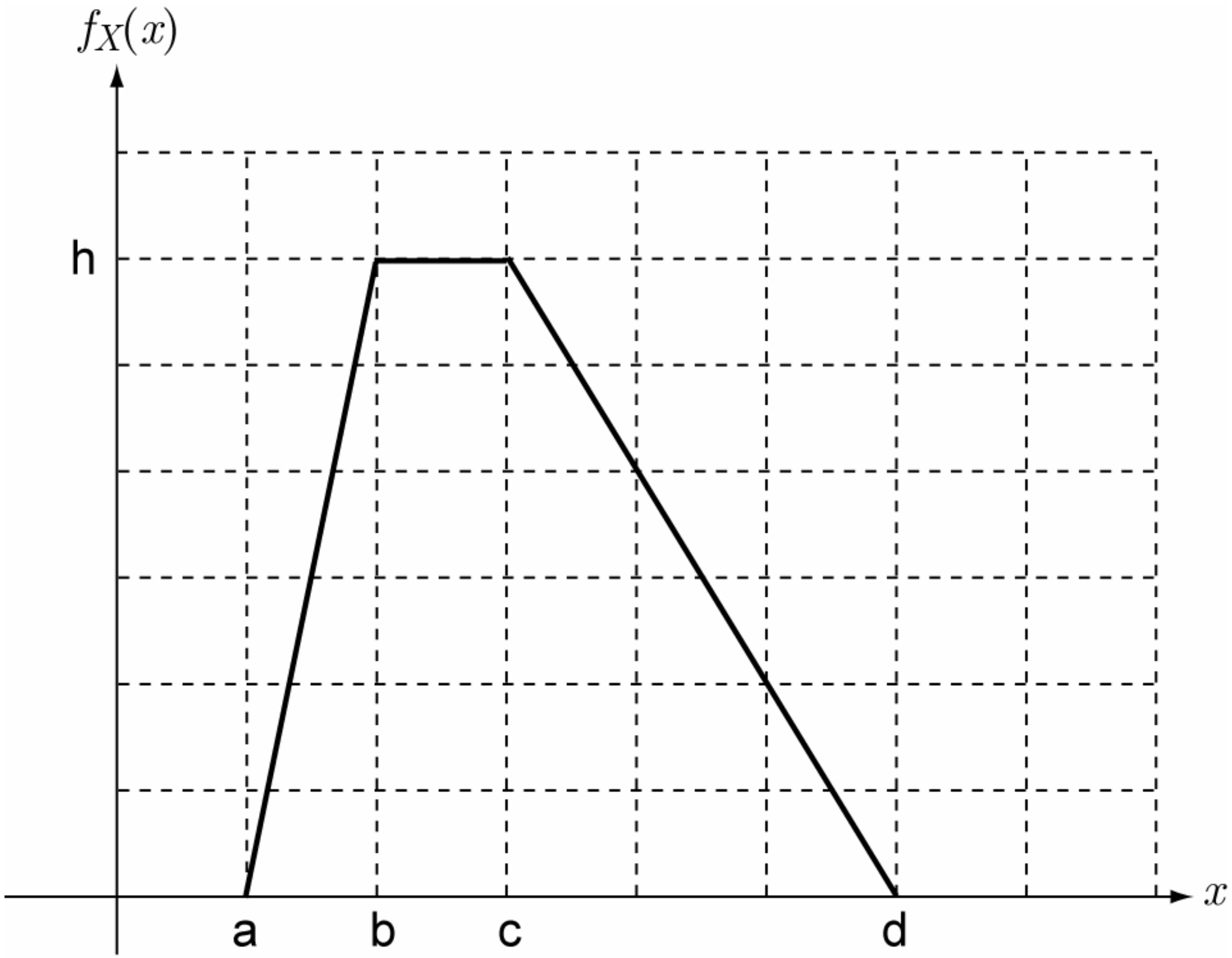
c. Calculate the mean value.

d. Calculate the value of the median.

e. Obtain graphically the median of the pdf. Discuss how the mean value may be evaluated graphically.



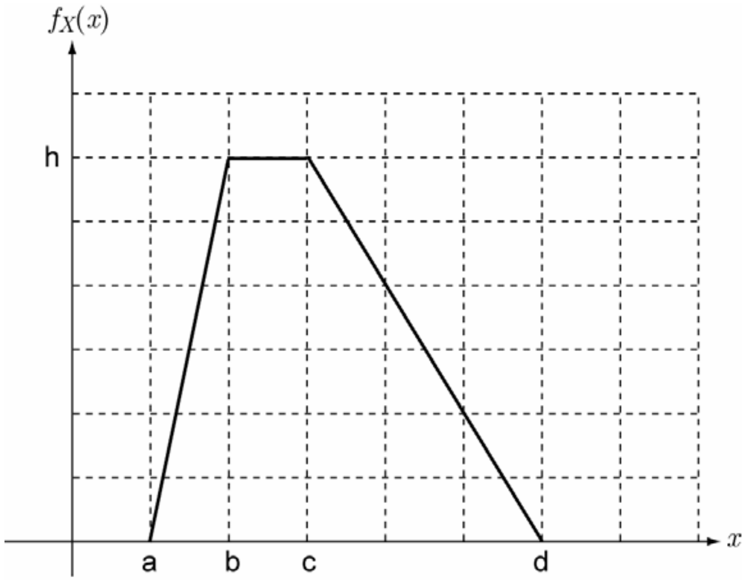
First think along with the definition, then think it again graphically.



Solution 4.2

a. determine analytically the PDF and the CDF.

PDF – Probability Density Function



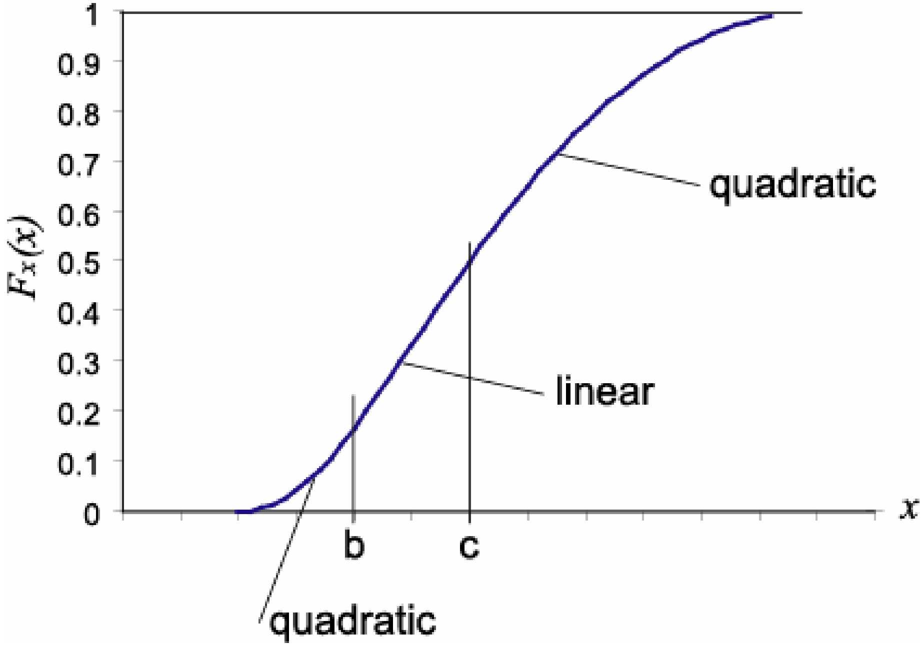
$$f_X(x) = \begin{cases} 0 & x < a \\ h \cdot \dfrac{(x-a)}{(b-a)} & a \le x < b \\ h & b \le x < c \\ h \cdot \dfrac{(x-d)}{(c-d)} & c \le x < d \\ 0 & d \le x \end{cases}$$

Solution 4.2

a. determine analytically the PDF and the CDF.

CDF – Cumulative Distribution Function

$$F_X(x) = \int_\Omega f_X(x)dx$$

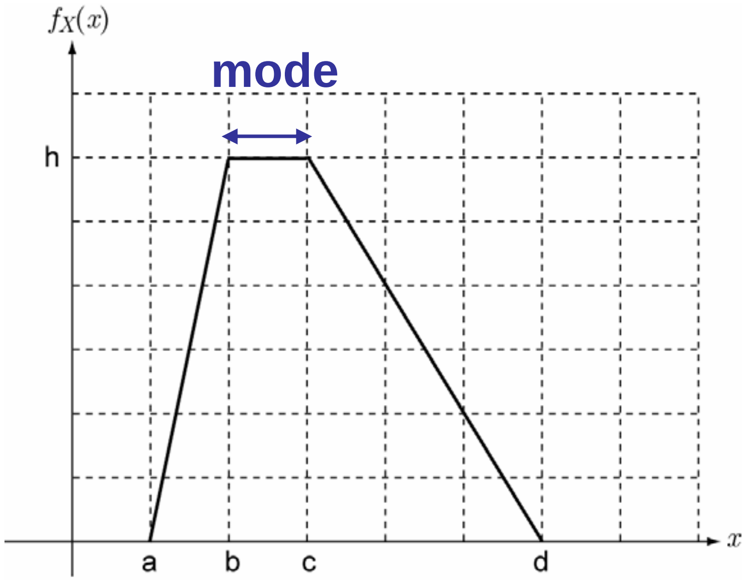


$$F_X(x) = \begin{cases} 0 & x < a \\ h \cdot \frac{(x-a)^2}{2 \cdot (b-a)} + C_1 & a \le x < b \\ h \cdot x + C_2 & b \le x < c \\ h \cdot \frac{(x-d)^2}{2 \cdot (c-d)} + C_3 & c \le x < d \\ C_4 & d \le x \end{cases}$$

The four constants can be calculated by using the boundary conditions

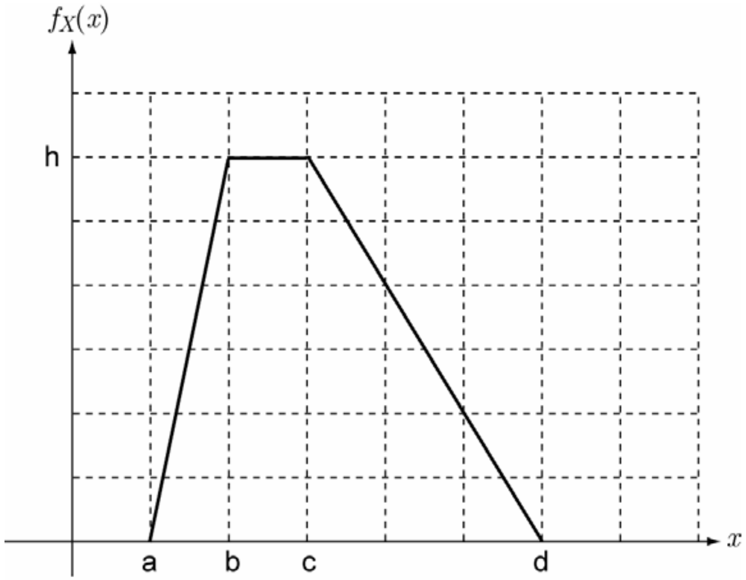
Solution 4.2

b. define the mode and the parameter h. (a=1, b=2, c=3, d=6)



Solution 4.2

b. define the mode and the parameter *h*. (a=1, b=2, c=3, d=6)



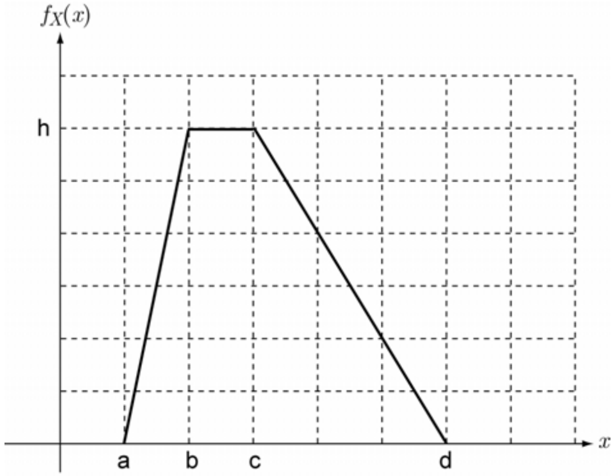
$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

Area under the density function

$$\frac{(d-a)+(c-b)}{2} \cdot h = 1 \Rightarrow ......h = ...$$

Solution 4.2

c. Calculate the value of the mean (a=1, b=2, c=3, d=6)



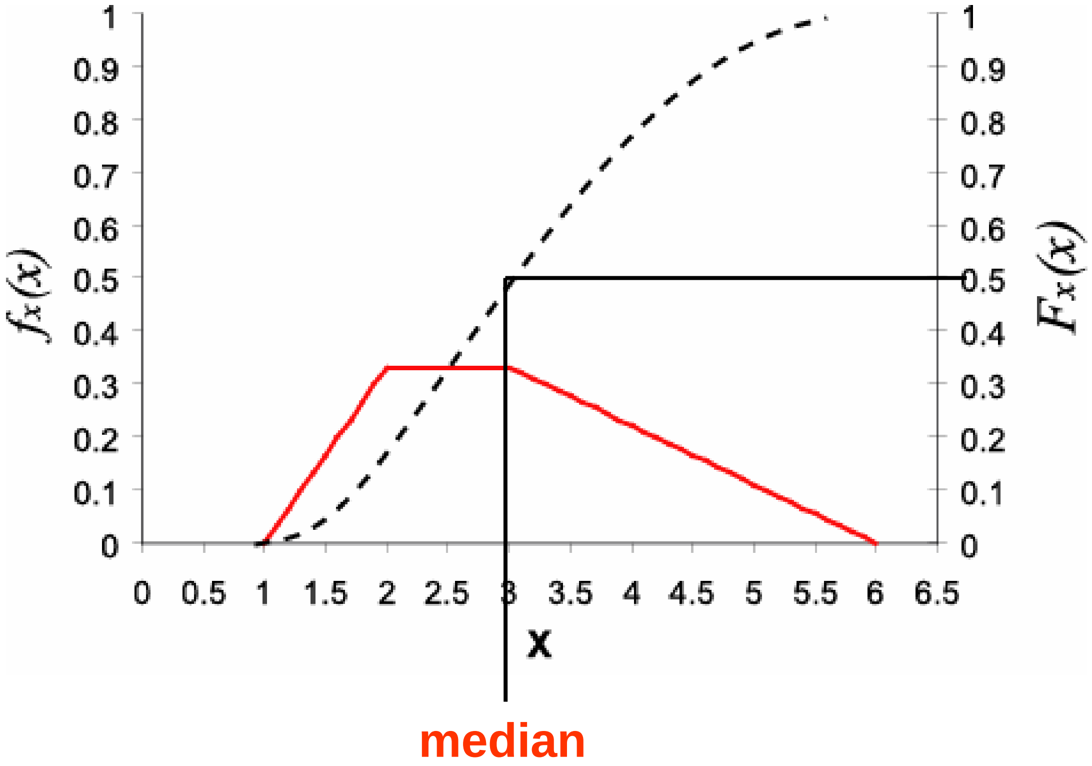
$$f_X(x)=\begin{cases} 0 & x<a \\ h\cdot\frac{(x-a)}{(b-a)} & a\le x<b \\ h & b\le x<c \\ h\cdot\frac{(x-d)}{(c-d)} & c\le x<d \\ 0 & d\le x \end{cases} \Rightarrow f_X(x)=\begin{cases} 0 & x<1 \\ \frac{(x-1)}{3} & 1\le x<2 \\ \frac{1}{3} & 2\le x<3 \\ -\frac{(x-6)}{9} & 3\le x<5 \\ 0 & 5\le x \end{cases}$$

$$\mu_x = E[x] = \int_{-\infty}^{\infty} x\cdot f_x(x)\cdot dx = \int_1^2 \frac{x\cdot(x-1)}{3}dx + \int_2^3 \frac{x}{3}\cdot dx + \int_3^6 \frac{-x\cdot(x-6)}{9}dx = ...$$

Solution 4.2

d. Calculate the value of the median.

Graphically from the CDF



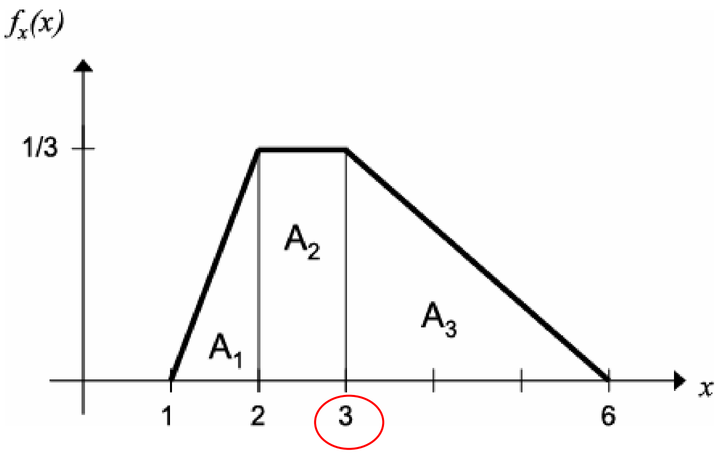
Analytically

$$P(X \leq x) = \int_1^x f_X(x)\ dx = 0.5$$

Solution 4.2

e. Obtain graphically the median of the pdf. Discuss how the mean value may be evaluated graphically.

Graphically from the PDF



$$A_1 = (2-1)\cdot\frac{1}{3}\cdot\frac{1}{2} = \frac{1}{6}$$

$$A_2 = (3-2)\cdot\frac{1}{3} = \frac{1}{3}$$

$A_1 + A_2$: 0.5

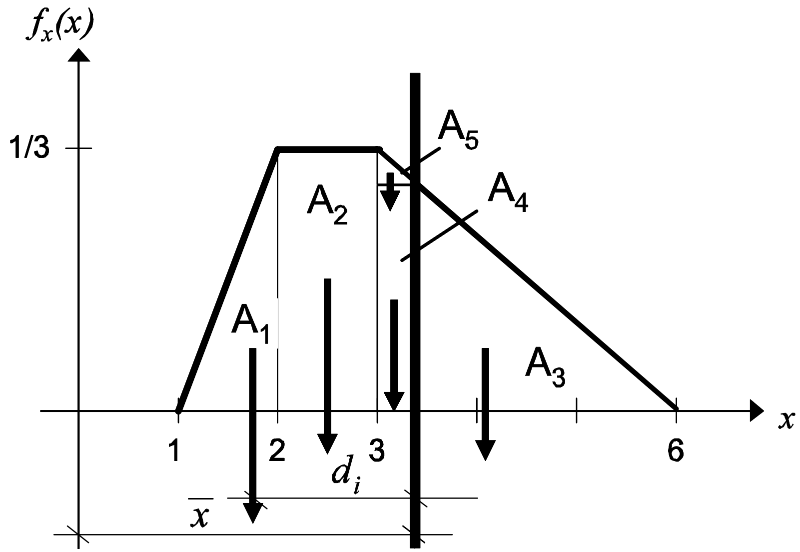
$$A_3 = (6-3)\cdot\frac{1}{3}\cdot\frac{1}{2} = \frac{1}{2}$$

Median: location at which the area under the density function is equal to 0.5

Solution 4.2

e. Obtain graphically the median of the pdf. Discuss how the mean value may be evaluated graphically.

Graphically from the PDF



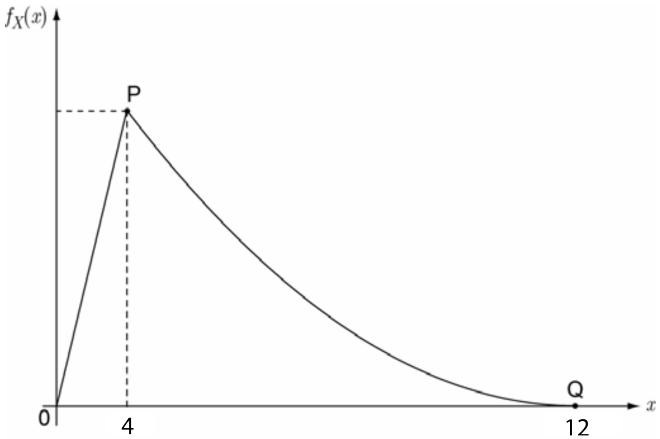
$$\sum_{i=1}^{5} A_i \cdot d_i = 0$$

1. **Estimate moments for each shape**
2. **Take equilibrium around the hypothesized location of the center of gravity**

Mean: center of gravity of the shape of the probability density function.

Exercise 4.3 (Group exercise- to be presented on 19.04.07)

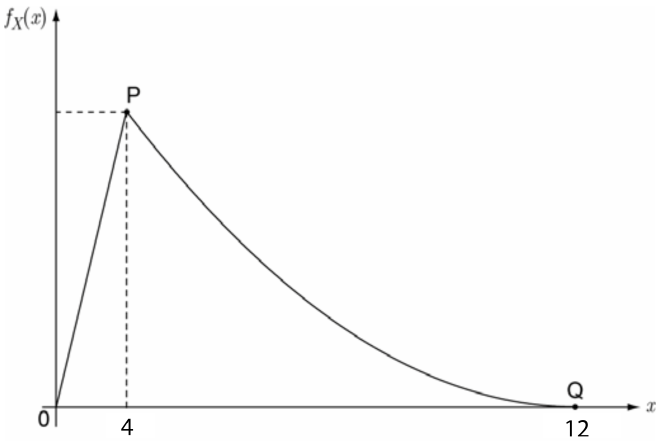
The probability density function of a random variable X is shown in Figure 4.3.1. In the interval [0, 4] the function is linear and in the interval [4, 12] the function is parabolic which is tangent to x-axis at point Q.



a. Determine the coordinate of point P(x,y) and then describe the probability density function.
b. Describe and draw the cumulative distribution function of X with some characteristic numbers in the figure.
c. Calculate the mean value of X.
d. Calculate $P[X>4]$.

Exercise 4.3 (Group exercise- to be presented on 19.04.07)

a. Determine the coordinate of point P(x,y) and describe the probability density function.

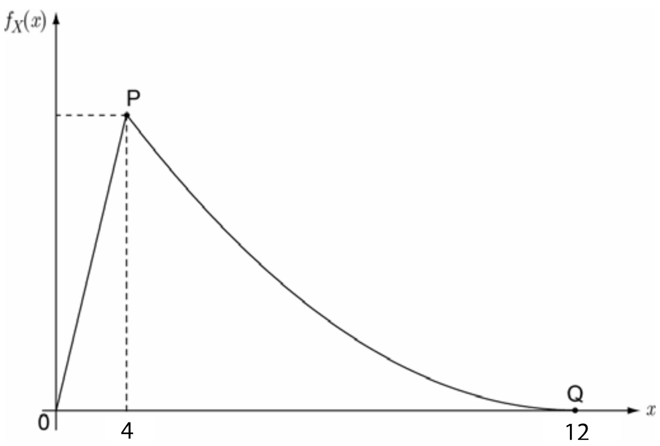


Steps:

- **Define the pdf in the interval [0,12]**
- **Find coordinates of P by remembering that the area under the density function is equal to 1!**

Exercise 4.3 (Group exercise- to be presented on 19.04.07)

b. Describe and draw the cumulative distribution function of X with some characteristic numbers in the figure.



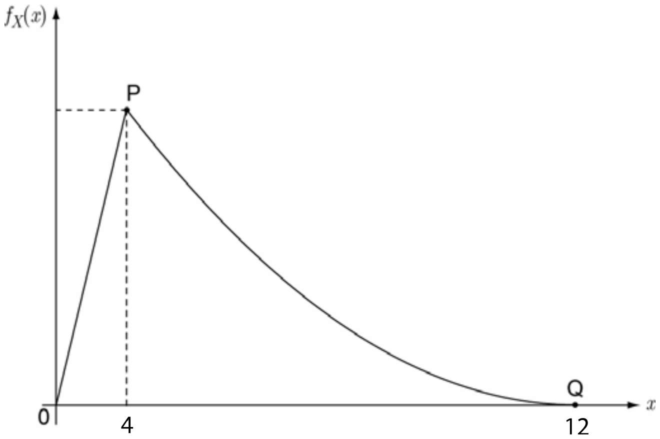
Steps:

1. $\int_{\Omega} f_X(x)dx = 1$

2. Draw…!

Exercise 4.3 (Group exercise- to be presented on 19.04.07)

c. Calculate the mean value

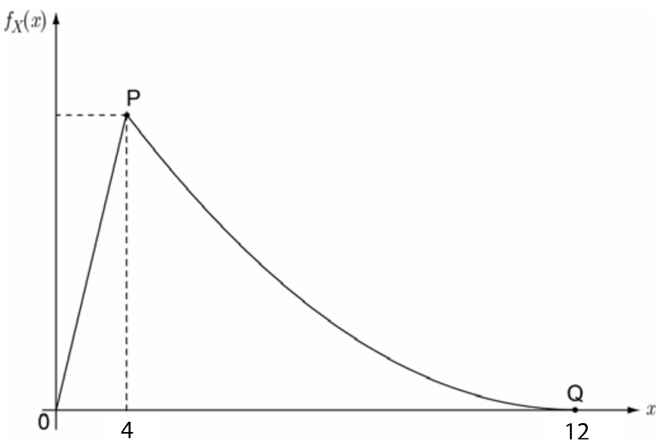


Steps (Remember Exercise 4.2):

1. $\mu_x = E\left[x\right]$

Exercise 4.3 (Group exercise- to be presented on 19.04.07)

d. Calculate $P[X>4]$.



Steps (Remember Exercise 4.2):

Exceedance probability $P[X>\alpha]$ is $1-P[X\leq\alpha]$

1. $P[X>4]=1-P[X\leq 4]$

How can this be expressed???