Histograms and Frequency Polygons

When data is grouped into classes, the best way to visualize the frequency distribution is by constructing a **histogram** (hist/histogram). A histogram is a type of bar graph, where classes are represented by rectangles whose bases are the class lengths and whose heights are chosen so that the areas of the rectangles are proportional to the class frequencies. If the classes have all the same length, then the heights will be proportional to the class frequencies.

A histogram shows the shape of a pdf (probability distribution/density function) or pmf (probability mass function) of data, checks for homogeneity, and suggests possible outliers.

A **frequency histogram** consists of columns, one for each class (bin), whose height is determined by the number of observations in the bin.

A **relative frequency histogram** has the same shape but a different vertical scale. Its column heights represent the proportion of all data that appeared in each bin.

If relative frequencies are considered (so the proportionality factor is N, the total number of observations), then the total areas of all rectangles will be equal to 1. For a large volume of data grouped into a reasonably large number of classes, the histogram gives a rough approximation of the density function (pdf) of the population from which the sample data was drawn.

An alternative in that sense (the sense of roughly approximating the shape of the density function) to histograms are **frequency polygons**, obtained by joining the points with coordinates (x_i, f_i), $i = \overline{1, n}$ ($x-$coordinates are the class marks and $y-$coordinates are the class frequencies).

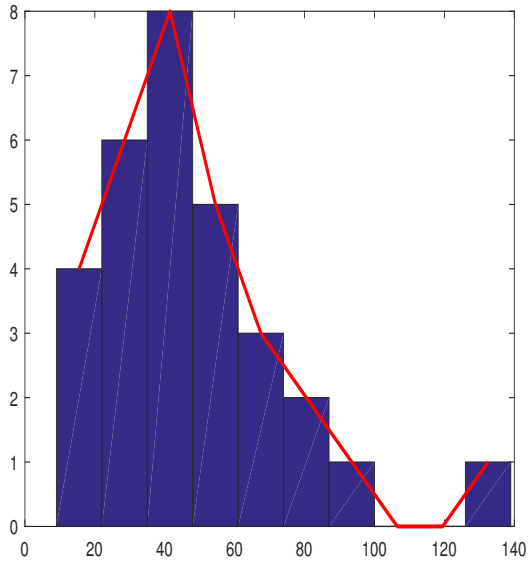
Example 4.4. Recall Example 4.3 from last time, about the CPU times (in seconds) for $N = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

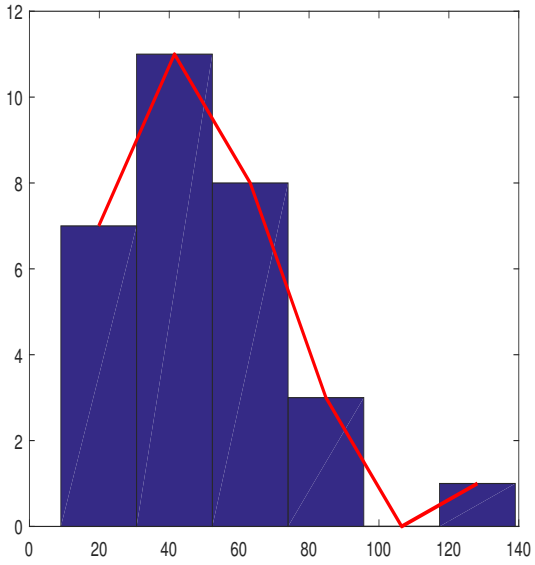
We constructed the grouped frequency distribution tables for these data for $n = 10$ and for $n = 6$ classes. Figure 1 shows the corresponding histogram and frequency polygon for grouped data ((a) and (b)). Also in Fig. 1, we show histograms for $n = 4$ and $n = 12$ bins, respectively. It is obvious that $n = 4$ is too small and $n = 12$ is too large for the number of bins. The values $n = 6$ and $n = 10$ seem to be the best (in terms of the information they provide), especially $n = 10$.

For 10 classes, let us take a closer look, see Figure 2. What information can we draw from these histograms?

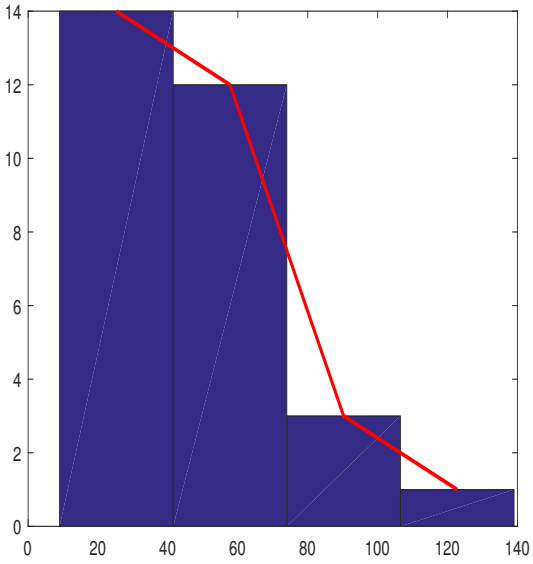
- the continuous distribution (continuous because time varies continuously) of the CPU times



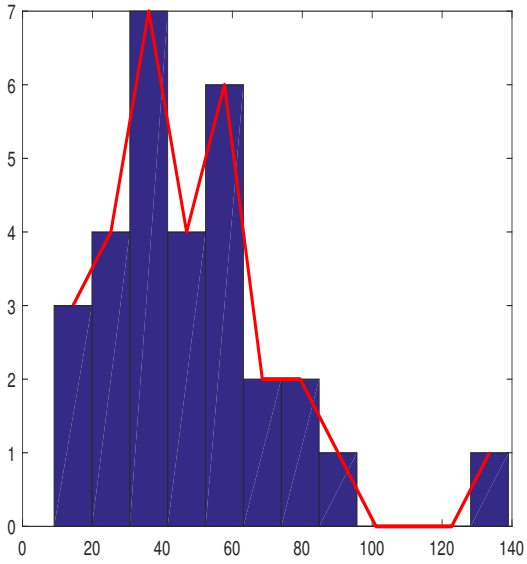
(a) $n = 10$ bins



(b) $n = 6$ bins



(c) $n = 4$ bins



(d) $n = 12$ bins

Fig. 1: Histograms and Frequency Polygons, Example 4.4

is not symmetric, it is skewed to the right, as we see 5 columns to the right of the highest column and only 2 columns to the left;

- the value 139 stands alone suggesting that it is in fact an outlier;
- a Gamma family of distributions seems appropriate for CPU times, see the dashed curve in Figure 2;
- there is no indication of heterogeneity; all data points except $x = 139$ form a rather homogeneous group that fits the sketched Gamma curve.

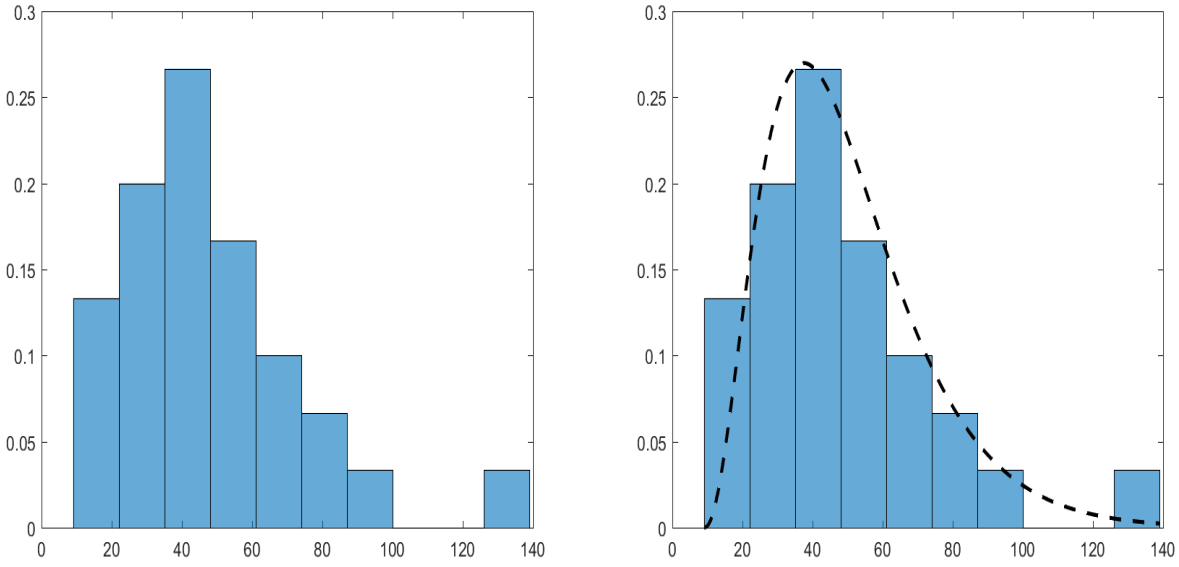


Fig. 2: Approximation of the pdf, Example 4.4

Stem-and-Leaf Plots

Stem-and-leaf plots are similar to histograms, although they carry more information. Namely, they also show how the data are distributed *within columns*. To construct a stem-and-leaf plot, we need to draw a stem and a leaf. The first one or several digits form a "stem", and the next digit forms a "leaf". Other digits are dropped; in other words, the numbers get rounded. For example, the number 139 can be written as

$$13 \mid 9$$

with 13 going to the stem and 9 to the leaf, or as

$$1 \,\big|\, 3$$

with 1 joining the stem, 3 joining the leaf, and the digit 9 being dropped. In the first case, the leaf unit equals 1 and the stem unit is 10, while in the second case, the leaf unit is 10 and the stem unit is 10^2, showing that the (rounded) number is not 13, but 130. The stem and leaf units *must be carefully specified* for each such plot.

Example 4.5. For the CPU times in Example 4.4 (sorted increasingly),

$$\begin{array}{cccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\ 36 & 36 & 37 & 38 & 42 & 43 & 46 & 48 & 54 & 55 \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 \end{array}$$

let us draw a stem-and-leaf plot with leaf unit 1 (i.e., the last digits form a leaf). The remaining digits go to the stem. Each CPU time is then written as

$$10 \text{ “stem”} + \text{ “leaf”},$$

making the following stem-and-leaf plot

0	9
1	5 9
2	2 4 5
3	0 4 5 5 6 6 7 8
4	2 3 6 8
5	4 5 6 6 9
6	2 9
7	0
8	2 2 9
9	
10	
11	
12	
13	9

Turning this plot by 90 degrees counterclockwise, we get a histogram with $10-$unit bins (because each stem unit equals 10). Thus, all the information seen on a histogram can be obtained here too. In addition, now we can see *individual* values within each column.

Stem-and-leaf plots can also be used to compare two samples. For this purpose, one can put two leaves on the same stem.

Example 4.6. The following two samples represent transmission times (in seconds) of signals - known as "pings"- from two different locations.

L1: $0.0156,\ 0.0396,\ 0.0355,\ 0.0480,\ 0.0419,\ 0.0335,\ 0.0543,\ 0.0350,$
$0.0280,\ 0.0210,\ 0.0308,\ 0.0327,\ 0.0215,\ 0.0437,\ 0.0483,$
L2: $0.0298,\ 0.0674,\ 0.0387,\ 0.0787,\ 0.0467,\ 0.0712,\ 0.0045,\ 0.0167,$
$0.0661,\ 0.0109,\ 0.0198,\ 0.0039.$

Let us sort the two samples in increasing order.

L1: $0.0156,\ 0.0210,\ 0.0215,\ 0.0280,\ 0.0308,\ 0.0327,\ 0.0335,\ 0.0350$
$0.0355,\ 0.0396,\ 0.0419,\ 0.0437,\ 0.0480,\ 0.0483,\ 0.0543,$
L2: $0.0039,\ 0.0045,\ 0.0109,\ 0.0167,\ 0.0198,\ 0.0298,\ 0.0387,\ 0.0467$
$0.0661,\ 0.0674,\ 0.0712,\ 0.0787.$

Since all numbers start with $0.0...$, we choose a stem unit of 0.01, a leaf unit of 0.001 and drop the last digit. We construct the following two stem-and-leaf plots (two in one), one to the left (L1) and one to the right (L2) of the stem.

						0	3	4	
					5	1	0	6	9
			1	1	8	2	9		
0	2	3	5	5	9	3	8		
		1	3	8	8	4	6		
					4	5			
						6	6	7	
						7	1	8	

Looking at these two plots, we can see about the same average ping from the two locations. Also, we realize that the first location has a more stable connection, because its pings have lower variability (i.e., lower variance).

Scatter Plots and Time Plots

Scatter plots are used to see and understand a relationship between two variables. These can be temperature and humidity, experience and salary, age of a network and its speed, number of servers and the expected response time, etc. To study the relationship, both variables are measured on each sampled item. For example, temperature and humidity during each of n days, age and speed of n networks, or experience and salary of n randomly chosen computer scientists are recorded. Then, a **scatter plot** consists of n points on an (x, y)-plane, with x- and y-coordinates representing the two recorded variables.

Example 4.7. Protection of a personal computer largely depends on the frequency of running antivirus software on it. One can set to run it every day, once a week, once a month, etc. During a scheduled maintenance of computer facilities, a computer manager records the number of times the antivirus software was launched on each computer during 1 month (variable X) and the number of detected viruses (worms) (variable Y). The data for 30 computers are given in the table below.

X	30	30	30	30	30	30	30	30	30	30	30	15	15	15	10
Y	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0

X	10	10	6	6	5	5	5	4	4	4	4	4	1	1	1
Y	0	2	0	4	1	2	0	2	1	0	1	0	6	3	1

Is there a connection between the frequency of running antivirus software and the number of worms in the system? A scatter plot of these data is given in Figure 3(a). It clearly shows that the number of worms reduces, in general, when the antivirus is employed more frequently. This relationship, however, is not certain because no worm was detected on some "lucky" computers although the antivirus software was launched only once a week (4 times a month) on them.

Looking at the scatter plot in Figure 3(a), the manager realized that a portion of data is hidden there because there are identical observations. For example, no worms were detected on 8 computers where the antivirus software is used daily (30 times a month). Then, Figure 3(a) may be misleading. When the data contain identical pairs of observations, the points on a scatter plot are often depicted with either numbers or letters (e.g., "A" for 1 point, "B" for two identical points, "C" for three, ..., "H" for eight, etc.). You can see the result in Figure 3(b).

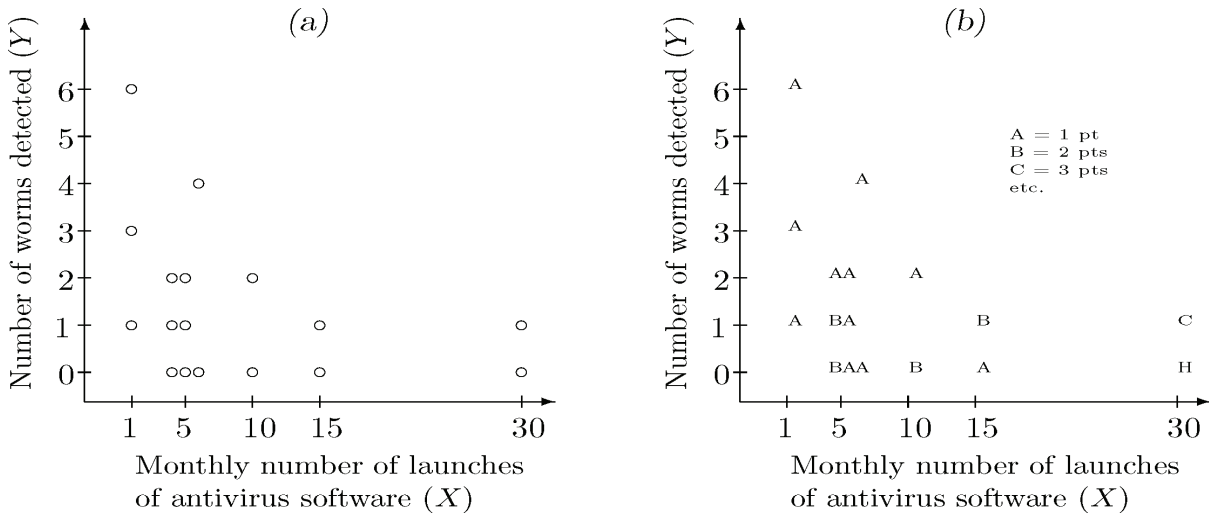


Fig. 3: Scatter plots for Example 4.7

When we study time trends and development of variables over time, we use **time plots**. These are scatter plots with x-variable representing time.

Example 4.8. Here is how the world population increased between 1950 and 2012 (Figure 4). We can clearly see that the population increases at an almost steady rate.

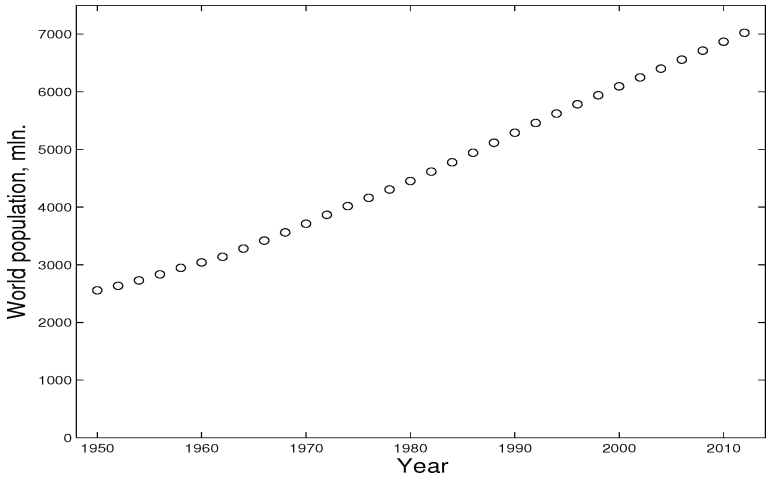


Fig. 4: Time plot of the world population in 1950–2012, Example 4.8

The actual data will be given and studied, later on in Chapter 5 (Correlation and Regression). We will estimate the trends seen on time plots and scatter plots and even make forecasts for the future.

Chapter 3. Calculative Descriptive Statistics

In the previous chapter we have considered some graphical methods for getting an idea of the shape of the density function of the population from which the sample data was drawn. Some characteristics, such as symmetry, regularity can be observed from these graphical displays of the data. Next, we consider some statistics that allow us to summarize the data set analytically. Simple **descriptive statistics** measuring the location, spread, variability and other characteristics can be computed immediately. It is hoped that these will give us some idea of the values of the parameters that characterize the entire population from which the sample was pooled. We are looking mainly at two types of statistics: *measures of central tendency*, i.e. values that locate the observations with highest frequencies (so, where most of the data values lie) and *measures of variability*, that indicate how much the values are spread out.

1 Measures of Central Tendency

These are values that tend to locate in some sense the "middle" of a set of data. The term "average" is often associated with these values. Each of the following measures of central tendency can be called the "average" value of a set of data.

1.1 Mean

Definition 1.1. *The **(arithmetic) mean*** ($\boxed{\text{mean}}$) *of the data $x_1, \ldots, x_N$ is the value*

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{1.1}$$

For grouped data, $\begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}$,

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{n} f_i x_i.$$

Remark 1.2. Some immediate properties of the arithmetic mean are the following:

1. The sum of all deviations from the mean is equal to 0. Indeed,

$$\sum_{i=1}^{N} (x_i - \overline{x}) = \sum_{i=1}^{N} x_i - N\overline{x} = 0.$$

2. The mean minimizes the mean square deviation, i.e. for every $a \in \mathbb{R}$,

$$\sum_{i=1}^{N} (x_i - a)^2 \geq \sum_{i=1}^{N} (x_i - \overline{x})^2 .$$

A straightforward computation leads to

$$\begin{aligned} \sum_{i=1}^{N} (x_i - a)^2 &= \sum_{i=1}^{N} [(x_i - \overline{x}) - (a - \overline{x})]^2 \\ &= \sum_{i=1}^{N} (x_i - \overline{x})^2 - 2(a - \overline{x}) \sum_{i=1}^{N} (x_i - \overline{x}) \\ &+ N (a - \overline{x})^2 \\ &\geq \sum_{i=1}^{N} (x_i - \overline{x})^2 , \end{aligned}$$

since the second term is 0 and the third term is always nonnegative.

Example 1.3. Let us recall the example from the previous chapter (Example 4.3), where to evaluate the effectiveness of a processor, a sample of CPU times for $N = 30$ randomly chosen jobs (in seconds) was considered:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19

The *mean* CPU time is

$$\overline{x} = \frac{70 + 36 + \ldots + 56 + 19}{30} = 48.2333 \text{ seconds.}$$

We may conclude that the mean CPU time of *all* the jobs handled by that particular processor is about the same, "near" 48.2333 seconds. In other words, we try to estimate the *population mean* by the *sample mean*. How good would that approximation be? We will learn later how to assess the

accuracy of our estimates.

Example 1.4. Let us assume that the value $x = 139$ (that seemed extreme, out of place, when we looked at the histogram) was *not* in this sample. Then the mean would be

$$\overline{x}_1 \ = \ 45.1034,$$

somewhat lower.
Now, in the other direction, let us suppose that the CPU time of one more job (a heavier one) is recorded and it is found to be 30 minutes $= 1800$ seconds. The mean of the new sample is

$$\overline{x}_2 \ = \ 104.7419 \text{ seconds},$$

way larger than the first value!

1.2 Median

One disadvantage of the sample mean is its *sensitivity to extreme observations*. As we have seen in the previous example, one extreme value can significantly shift the value of the mean, to the point where it becomes almost irrelevant.
The next measure of location is the *median*, which is much less sensitive than the mean.

Definition 1.5. *The **median*** ($\boxed{\text{median}}$) *is the value $\overline{M}$ that divides a set of ordered data X into two equal parts, i.e. the value with the property that it is exceeded by at most a half of observations and is preceded by at most a half of observations.*

A sample is always *discrete*, since it consists of a finite number of observations. Then, computing a sample median is similar to the case of discrete distributions. In simple random sampling, all observations are equally likely, and thus, equal probabilities on each side of a median translate into an equal number of observations. There are two cases, depending on the sample size N.

If the sorted primary data is

$$x_1 \leq \ldots \leq x_N,$$

then

$$\overline{M} = \begin{cases} x_{k+1}, & \text{if } N = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, & \text{if } N = 2k \end{cases}.$$

Remark 1.6. The median may or may not be one of the values in the data.

Example 1.7. Let us find the median for the data in Example 1.3 (the CPU times).

Since there are $N = 30$ observations, there are two middle values, the 15th and the 16th entries.

$$\begin{array}{cccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 \\ 36 & 36 & 37 & 38 & \mathbf{42} & \mathbf{43} & 46 & 48 & 54 & 55 \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 \end{array}$$

Then the median is $\overline{M} = 42.5$.

Remark 1.8. For an even number of observations, the median can be chosen to be any number between the two middle values. So in the previous example, we could say that any number in the interval $(42, 43)$ is a median.

Example 1.9. Let us add again the extreme value of 30 minutes $= 1800$ seconds. The new sample

$$\begin{array}{ccccccccccc} 9 & 15 & 19 & 22 & 24 & 25 & 30 & 34 & 35 & 35 & \\ 36 & 36 & 37 & 38 & 42 & \mathbf{43} & 46 & 48 & 54 & 55 & \\ 56 & 56 & 59 & 62 & 69 & 70 & 82 & 82 & 89 & 139 & 1800 \end{array}$$

has 31 observations, there is only one middle value (the 16th entry), so the median of the new sample is

$$\overline{M}_2 \ = \ 43.$$

Notice that the new value differs very little from the previous one and is *still relevant*, unlike the mean. So the median is a *robust* statistic, not being influenced (so much) by outliers.

1.3 Mode

Definition 1.10. *A **mode** Mo of a random variable X is a value with the highest pdf, i.e., it is the point with the highest concentration of probability, $Mo = \operatorname{argmax}\{f(x)\}$. A **sample mode**, $\overline{x}_{mo}$, of a set of data is a most frequent value.*

Remark 1.11. Notice from the wording of the definition that the mode may not be unique. A distribution can have one mode $-$ **unimodal**, two modes $-$ **bimodal**, three modes $-$ **trimodal**, or more $-$ **multimodal**.
When the pdf of a continuous distribution has multiple local maxima, it is common to refer to *all* of the local maxima as modes of the distribution.
If every value occurs only once in a sample, we say that there is **no mode**.

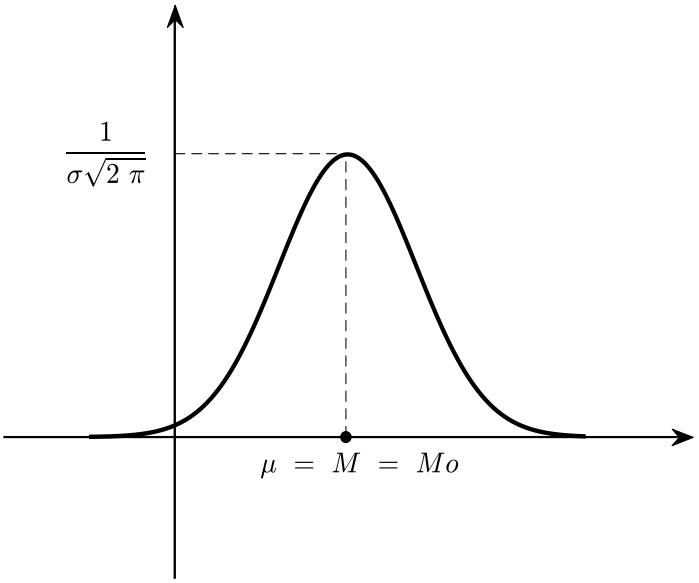


Fig. 5: Normal Distribution (unimodal)

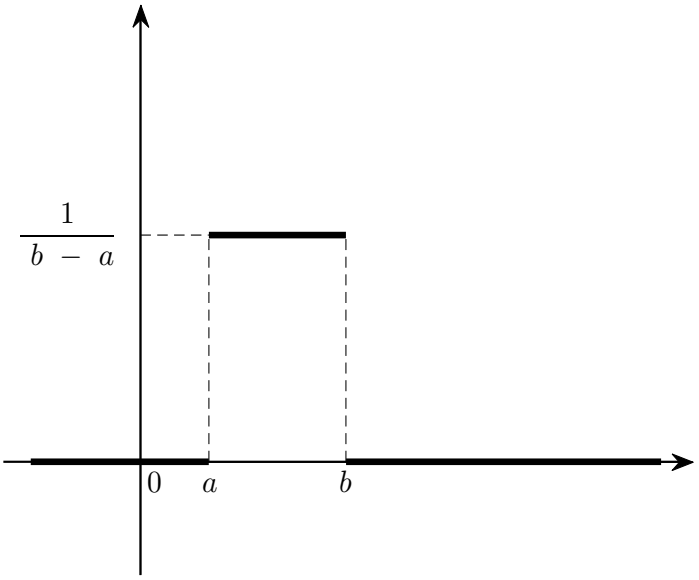


Fig. 6: Uniform Distribution (multimodal)

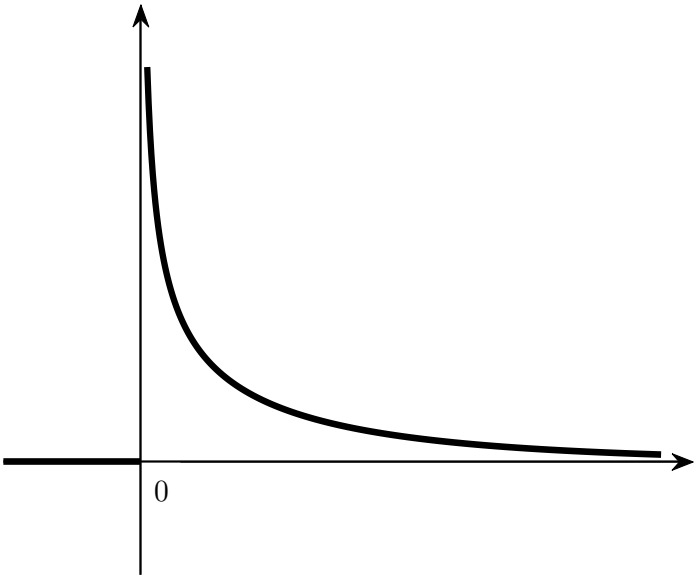


Fig. 7: χ^2 Distribution (no mode)

For data drawn from symmetric distributions, we have

$$\overline{x} = \overline{M} = x_{mo}.$$

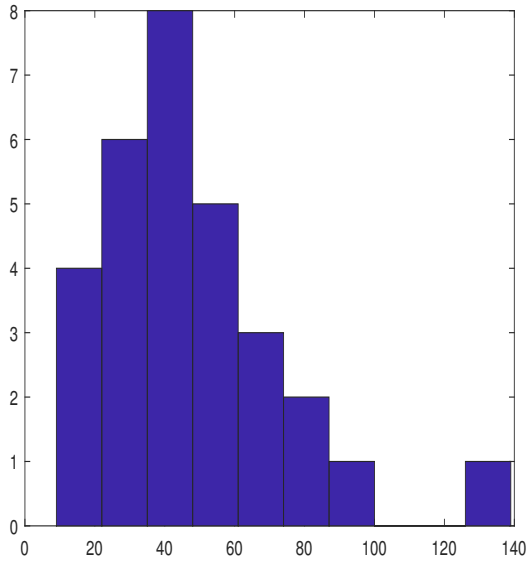
This is true, for instance, for the Normal distribution which is unimodal (Figure 5). For a Uniform $U(a, b)$ distribution, *all* values in the interval $[a, b]$ are modes (Figure 6), while the $\chi^2(1)$ distribution (with $\nu = 1$ degree of freedom) has no mode (Figure 7).
In general,

$$x_{mo} \approx \overline{x} - 3(\overline{x} - \overline{M}).$$

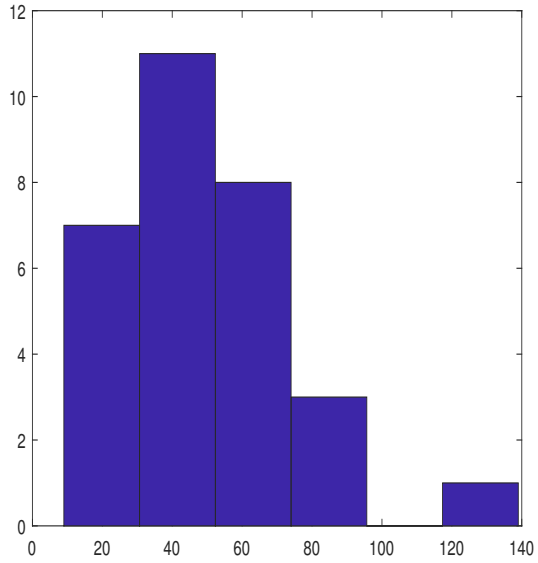
This empirical formula was given by K. Pearson.

Example 1.12. In our example about the CPU times (Example 1.6), the values $35, 36, 56$ and 82 appear twice, while all the other values have a frequency of 1. So all four are modes, this is multimodal data.

9	15	19	22	24	25	30	34	**35**	**35**
36	**36**	37	38	42	43	46	48	54	55
56	**56**	59	62	69	70	**82**	**82**	89	139



(a) $n = 10$ bins



(b) $n = 6$ bins

Fig. 8: Modal class

If we group the data into 10 classes, then the *modal class* is the third one, $(35, 48]$, with modal mark 41.5 (Figure 8(a)). If we have only 6 classes, then the second one is the modal class, $[30.7, 52.4)$, with mark 41.55 (Figure 8(b)).

2 Measures of Variability

Once we have located the central values of a set of data, it is important to measure the *variability*, whether the data values are tightly clustered or spread out. At the heart of Statistics lies variability: measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability and making decisions in the presence of it. We need to know how "unstable" the data is and how much the values differ from its average or from other middle values. These numbers will have small values for closely grouped data (little variation) and larger values for more widely spread out data (large variation).

The measures of variation will also help us assess the reliability of our estimates and the accuracy of our forecasts.

2.1 Quantiles, percentiles and quartiles

Consider the primary data $X = \{x_1, \ldots, x_N\}$. The first two measures of variation give a very general idea of the spread in the data values.

Definition 2.1. *The **range*** (range) *of X is the difference*

$$x_{max} - x_{min}.$$

If the values of X are sorted in increasing order, then the range is $x_N - x_1$.

Definition 2.2. *The **mean absolute deviation*** (mad) *of X is the mean of the absolute value of the deviations from the mean, i.e. the value*

$$MAD_1 = \frac{1}{N}\sum_{i=1}^{N} |x_i - \overline{x}|.$$

*The **median absolute deviation*** (mad) *of X is the median of the absolute value of the deviations from the median, i.e. the value*

$$MAD_2 = \text{median}\{|x_i - \overline{M}|\}.$$

Like the median, the median absolute deviation is not influenced by extreme values, whereas the mean absolute deviation is.

Next, following the idea behind the definition of the median, we define values that divide the data into certain percentages. We simply replace 0.5 in its definition by some probability $0 < p < 1$.

Definition 2.3. *Let X be a set of data sorted increasingly, $p \in (0, 1)$ and $k = 1, 2, \ldots, 99$.*

(1) *A **sample** p-**quantile*** (quantile) *is any number that exceeds at most $100p\%$ of the sample and is exceeded by at most $100(1-p)\%$ of the sample.*

(2) *A k-**percentile*** (prctile) *P_k is a $(k/100)$-quantile. So, P_k exceeds at most $k\%$ and is exceeded by at most $(100-k)\%$ of the data*

(3) *The **quartiles** of X are the values*

$$Q_1 = P_{25}, \;\; Q_2 = P_{50} = \overline{M} \;\text{ and }\; Q_3 = P_{75}.$$

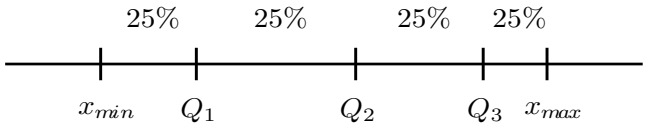


Fig. 9: Quartiles

Definition 2.4. *Let X be a set of sorted data with quartiles Q_1, Q_2 and Q_3.*

(1) *The **interquartile range*** (iqr) *is the difference between the third and the first quartile*

$$IQR = Q_3 - Q_1. \tag{2.2}$$

(2) *The **interquartile deviation** or the **semi interquartile range** is the value*

$$IQD = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}. \tag{2.3}$$

(3) *The **interquartile deviation coefficient** or the **relative interquartile deviation** is the value*

$$IQDC = \frac{IQD}{\overline{M}} = \frac{Q_3 - Q_1}{2Q_2}. \tag{2.4}$$

Remark 2.5.

1. The interquartile deviation is an absolute measure of variation and it has an important property: the range $\overline{M} \pm IQD$ contains approximately 50% of the data.

2. The interquartile deviation coefficient $IQDC$ varies between -1 and 1, taking values close to 0 for symmetrical distributions, with little variation and values close to ± 1 for skewed data with large variation.

Example 2.6. Consider again the CPU times (in seconds) for $N = 30$ randomly chosen jobs (sorted ascendingly):

9	15	19	22	24	25	30	**34**	35	35
36	36	37	38	42	43	46	48	54	55
56	56	**59**	62	69	70	82	82	89	139

Solution For this example, the range is

$$139 - 9 = 130 \text{ seconds}$$

and the mean and median absolute deviations are

$$\begin{aligned} MAD_1 &= 19.6133, \\ MAD_2 &= 13.5. \end{aligned}$$

To determine the quartiles, notice that 25% of the sample equals $30/4 = 7.5$ and 75% of the sample is $90/4 = 22.5$ observations. From the ordered sample, we see that the 8th element, 34, has 7 observations to its left and 22 to its right, so it has *no more* than 7.5 observations to the left and *no more* than 22.5 observations to the right of it. Hence, $Q_1 = 34$.

Similarly, the third quartile is the 23rd smallest element, $Q_3 = 59$. Recall from Example 1.7 that the second quartile (the median) is $Q_2 = \overline{M} = 42.5$. Then

$$\begin{aligned} IQR &= 59 - 34 = 25, \\ IQD &= IQR/2 = 12.5, \\ IQDC &= IQD/Q_2 = 0.2941. \end{aligned}$$

The interval

$$\overline{M} \pm IQD = [30, 55]$$

contains 14 observations.

The value of the $IQDC$ is close neither to 0, nor to the values ± 1. So the data doesn't show strong symmetry or strong asymmetry. This may be due to the extreme values 9 and/or 139.

■

Example 2.7. A computer maker sells extended warranty on the produced computers. It agrees to issue a warranty for x years if it knows that only 10% of computers will fail before the warranty expires. It is known from past experience that lifetimes of these computers have a Gamma distribution with parameters $\alpha = 60$ and $\lambda = 1/5$ years. Compute x and advise the company on the important decision under uncertainty about possible warranties.

Solution We just need to find the tenth percentile of the specified Gamma distribution and let $x =$

P_{10}. In Matlab, that would be computed (as the *inverse* of the cdf) by

$$x = gaminv(0.1, 60, 1/5) = 10.0624.$$

Thus, the company can issue a 10-year warranty rather safely.

■

Remark 2.8. For populations or very large data sets, calculating exact percentiles can be computationally very expensive since it requires sorting all the data values. Machine learning and statistical software use special algorithms (such as linear interpolation) to get an approximate percentile that can be calculated very quickly and is guaranteed to have a certain accuracy.