Chapter 5. Graphical Display and Descriptive Statistics

> The greatest moments are those when you see the result pop up in a graph or in your statistics analysis - that moment you realize you know something no one else does and you get the pleasure of thinking about how to tell them.
>
> Emily Oster

What is statistics?

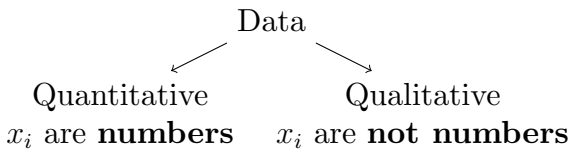
So far we have been talking about *theoretical* concepts, such as random variables and their probability distributions. In the matter of fact we are not given them. For example, the true distribution of a variable is usually not achievable for us. Consider random variable X is a monthly salary of a Russian citizen. How to get the distribution of this variable? First we need to gather the data about the salary of all citizens. So, now we are switching to *practice* and start working with real data. Each value that was gathered is called an **observation**. If we will get information about salary for *all* the Russian citizens, we would have the **population** of X.

An observation is denoted by x_i. For example, if Masha's sister salary is 40 000 roubles, then $x_{\text{sister}} = 40\ 000$ is one observation. All existent observations x_i, $i = 1, 2, \ldots, N$, contribute the population where N is the population size – total number of citizens. If so, we would be able to construct the true probability distribution of X, calculate probabilities (relative frequencies) for each value of X. This is called a population distribution.

However, in many cases it would require too large amount of time and money to get information about the whole population. Then, we only take a small but representative part of it to analyze X. This is what we call **a sample of X** - a set of observations x_i, $i = 1, 2, \ldots n$, where n is the **sample size**. It is much smaller than population size ($n >> N$). For example, we can use the data on salaries of 1000 randomly chosen citizens – a sample of size 1000 - to analyze monthly salary in Russia. Based on such data you can construct the **sample distribution** of X – observed relative frequencies for different values of X. In statistics we usually work with samples, as population data is rarely available. You will learn more on sampling in chapter 6.

When you are looking at a dataset, remember that it can represent the whole population or only a sample taken from it. It is important to distinguish between these two types of datasets, since they should be analyzed differently.

Types of data. According to the basic classification data can be quantitative or qualitative.



Another name for quantitative data is **numerical data**. It could be discrete (number of assignments submitted) or continuous (measured level of temperature). Qualitative data is **non-numerical** or **categorical**, it consists of a finite set of categories, such as "yes" or "no". Examples of qualitative data are: eye color, student's name.

Start your work with data. After the data are collected, what is the next step? The raw dataset usually looks messy so that you cannot make sense of what is observed. Therefore, the data should first be organized properly. We can provide a quick insight into data with a graph, or we can summarize data using quantitative measures, called descriptive statistics. This chapter will analyze these tools.

The choice of a particular instrument depends, above all, on the type of data you work with. There are different methods of graphical representation available for both qualitative and quantitative data. Most of descriptive statistics can be calculated only for quantitative data.

Graphical representation and descriptive statistics are the two ways to meaningfully describe and analyze the observed distribution of a random variable whether it is based on *population*, or on *sample* data.

Generally speaking, Statistics consists of two branches. **Descriptive Statistics** summarizes and describes the data while **Inferential Statistics** allows making further generalized conclusion. We will proceed the course afterwards with the inferential statistics.

Graphical Representation of Data

> You can't do anything you could not picture yourself doing.
>
> unknown

Graphs are used to provide quick visual representation of data. The first immediate impression helps to notice any patterns in data such as its shape, dispersion and location of the center. It allows us to make brief conclusions from the first glance. We are going to analyze the following graphs:

1. Dot plot
2. Bar chart
3. Stem-and-leaf plot
4. Histogram (labeling both frequency and relative frequency)
5. Cumulative frequency plot
6. Boxplot (discussed in the descriptive statistics section of this chapter)

Dot Plot

Dot plot is a very intuitive graph which represents the number of observations in each category by plotting the corresponding number of dots. It could be applied to all types of data, both qualitative and quantitative.

The dataset presented on the left graph shows the number of snakes of different species in a zoo. This is a dot plot for qualitative data. From this graph for instance you can say that there are 2 anacondas in this Zoo. The next graph shows dot plot for the quantitative data. The dataset was obtained in the following way: 15 people were asked if they have siblings and if yes, how many. The dotplot below pictures the results.

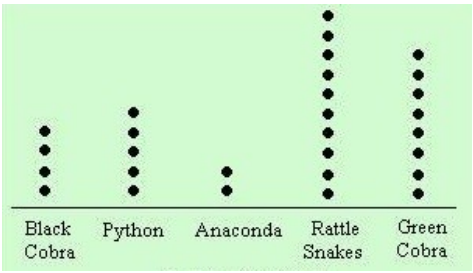


Figure 1: dotplot of "Types of snakes in the Zoo"

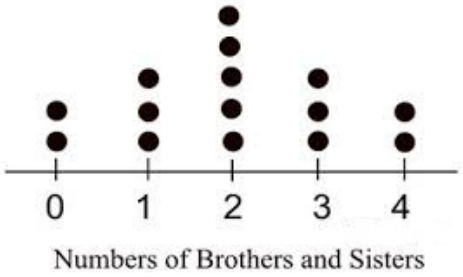


Figure 2: dotplot of "Number of brothers and sisters"

Bar Chart

Bar chart shows the number of observations in different categories through the *heights* of the corresponding bars. Sizes of categories can be measured in absolute (number of observations) or relative terms (proportions of all observations). The bar chart below shows the distribution of colors of M&M candies in a pack. For example, we can infer that there were 5 red candies.

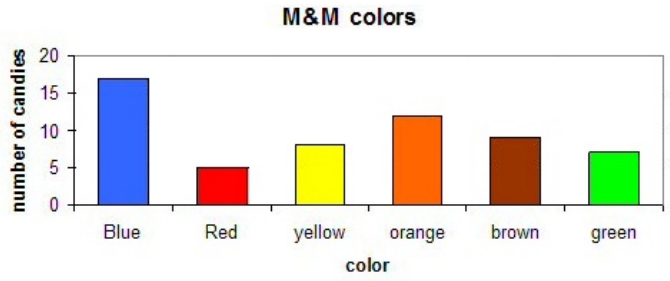


Figure 3: bar chart

Stem and Leaf Plot

Stem-and-leaf plot or just stemplot is a graph representing all observations in a set, sorted and grouped. It took its name from the way it looks like - the leaves are "growing" on the stem.

SPRINTERS.Masha is preparing for athletic competition at HSE. In order to evaluate her chances, she gathered data on sprint results in seconds of 20 best university athletes. She's got the following dataset: 10.17, 10.23, 10.25, 10.28, 10.31, 10.32, 10.34, 10.35, 10.41, 10.44, 10.45, 10.46, 10.49, 10.52, 10.55, 10.64, 10.68, 10.69, 10.71, 11. Below is the stem and leaf plot representing this data.

stem	*leaves*
10.0	
10.1	7
10.2	3 5 8
10.3	1 2 4 5
10.4	1 4 5 6 9
10.5	2 5
10.6	4 8 9
10.7	1
10.8	
10.9	
11.0	0

key: 10.7|1 = 10.71 seconds

Figure 4: time for 100-meters sprint

The graph contains the stem-values on the left and leaf values on the right. They are separated by a vertical line. Each leaf represents one observation. It can be

read as the sum of stem and leaf value. For example notation 10.2|3 means 10.23 = 10.2+0.03.

The number of leaves equals the number of observations. If you have repeated values, like 42, 42, 45, you would plot it as 4| 2 2 5, you cannot skip any of them. Also note, that even if some stems have no leaves, you could not skip them! You have to plot those stem values anyway, like 10.8 for instance. Otherwise, you will not be able to infer this missing space from the graph.

The main advantage of stem and leaf plot, over other types of graph, is that it shows the dataset without loss of information. You can recover the value of each particular observation, which is not possible with the histogram for example.

Histogram

Histogram is a graph showing the distribution of observed values by putting them in groups of equal length. Therefore, in order to draw a histogram you should first divide your data into *groups*, or *intervals*. Let us check how to do it on the **"SPRINTERS"** dataset.

The values are ranging from 10.17 to 11. Let us divide our range in 10 segments, and check how many observations are there in each segment. We will have the length 0.1 of each interval. Consider the interval $[10.1, 10.2]$. It has one element – 10.17. Consider the next interval $[10.2, 10.3]$. It has 3 elements – 10.23, 10.25 and 10.28. Continue the same procedure for all the other intervals. What to do if an element falls on the boarder of two intervals? Choose the unique strategy for all intervals which border, right or left, you are going to include. Here we will include the left boarder, namely [10.1, 10.2).

The number of observations in each interval is called **frequency**, sometimes denoted by f. Plot the frequency on the histogram by corresponding bars with width 0.1 equal to chosen interval length.

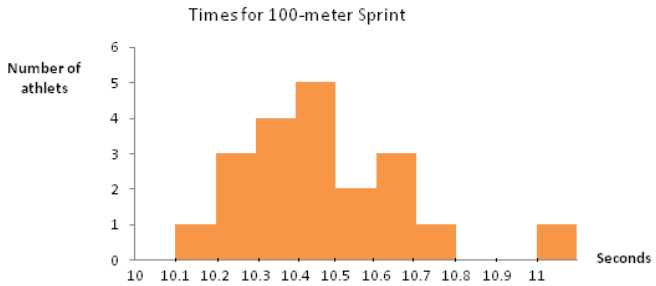


Figure 5: histogram time for 100-meters sprint

You can see from the graph that interval $[10.3, 10.4)$ has 4 elements, $[10.4, 10.5)$ – 5 elements, $[10.5, 10.6)$ – 2, and so on. Interval $[10.8, 10.9)$ has no elements.

How to choose the interval length? We look at the whole range of values from minimum to maximum and decide on how many intervals it should be divided into. There is no strict rule about the number of intervals, but crucial is that they are

of the *same length*. It is up to you how many intervals it would be, keep in mind that too few number makes the graph less informative, and too many – difficult to perceive. In simple problems it is usually optimal to construct from 5 to 10 columns.

The frequency can be measured in absolute terms, as above, or in relative terms as well. For instance, there are 5 observations in the interval $[10.4, 10.5)$. If you divide 5 by the total number of observations, you will get the relative frequency $5/20 = 0.25$. Frequency answers the question how many observations fall into an interval, and relative frequency – the proportion of values. Recall from chapter 1, that relative frequency is the proportion of number of times some event occurred in n observations. Here it means the proportion of athletes with results belonging to some interval in the total number of athletes. The histogram in relative frequencies is shown below.

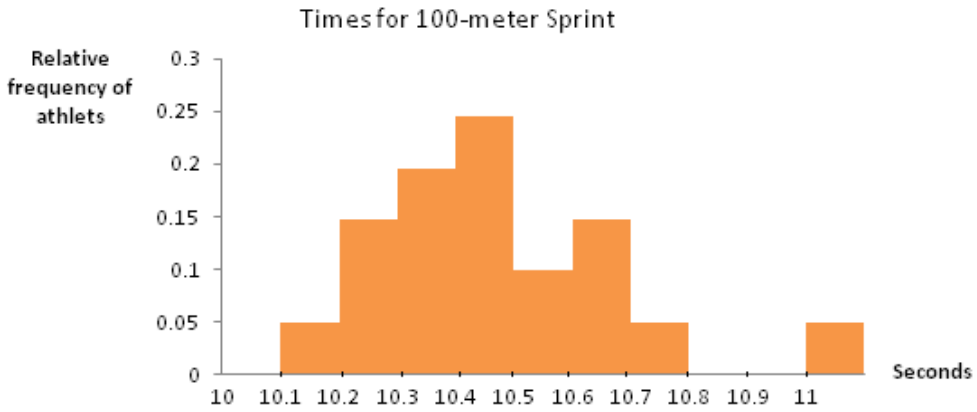


Figure 6: histogram time for 100-meters sprint, relative frequency

The shape of the histogram did not changed. This is always the case transferring to relative frequencies.

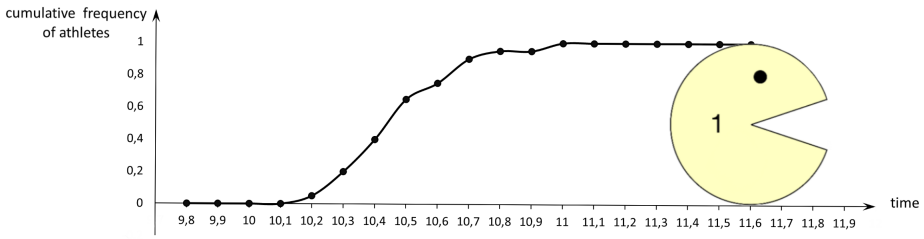
A histogram can only be constructed for quantitative data. Its bars stay close to each other unless there is a gap there in your dataset.

If you have *qualitative* data analogous graph must be called a bar chart. You can't use the term histogram in this case. It differs, except for the type of data,

by always having spaces between the bars. These spaces emphasize the idea that categories represent different realizations of some quality, which cannot be put on axis of continuous values. For example, think of a bar chart for some students' university faculties. Values on X-axis are department names, Economy, Medicine, Media, etc. Y-Axis shows frequency. It is necessary to keep distance between categories to show there is no "value" between Media and Medical. Contrary, columns of histogram are side to side. Space between them should be interpreted as a gap. In the histogram above there is no spaces between columns except for one place, because there are no observations between 10.8 and 11. Histogram allows to see patterns of data distribution, such as symmetry.

Cumulative frequency plot

Cumulative frequency plot, or ogive, shows the *accumulated* frequency up to a reference value. In other words it shows what percent of observations are no higher than each possible value. Cumulative frequency plot for the "SPRINTERS" dataset is presented below.



Look at the point (10.4, 0.4). It means that 40% of sprinters run as fast as 10.4 sec, their resulting time is no higher than 10.4 sec. How did we get the value 0.4? Go back to the histogram. To the left of 10.4 there are three columns, showing relative frequencies 0.05, 0.15 and 0.2. If we sum up all the frequencies, we'll get the total of 0.4, the cumulative frequency up to 10.4 sec.

Ogive can be used to infer relative frequencies as well. What percent of sprinters has shown time above 10.3 seconds? The point (10.3, 0.2) suggests that 0.2 of all athletes had resulting time below the 10.3 seconds. So, the other $1 - 0.2 = 80\%$ of all athletes ran slower than 10.3. What percent of athletes had time between 10.3 and 10.4 seconds? It is the difference of frequency below 10.4 sec and below 10.3 sec, namely $0.4 - 0.2 = 0.2$. You can check this answer looking at the column between these values on the histogram.

Note that cumulative frequency plot is always a non-decreasing graph.

Recall that you are already familiar with the concept of cumulative distribution and its graph from the chapters 2 and 4. It shows the probability that a variable does not exceed some value. For example, below is the table derived on page ... of Chapter 2 for the "DRIVING LICENSE" example.

The lowest row in the table contains cumulative probabilities for the corresponding values of X. As you can see, cumulative probability of value x is the sum of its

X	1	2	3	4	5	6	7	8
$P(X = x)$	0.05	0.1	0.3	0.25	0.15	0.07	0.05	0.03
$P(X \leqslant x)$	0.05	0.15	0.45	0.7	0.85	0.92	0.97	1

probability and probabilities of all values below x, e.g. $P(X \leqslant 3) = 0.45 = 0.3 + 0.1 + 0.05$. Plotting these values, you will get the cumulative probability plot.

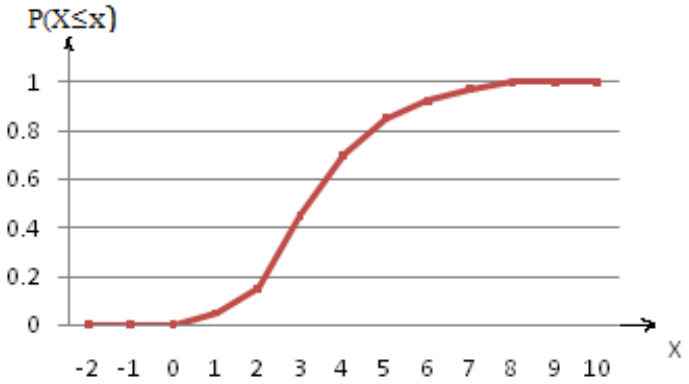


Figure 7: cumulative probability plot

As you can see it is a non-decreasing function, which values increase from 0 to 1 as x increases from the value below minimum up to the maximal value. At each possible value cumulative probability function rises by probability of this value. E.g. $P(X \leqslant 3)$ is higher than $P(X \leqslant 2)$ by exactly $P(X = 3)$. Here it is constructed based on the known probability distribution. The only difference with the previous graph is that you don't really have true probabilities for the times shown by sprinters, you only have observed values. Therefore you should construct the graph based on the *relative frequencies*. The resulting graph is called then cumulative frequency plot.

About graphs

There is one more type of graph used in this course, called a box plot. It will be introduced in the second section of this chapter.

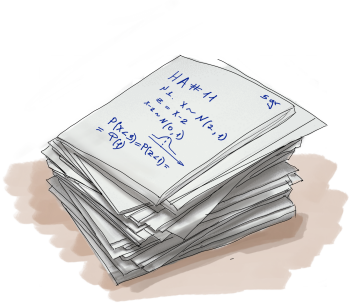
We've discussed several approaches to the graphical representation. As you have seen the same data could be represented on different graphs. This topic seems quite simple, although don't get too relaxed. First, you should accurately choose the graph type according to the type of data you have. Second, graphs are drawn to reveal important features of data. Any graph simplifies the dataset emphasizing some patterns of it, and ignoring others. So, if a graph is not fitted accurately, it can be misleading. Careless choice and interpretation of a graph may result in wrong understanding of data. When solving problems always make sure that your graph is clear and accurate leaving no doubt on what information it provides. Do not forget to label your graph. AP Statistics scoring guide harshly penalize for the lack of titles!

Descriptive Statistics

> All right everyone, line up alphabetically according to your height.
>
> Casey Stengel

Descriptive statistics summarizes data using quantitative characteristics. Just looking at a data set, what can you say about its characteristics? For instance, look at the following data set.

HOME ASSIGNMENTS. ICEF Statistics course requires weekly home assignments. Masha is interested how well her classmates perform this task. For it she gathered data on the total number of home assignments on Statistics submitted during the 1st semester by each of the 25 students in her group. She's got the set of 25 values: 23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4, 30, 20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32.



How well students perform in general? What is the average number of completed assignments, whether results are concentrated at some number or dispersed, or whether there is a gap between low and good students? The data set itself doesn't provide an obvious insight. It has to be first put in order and then processed. So, you need some instruments for processing. Descriptive statistics are those instruments to progress with your data. It is a formula to produce a single number (e.g. an average score) out of the whole set of numbers (e.g. set of students' scores). Using descriptive statistics you can analyze datasets and compare them with the others. For example, you can compare AP scores of Statistics of ICEF students with the applicants from University of Singapore.

We can divide descriptive statistics into three main categories. The first category describes the location of observed numbers along data set. They answer questions about the *center* of the dataset (mean, median and mode), its extreme values (minimum and maximum) and intermediate values (quartiles and percentiles).

The second category of descriptive statistics represents the *variability,* or spread among observations. It answers the questions of typical variability (standard deviation, variance), extreme variability (Range) and spread of middle observations (Interquartile range).

The third category of descriptive tools is used to describe the *shape* of data distribution (such as symmetry).

Measures of Center

Intuitively, the first step in describing any dataset is finding its center. We can call center a typical value. It exist three ways to do that, with mean, median or mode,

sometimes called for ease MMM.

Mean

Mean is the arithmetic average of the all observed values. It is the sum of values divided by their total number.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Masha wants to find the mean number of submitted home assignments in her group. That would be: $\overline{X} = \frac{23+45+...+32}{25} = 35.08$. Check whether she has calculated mean correctly.

When mean is calculated on sample data it is called a **sample mean**. Contrary, population mean $\mu_x = E(X)$ introduced in Chapter 2 is calculated on population data. It requires the knowledge of all elements in the population, namely the distribution of X. You will learn more on this distinction in Chapter 8.

If some observations occur in dataset more than once, that is have frequency f_i more than one, it is more convenient to use another formula. You first multiply observation by the number of times it appears and then sum up all the products. It is provided below, where f_i is the frequency of every observation x_i.

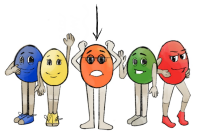
$$\overline{X} = \frac{\sum_{i}^{m} x_i f_i}{n}$$

Median

Median is the number that *separates* lower half of the sample from the upper half. If all observations are ordered and divided into 2 equal parts the median would be the *border* between them. Then 50% of observations lie above the median and the other 50% below.

How to find the median?

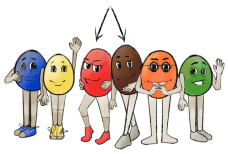
First, you need to put all observations in ascending order, from lowest to highest, and assign them numbers from 1 to n. As a result, the 1st observation is the minimum and the nth observation is the maximum. If the number of observations is *odd*, the median is very easily found - it is the one in the middle of the list. Just like the orange candy on the picture to the right.



Formally, if the number of observations is odd the *median* is the *observation* with number $\frac{n+1}{2}$ in the list of sorted observations.

For example, if you have 5 observations, median is the 3rd observation, since $\frac{5+1}{2} = 3$.

If the number of observations is *even*, you have 2 observations in the middle. On the picture to the right those are the orange and the red candies. Median equals to the average of those two central observations.



Formally, if the number of observations is even the *median* is the *average of two central observations*, those numbered $\frac{n}{2}$ and $\frac{n}{2}+1$ in the list of sorted observations.

Then, if you have 6 observations for example, median equals to the arithmetic mean of the 3$^{\text{rd}}$ and the 4$^{\text{th}}$ observation.

Recall "HOME ASSIGNMENTS" dataset. To find the median, first Masha puts the observations in ascending order. The number of observations $n = 25$ is odd. Thus, median equals to the observation with number $\frac{25+1}{2} = 13$ in the list. This way Masha finds that median $= x_{13} = 38$.

4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, (38), 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56

Mode

Mode is the observation with the maximum frequency. It is the most *popular* observation. If most of people wear red boots, "red boots" would be the mode because of their maximum frequency.

In the above example mode equals 44 – the only number occurred three times.

Basically, mode is not exactly the characteristic of center. There may be more than one mode in the distribution! If numbers of occurrences of several different elements are equal or almost equal, all of them can be viewed as modes.

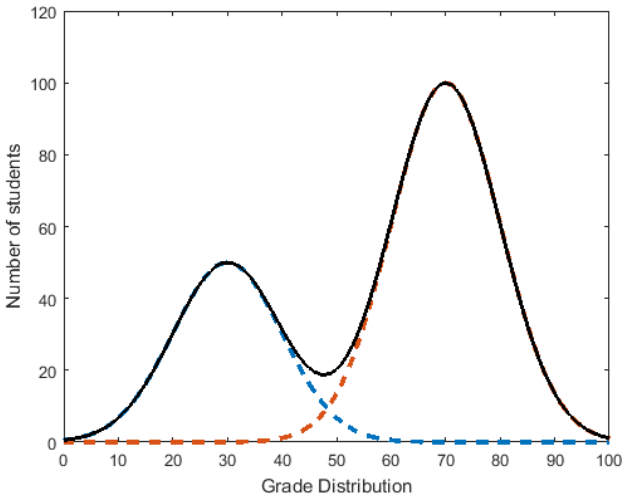


Figure 8: bimodal distribution

It may happen that the most frequent are observations with high values, that is the right side of a set. Then the mode will not be a good characteristics of a center.

The mode can be found both for quantitative and qualitative data, e.g., in the zoo example for distribution of snake species, rattle is the mode.

Mean or Median: Which is better?

Mean gives equal weight to all of the observations in a dataset. Therefore quite a few highly extreme observations, very large or very small, pull the characteristic of center towards them and away from more *typical* observations. So, the mean is a *value-sensitive* characteristic. The median though does not have such a problem. The high values of extreme observations do not affect the general order, so do not affect the median value.

STARTUP Consider an example that illustrates this. The dataset represents monthly income of 29 employees working for start-up project in Moscow (in thousands of roubles):

13, 15, 21, 21, 22, 25, 25, 25, 27, 28, 28, 30, 30,33, 34, 35, 35, 35, 39, 40, 40, 40, 41, 45, 45, 50, 55, 63, 1235.

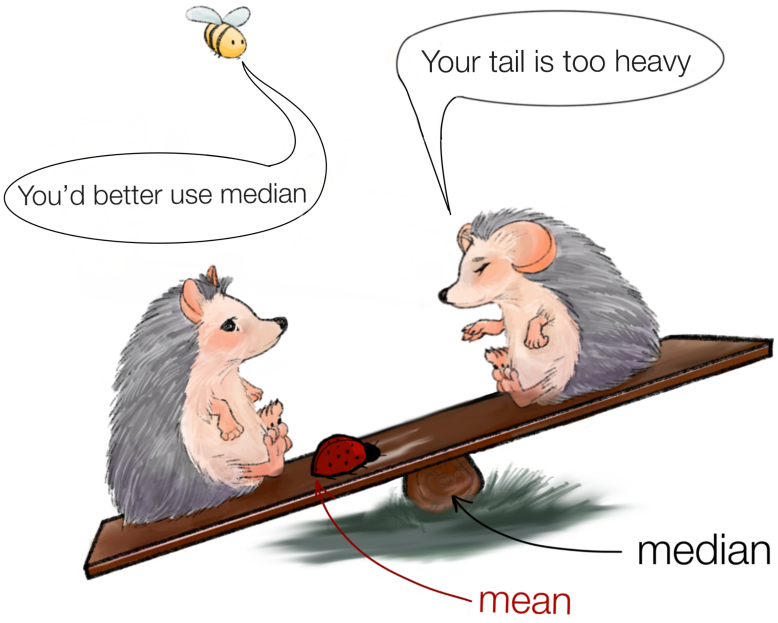
The latest observation is the monthly income of the owner of the enterprise. He earns more than a million roubles per month.

Let's calculate the mean and the median for monthly income of a worker. $\overline{X} = 75$, median $= 34$. They differ substantially. Which one is a better estimate of a typical income in this dataset?

In this case median is better. 34 000 roubles seems to be a nice measure of center, while mean 75 000 roubles seems a bit unrealistic. No one has income close to 75 000 roubles and thus 75 000 cannot be viewed as a typical income.

Even one millionaire in the dataset can substantially skew the value of mean income upwards up.

This kind of difference is typical only for skewed distributions with the so-called "heavy tails". In this case median is a better measure of central value. For symmetric data distributions mean and median produce approximately equal numbers.



However, even for skewed distributions mean is not a non-sense. It does not provide the value of typical observation, but it represents other important information.

For example, comparing start-up projects you may want to compare their income *per worker*, to learn which one is more effective in terms of profit-to-labour balance. Then, you would prefer mean over median. Analogically economists use GDP per capita to compare welfare of countries. If you compare absolute values of GDP then large countries inevitably will be viewed as most wealthy. This does not take into account that large countries also have more citizens. Therefore total GDP is divided by population size so that the amount of wealth, hypothetically available to a citizen in different countries, can be meaningfully compared. Thus, median is better when you search for the central value of X, while mean is better as a measure of overall average value.

Other reference points

Minimum and **maximum** are the extreme values of a dataset – the lowest and the highest observations.

Quantile is a value that separates dataset. Quantiles are the cut points dividing the dataset into parts of equal size, that is having equal frequencies. If distribution is continuous these parts are intervals. Depending on how many parts quantile separates the set there exist different types of quantiles: quartiles, percentiles, deciles. There are several ways to calculate quantiles, analyzed below.

Quartiles

Quartiles separate dataset in 4 parts, by one quarter each. There is one less quantile than the number of groups created. Thus quartiles are the three cut points that will divide a dataset into four equal-sized groups. The three quartiles are: lower, upper and middle.

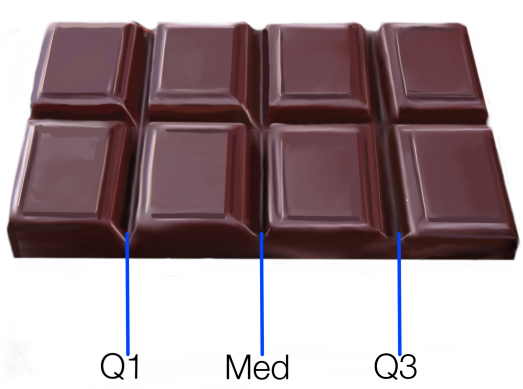
Lower quartile (LQ) is the value that separates the lower 25% of observations from the higher 75% of observations. It is also called the **1st quartile** Q1.

If you put all observations in ascending order and divide them into 4 parts of equal size (quarters), lower quartile would be the value between the highest observation in the first quarter and the lowest observation in the second quarter. You can also view lower quartile as a median of the *lower half* of your observations.

Upper quartile (UQ) separates the lower 75% of observations from the higher 25% of observations. It is also called the **3rd quartile** Q3. It is the threshold between the third and fourth quarters of sorted observations. It can be viewed as the median of the *upper half* of observations.

Finally, second quartile is the value that separates the 2nd and the 3rd quarters or, equivalently, the lower and the upper halves of all observations. That coincides with the definition of median! Therefore, the *median* is also can be called a **2nd quartile** Q2. Median is the middle (second) quartile!

The quartiles in a dataset then will look like this: Q1, Med (Q2), Q3.



How to calculate quartiles?

Masha wants to find lower and upper quartiles for her "HOME ASSIGNMENTS" dataset.

Method 1. By this method quartiles are just the medians of first and second halves.

1. She puts the numbers in *ascending* order and finds the *median*. We've already shown that median $= x_{13} = 38$.

2. Median *divides the set into 2 parts*: to the right and to the left of it (the median itself should be excluded from both parts). Thus, the lower half is: 4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36. The upper half is: 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56.

3. Masha finds LQ as the *median of the lower half* and UQ as the *median of the upper half* of observations. Both halves contain 12 observations, which is an even number. Therefore, median is between observations with number $\frac{12}{2}$ and $\frac{12}{2}+1$. Thus, LQ and UQ lie between 6^{th} and 7^{th} observations in the lower and the upper datasets correspondingly. LQ $= \frac{x_6+x_7}{2} = \frac{23+30}{2} =$ **26.5** and the upper quartile is $UQ = \frac{44+45}{2} =$ **44.5**.

4, 11, 12, 20, 23, 23, [Q1 is here] 30, 31, 32, 33, 34, 36, (38), 38, 40, 41, 44, 44, 44, [Q3 is here] 45, 47, 48, 49, 54, 56

12+1+12 total number of observations

Method 2. There is another method of finding LQ and UQ which is based on the more general notion of percentile.

Percentiles

Percentiles divide set into 100 parts. Each **percentile** is the number that divides set into 2 parts, the given proportion p of observations is below the percentile, and the other $(1-p)$ proportion is above the percentile. For example, the lower quartile is 25^{th} percentile, since it separates 25%. The median is the 50^{th} percentile and so on.

Formal definition is the following: $P(X \leqslant x_a) = p$, where x_a is the percentile and p is the fixed probability. This fixed probability is set to name the percentile and is also called the level of a percentile.

There is a general method for calculating the p-percentile.

1. **Put observations in ascending order**:

   4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, 38, 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56.

2. **Calculate** the number $p \cdot (n + 1)$, where n is the number of observations in the dataset. For example, to find LQ for the "HOME ASSIGNMENTS" dataset: $n = 25$, $p = 0.25$ the resulting number is: $p \cdot (n + 1) = 0.25 \cdot (25 + 1) = 6.5$.

3. **Divide the number into integer part k and fractional part a**. In our example, $6.5 = 6 + 0.5$. Thus, $k = 6$, $a = 0.5$. $k + a$ indicates the location of the percentile in the sorted list of observations, e.g. 6.5 means that LQ is between 6^{th} and 7^{th} observations.

4. Find the percentile as $x_p = x_k + a \cdot (x_{k+1} - x_k)$.

   Thus, $LQ = x_{0.25} = x_6 + 0.5 \cdot (x_7 - x_6) = 23 + 0.5 \cdot (30 - 23) = \mathbf{26.5}$. Since LQ is between 6^{th} and 7^{th} observations, it equals x_6 plus some proportion a of the distance between x_7 and x_6.

$$p \cdot (n + 1) = k + a, \;\; x_p = x_k + a \cdot (x_{k+1} - x_k)$$

In the same way for median: $p \cdot (n + 1) = 0.5 \cdot 26 = 13$, $k = 13$, $a = 0$, median $= x_{0.5} = x_{13} + 0 \cdot (x_{14} - x_{13}) = x_{13} = 38$.

For the UQ: $p \cdot (n + 1) = 0.75 \cdot 26 = 19.5$, $k = 19$, $a = 0.5$, $UQ = x_{0.75} = x_{19} + 0.5 \cdot (x_{20} - x_{19}) = 44 + 0.5 \cdot (45 - 44) = \mathbf{44.5}$.

Note, different methods may produce slightly different values of quartiles. That happens when quartile is between two observations x_{k+1} and x_k (or $a \neq 0$). That's Ok. Quartile is the number which lies between the two quarters of observations. Strictly speaking, any number between x_{k+1} and x_k satisfies this definition.

However, it is conventional to give a single number for a quartile, therefore you should apply any but one method to produce the answer. Remember to clearly explain it.

Variability of dataset

Once the center is defined, the next important question is how large the difference between valuesin the dataset is. This is referred to as variability or spread among the observations. Do you remember the example about the choice among three assets with different risks (page 5 of Chapter 2)? We have shown that although mean result is important, another factor that should be taken into account is the variability.

Range

The extreme (the largest possible) difference between the numbers in the dataset is the difference between the maximum and the minimum values. It is called Range:

$$\text{Range} = X_{max} - X_{min}$$

In Masha's example Range is $56 - 4 = 52$.

Interquartile Range

The Interquartile range, IQR, is the difference between the Upper and Lower quartiles. It answers the question what is the *range of middle 50%* of observations, in other words shows the variability of middle observations.

$$\text{IQR} = \text{UQ} - \text{LQ}$$

Let's calculate IQR for the homeworks dataset: $\text{IQR} = 44,5 - 26,5 = 18$.

Standard Deviation

Standard deviation answers the question how far do observations typically fall from its mean. It is the square root of variance, namely the average squared deviation:

$$s_x = \sqrt{Var(X)} = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1}}$$

You can also use the more convenient version of the formula for variance:

$$Var(X) = \frac{\sum x^2 - n\overline{X}^2}{n-1}$$

Proof

$$\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} (X_i^2 - 2X_i\overline{X} + \overline{X}^2) = \sum_{i=1}^{n} {X_i}^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2 =$$

$$\sum_{i=1}^{n} {X_i}^2 - 2\overline{X} \cdot n\overline{X} + n\overline{X}^2 = \sum_{i=1}^{n} {X_i}^2 - n\overline{X}^2$$

Thus, $\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum x^2 - n\overline{X}^2}{n-1}$

What does it mean? Well, let's try to derive this formula together with Masha. Imagine, that you are trying to invent a formula to evaluate variability between observations. You might prefer to take a center as the reference point and measure the distance from each observed value x_i to the center. Thus, you get a set of deviations of each observation i from the mean: $(x_i - \overline{X})$. Masha says that variability in X may well be measured by average deviation. What if we simply calculate the mean of deviations – sum of them divided by n? For a big observation deviation is positive (it lies above the mean $\overline{X}$), and for a small observation it is negative. Then, the simple average of all deviations would reduce to zero: $\frac{\sum(X_i - \overline{X})}{n} = 0$. So, it does not work this way.

In order to avoid this problem and get the value that would reflect the overall scale of deviations, we suggest to take each deviation squared: $\sum (X_i - \overline{X})^2$. Thus, each summand is now positive and the sum of deviations will not reduce to zero. Now, to get an average deviation the sum $\sum (X_i - \overline{X})^2$ should be divided by the number of observations. Well, that's true for the case when you have the data on the whole population. Then, you should use the formula for population variance: $\sigma^2 = Var(X) = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$. However, the formula above is for sample variance, and we divide by $(n-1)$ instead of n.

Variance provides the result in squared units. E.g. for the "home assignments" dataset variance approximately equals 182 homeworks squared. The squared answer is very inconvenient for interpretation. Standard deviation solves this problem by taking the square root of Var.

Standard deviation is measured in the same units as the variable X itself. It can be easily interpreted. For example, standard deviation of the number of home assignments is approximately 13.49. We also know that $\overline{X} = 35.08$. It means that, while students submitted 35 home assignments on average (exact value 35.08), typical result was observed to vary from 35 by 13,5 assignments approximately.

Note that both s and $Var(X)$ are non-negative.

Depending on what you are working with, the whole available population or only a part of it that we call sample, you need to choose different formulas for variance and standard deviation.

Which formula for variance to choose?

you have the whole *population* — $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$

you have only *sample* — $s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$

No, hedgehog. We won't answer your question here.

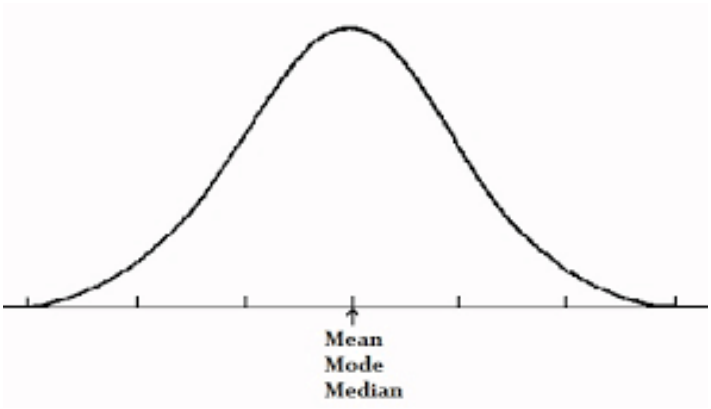
It happens that when the true mean μ is unknown, the right formula provides more accurate values. s is called a **sample standard deviation** and s^2 is a sample variance. In Chapter 2 you've learned about population variance or variance of a random variable. It is the expected value of the squared deviation of X from its population mean μ, $Var(X) = \sigma^2 = E[(X-\mu)^2]$. We've also shown that for the dataset containing the whole population you should use the formula $Var(X) = \sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$. Check that they are equal: $\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N} = E[(X-\mu)^2]$ Note that sample variance s^2 and population variance σ^2 are not the same. The proof will be provided in Chapter 8.

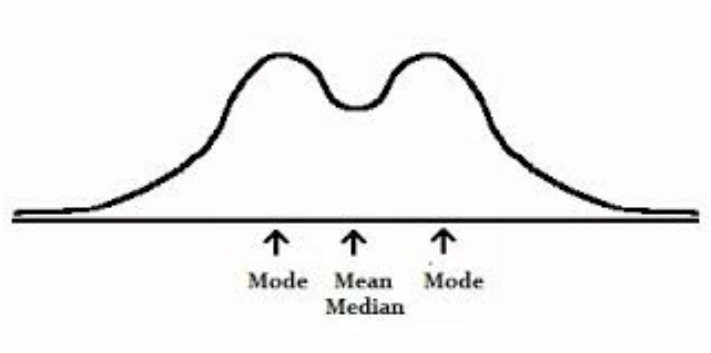
Additional intuition: why squared deviations? Squaring gives different "weights" to deviations of different size, introducing a "penalty" to large deviations. What is meant by a penalty? Note, that deviation of $\frac{1}{2}$ contributes $\frac{1}{4}$ to the variance value, while a deviation of 10 contributes the value of 100. Thus, the sum is very sensitive to large deviations, while allowing small ones to make only a tiny influence. This has a very important practical implication. In practice we don't care about small mistakes, which do not introduce significant risk for the accuracy of overall economic evaluation. However, big deviations may completely change the result and economic decisions based on it.

Shape

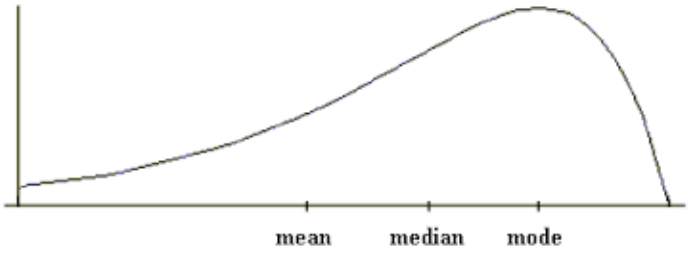
As we've seen in the 1st section of this chapter, visual appearance of the distribution (histogram, dotplot, stemplot, ...) is very informative. Therefore, when describing a dataset or comparing several datasets it is conventional to verbally describe the shape of distribution.

Distribution can be described as symmetric when the right half of a histogram looks as an approximate mirror image of a left half: Symmetric unimodal Symmetric bimodal





Otherwise, the distribution is described as non-symmetric. Sometimes non-symmetric distributions are described as being skewed to the left or to the right. If the right tail of a histogram looks much longer and thinner, the distribution is skewed *to the right*, or positively skewed towards the high values. Contrary, when the left tails is longer and thinner, the distribution is skewed *to the left*, or negatively skewed towards the lower values.



Note that the word 'skewed' can be interpreted as "stretched". That means from which side distribution is more stretched, to this side is the skewness. Another way to remember is distribution is skewed to the side from which distribution was hit to take this shape.

Sometimes the direction of skewedness is not visually obvious. Then you can help yourself comparing mean and median. Median is always located at a point which divides the graph into two parts with equal areas. It coincides with what we visually perceive as a "middle" of the dataset. Contrary, mean is value sensitive and it is

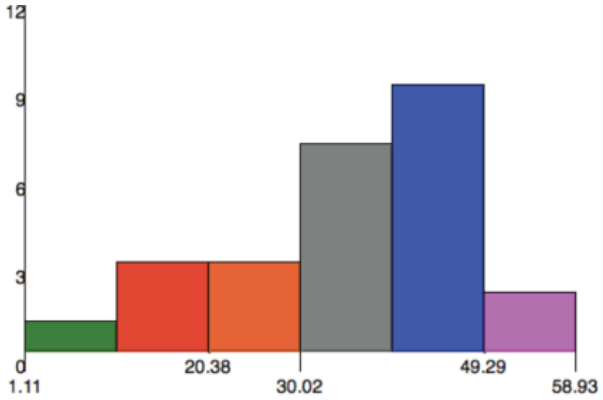
always skewed to the extremely small or extremely large observations. Thus, for right skewed distribution $\overline{X}$ is greater than the median and vice versa.

$$\overline{X} > \text{median significantly, the distribution is skewed to the right}$$

$$\overline{X} < \text{median significantly, the distribution is skewed to the left}$$

Thus, mean is located closer to the "tail" of a skewed distribution and goes to the direction of the skewedness.

Let's draw a histogram, for "home assignments" example:

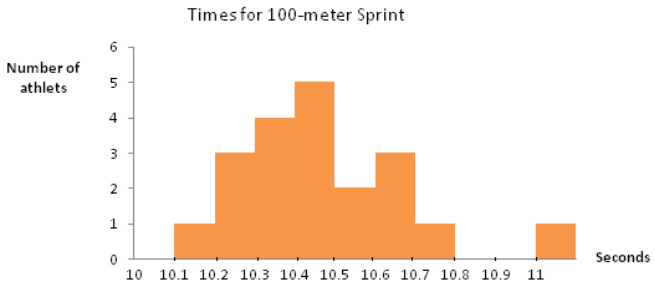


It can be seen that the distribution is skewed to the left. In such a case it must be true that $\overline{X}$ is smaller than the median. This is exactly the case since as we've calculated $\overline{X} = 35.08$ and median $= 38$.

Special features

Gaps

If the distribution contains an interval with no observations inside, it is said to have a gap in that interval. As you can see the "sprinters" example, there is a gap on the $[10,8; 11)$ interval.



Note that gaps can be seen on histograms, stem and leaf plots and dotplots. Boxplot does not reveal this feature of data.

Outliers

Outlier is an observation lying far enough from other observations in a data set. It reflects something extreme. For example, in a dataset on shoe size, outliers would be the size numbers of people who have extremely large or too small feet. The former have size number far above the others, the latter – far below. However, not all extremes are considered to be outliers. There exist the *cut* points to define an outlier. They are provided below and are called "fences":

$$\text{Lower fence} = LF = Q_1 - 1,5 \cdot IQR$$

$$\text{Upper fence} = UF = Q_3 + 1,5 \cdot IQR$$

Then, an outlier is an observation outside of these fences. Namely such an X_i which is above (to the right of) the upper fence $X_i > UF$ or below the lower fence $X_i < LF$. In other words, a value which is $1.5 \cdot IQR$ more than upper quartile, and $1.5 \cdot IQR$ less than lower quartile, $X_i > Q3 + 1,5 \cdot IQR$ and $X_i < Q1 - 1,5 \cdot IQR$.

Are there any outliers in the "home assignments" dataset?

Let's calculate the fences:

Lower fence $LF = Q_1 - 1.5 \cdot IQR = 26.5 - 1.5 \cdot 18 = -0.5$

Upper fence $UF = Q_3 + 1.5 \cdot IQR = 44.5 + 1.5 \cdot 18 = 71.5$.

There are no observations outside the interval $[-0,5; 71,5]$, thus the dataset contains no outliers.

You can also check that there is an outlier in the dataset on incomes of participants of a startup. Salary of the business owner is an outlier.

On graphs it is usually clearly seen which observation is a potential candidate for outlier, if such exist. They stand away from other observations. However you should every time formally check, if it indeed is an outlier.

What to do if you have an outlier?

1. An outlier may indicate an error in records (e.g. someone misrecorded 75 as 750). In this case just exclude outlier from the dataset.

2. An outlier may indicate specific features of data. In this case what to do with it depends on how descriptive statistics are going to be used. First, you can make calculations separately of outlier and then mention that there was such an observation. Alternatively, you can include it in your calculations and after that provide careful interpretation of the resulting numbers.

For example in the "startup" example both answers are acceptable:

1. Excluding a millionaire from the dataset we get $\overline{X} \approx 33.571$. We can use it keeping in mind that an extremely high value of income was also observed.

2. The mean of 74 is highly skewed towards the value of outlier: $X_{29} = 1235$.

Is it necessary to check the presence of outliers in any dataset before solving the problem? Not necessarily. You should first ask yourself the following questions:

1. Is it clearly stated in the conditions of a problem that you need to check the presence of outliers?

2. Do you suspect that there is an outlier in the dataset? For instance, in the "home assignments" example you are not expected to check that and it's Ok if you don't check. However, if you asked to describe the distribution of incomes in the "startup" it is highly required that you notice a single outstanding observation and check that it is an outlier.

The presence of outliers may strongly affect descriptive statistics, so if there are outliers, we need to be careful.

Boxplot

Boxplot is a graph summarizing descriptive statistics. It represents important descriptive statistics visually. Below is the descriptive statistics on **"HOME ASSIGNMENTS"** dataset and the graph summarizing it.

Min=4
LQ=26,5
Med=38
UQ=44,5
Max=56
UF = 71,5; LF = -0.5

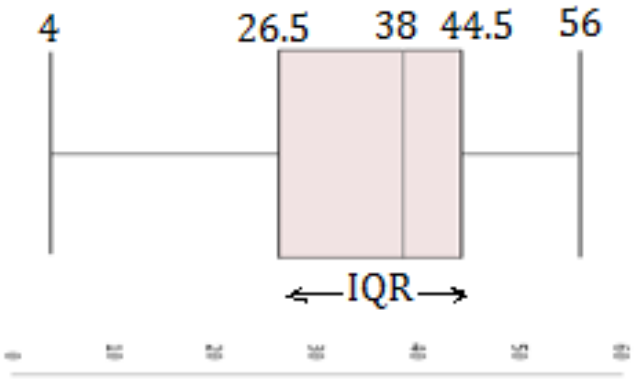


Figure 9: boxplot on number of home assigmnets

Sometimes it is also called a box-and-whisker plot. You plot a box with length eaual to IQR, and inside it you plot a line to denote the median. Outside of the box you plot the whiskers to minimum and maximum. If there exists an outlier, then you denote it as a single dot, and then the whisker stops at the lat observation thet is still inside the outliers fence. You can find an example of such a boxplot in problem of this chapter.

How to "read" shape from a boxplot
From the beginning it is not so easy to infer that shape of a distribution from a box plot, but it will just require you a bit of practice to get used. So, if a part of box is not long, that means the 25% of observations happened to be on smaller distance. That means the density of them would be high as exactly reflected on the left below. The long box and wishkers, however, reflect the skewedness. You can see this on the right side. That is the distribution below is skewed to the right.

Relative location of elements

Z-score
So far we've talked about methods of description of a dataset. Sometimes we face an opposite problem: to characterize an observation in the context of its environment or dataset. Z-scores and percentiles of observations serve this goal.

Z-score measures position of a point in the distribution. It is *the number of standard deviations* by which the observation stands away from the mean.

Precisely Z-score states by how many standard deviations a particular value is situated far from the mean (is higher than the mean). For example $z_i = 1$ states that the value z_i is 1 standard deviation greater than mean (to the right of it) and z-score $z = -1$ says that the value is 1 standard deviation below mean (to the left of it). Z-score: $z_i = \frac{x_i - \mu}{\sigma}$

Note that when population mean μ and population standard deviation σ are not known, you can use sample statistics in the formula instead: $z_i = \frac{x_i - \overline{X}}{s}$.

The procedure of converting the set of X_i's to the set of Z_i's is called **standartisation**. This is why in Chapter 5 Z is called a *standard* normal variable.

Intuition: $x_i - \overline{X}$ is the distance from mean. Dividing by s converts it to the number of standard deviations. Positive z_i indicates that observation x_i exceeds the mean, while negative z_i implies that x_i is smaller than the mean.

You can also do the reverse procedure. Given z_i you can find the initial observation x_i:

$$x_i = \mu + z_i\sigma$$

Compare two observations using Z-score. Standartisation (convertation to Z-scores) allows to compare observations from different populations.

SUMMER SCHOOL FRIENDS. Masha has got two friends from the LSE Summer School, Greta from Germany and Junko from Japan. It happened that after the seminars they started discussing which girl is taller. Greta (171 cm) turned out to be slightly taller than Junko (170cm), but the latter was arguing that she was the tallest in her class in Tokyo!

What is the right way to compare the girls' height?

Solution: It would not be correctly to compare heights of girls in absolute terms, because in general Japanese people tend to be relatively shorter compared to people of other nations.

Please, pay attention: These are two different variables!

It is not like comparing two values of the same variable – these are two values of different variables: one is measuring the height of a person from Japanse population, while the other – from a German population. Therefore, it is important not only to compare the girls' height values, but also to take their relative position in the distributions from which they are taken from. One reasonable way to do this is using z-scores.

The women's height in Germany and Japan have the parameters: $E(G) = 168$, $Var(G) = 15^2$ and $E(J) = 163$, $Var(J) = 10^2$.

Then z-scores of the girls are equal to:
German girl Greta: $z_G = \frac{171-168}{15} = 0.2$
Japanese girl Junko: $z_J = \frac{170-163}{10} = 0.7$

Z-scores are positive, indicating that both girls are higher than an average girl in each of the countries. But the z-scores also indicate the volume of that difference. Since z_J is by 0.5 of standard deviation (calculated as $0.7 - 0.2$) higher than z_G, we can say that Junko is *relatively* higher than Greta, taking into account her *initial* national *conditions*.

Percentile of an observation An alternative way to characterize the position of an observation in the population is finding its percentile. It is the *proportion* of observations in the population which *lies below* the observation of interest. It is just an inverse problem of finding an observation to be a percentile, discussed earlier.

Go back to "SUMMER SCHOOL FRIENDS" example. The height is normally distributed: $G \sim N(168, 15^2)$ and $J \sim N(163, 10^2)$.

Let's compare the percentiles of the girls' heights:
$P(G \leqslant 171) = P(z < 0,2) \approx 0.579$ 58-th percentile
$P(J \leqslant 170) = P(z < 0,7) \approx 0.758$ 76-th percentile

Thus, Greta is taller than 58% of girls in her country, while Junko is taller than 76% of girls in her country. Again, we come to the conclusion, that Junko is relatively taller.

Full score strategy

Comparison of distributions

1. Center.
2. Spread.
3. Shape.

\+ gaps/clusters/outliers and other peculiar features

Make conclusions in terms of the problem (which asset is better, which group of students is more successful, etc).

Examples are given in the problems section of this chapter.

You must be able to reproduce even being half awake

- dot plots can be used for any type of data
- stem and leaf plots are only for quantitative data
- histograms are for quantitative data, bar charts are for qualitative data
- size of intervals in histogram should be equal
- median separates two halves of a set
- quartiles Q1 and Q3 are the medians of lower and upper halves
- $Var(X) = \frac{\sum x^2 - n\overline{X}^2}{n-1}$
- outlier is an observation X_i such that $X_i > Q3 + 1,5 \cdot IQR$ and $X_i < Q1 - 1,5 \cdot IQR$.

Calculator Box

- To draw graphs: `Stat` → `DRAW` (set the required type of graph)
- To find all descriptive statistics: `Stat` → `Calc` → `1Var`
- To find $E(X)$, σ, $E(X^2)$:
  1. Put values of X into `List 1`, the corresponding probabilities into `List 2`
  2. `CALC` → `SET`. `1 Var XList List1` / `1 Var Freq List2` → `EXIT`
  3. `1VAR`

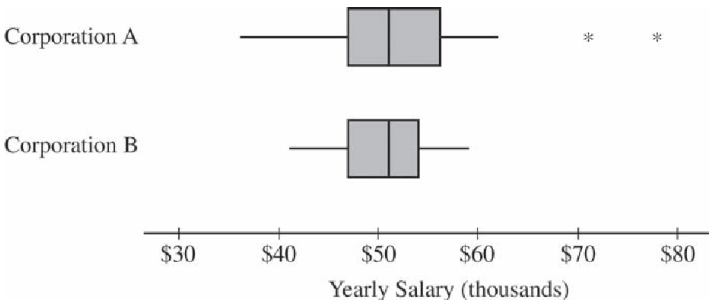
Now you have results:

$E(X)$	$\overline{X} = \sum X$
σ	σ_x
$E(X^2)$	$\sum x^2$

Sample AP Problems with solutions

Problem 1. AP 2015 №1

Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2009 the starting salary for an entry-level accountant position was \$36,000 a year at both corporations. At each corporation, data were collected from 30 employees who were hired in 2009 as entry-level accountants and were still employed at the corporation five years later. The yearly salaries of the 60 employees in 2014 are summarized in the boxplots below.



(a) Write a few sentences comparing the distributions of the yearly salaries at the two corporations.

(b) Suppose both corporations offered you a job for \$36,000 a year as an entry-level accountant.

(i) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

(ii) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

Solution:

(a) The median salary is approximately the same for both corporations.

The range of the salaries in Corporation A is greater than in Corporation B. Interquartile range is also slightly higher for Corporation A.

Distributions of salaries for both Corporations are quite symmetric.

Corporation A has two outliers, namely two highest salaries. Corporation B has no outliers.

(b) (i) 5 years after, at least several highest salaries are higher at Corporation A than at Corporation B. If I choose the offer from Corporation A, I might be able to make a higher salary in future. As could be seen from the graph there is probably higher possibility for career growth at Corporation A.

(ii) 5 years after, the minimum salary is higher at Corporation B rather than at Corporation A. It looks like at Corporation A some people are still

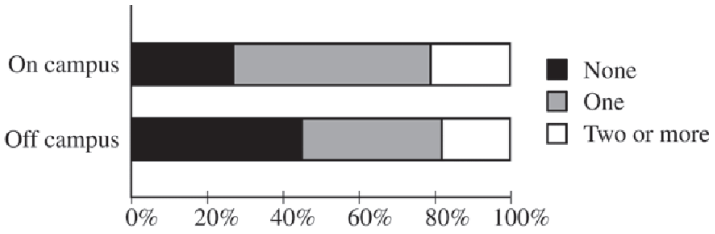
making the starting salary \$36,000 and never received a raise. Thus the promotion at Corporation B is probably more secured. At Corporation A in contrast there exists a possibility never to receive a raise in the salary.

Problem 2. AP 2014 №1 b

An administrator at a large university is interested in determining whether the residential status of a student is associated with level of participation in extracurricular activities. Residential status is categorized as on campus for students living in university housing and off campus otherwise. A simple random sample of 100 students in the university was taken, and each student was asked the following two questions.

- Are you an on campus student or an off campus student?
- In how many extracurricular activities do you participate?

The responses of the 100 students are summarized in the segmented bar graph shown.



Write a few sentences summarizing what the graph reveals about the association between residential status and level of participation in extracurricular activities among the 100 students in the sample.

Solution:

The graph reveals that on campus residents in this sample are more likely in general to participate in extra-curricular activities than off campus residents.

On campus residents have a greater proportion who participate in one activity (on campus: 0.515, off campus: 0.373) and a smaller proportion who participate in no extracurricular activities (on campus: 0.273, off campus: 0.448) than off campus residents. The proportions who participate in two or more extra-curricular activities are similar between the two groups but slightly greater for on campus residents (on campus: 0.212, off campus: 0.179).

Problem 3. AP 2014 №4 a

As part of its twenty-fifth reunion celebration, the class of 1988 (students who graduated in 1988) at a state university held a reception on campus. In an informal survey, the director of alumni development asked 50 of the attendees about their incomes. The director computed the mean income of the 50 attendees to be \$189,952. In a news release, the director announced, "The members of our class of 1988 enjoyed resounding success. Last year's mean income of its members was \$189,952!"

(a) What would be a statistical advantage of using the median of the reported incomes, rather than the mean, as the estimate of the typical income?

Solution:
The median is less affected by skewness and outliers than the mean. With a variable such as income, a small number of very large incomes could dramatically increase the mean but not the median. Therefore, the median would provide a better estimate of a typical income value.

Problem 4. AP 2013 №1 a
An environmental group conducted a study to determine whether crows in a certain region were ingesting food containing unhealthy levels of lead. A biologist classified lead levels greater than 6.0 parts per million (ppm) as unhealthy. The lead levels of a random sample of 23 crows in the region were measured and recorded. The data are shown in the stem-plot below.

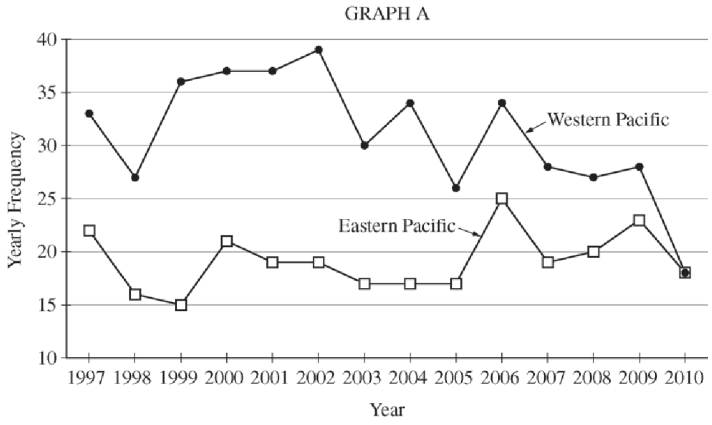
Lead Levels	
2	8
3	0
3	5 8 8
4	1 1 2
4	6 8 8
5	0 1 2 2 3 4
5	9 9
6	3 4
6	6 8
Key: 2|8 = 2.8 ppm	

(a) What proportion of crows in the sample had lead levels that are classified by the biologist as unhealthy?

Solution:
Four of the 23 crows in the sample had a lead level greater than 6.0 ppm. Therefore, the proportion of crows in the sample that were classified as unhealthy is $4/23 \approx 0.174$.

Problem 5. AP 2013 №6

Tropical storms in the Pacific Ocean with sustained winds that exceed 74 miles per hour are called typhoons. Graph A below displays the number of recorded typhoons in two regions of the Pacific Ocean – the Eastern Pacific and the Western Pacific – for the years from 1997 to 2010.



1. Compare the distributions of yearly frequencies of typhoons for the two regions of the Pacific Ocean for the years from 1997 to 2010.

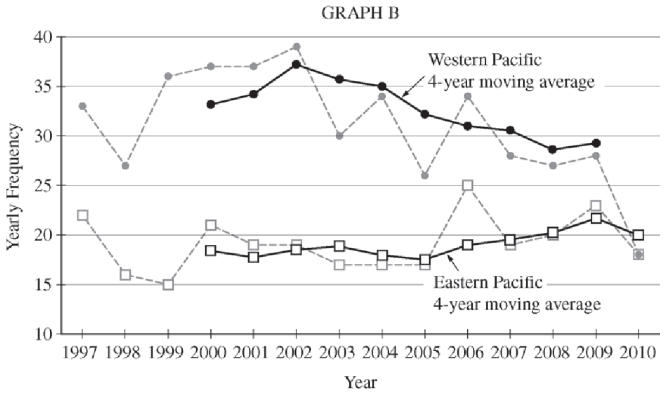
2. For each region, describe how the yearly frequencies changed over the time period from 1997 to 2010.

   A moving average for data collected at regular time increments is the average of data values for two or more consecutive increments. The 4-year moving averages for the typhoon data are provided in the table below. For example, the Eastern Pacific 4-year moving average for 2000 is the average of 22, 16, 15, and 21, which is equal to 18.50.

Year	Number of Typhoons in the Eastern Pacific	Eastern Pacific 4-year moving average	Number of Typhoons in the Western Pacific	Western Pacific 4-year moving average
1997	22		33	
1998	16		27	
1999	15		36	
2000	21	18.50	37	33.25
2001	19	17.75	37	34.25
2002	19	18.50	39	37.25
2003	17	19.00	30	35.75
2004	17	18.00	34	35.00
2005	17	17.50	26	32.25
2006	25	19.00	34	31.00
2007	19	19.50	28	30.50
2008	20	20.25	27	28.75
2009	23	21.75	28	29.25
2010	18	20.00	18	

3. Show how to calculate the 4-year moving average for the year 2010 in the Western Pacific. Write your value in the appropriate place in the table.

4. Graph B below shows both yearly frequencies (connected by dashed lines) and the respective 4-year moving averages (connected by solid lines). Use your answer in part (c) to complete the graph.



5. Consider graph B.

   (i) What information is more apparent from the plots of the 4-year moving averages than from the plots of the yearly frequencies of typhoons?

   (ii) What information is less apparent from the plots of the 4-year moving averages than from the plots of the yearly frequencies of typhoons?

Solution:

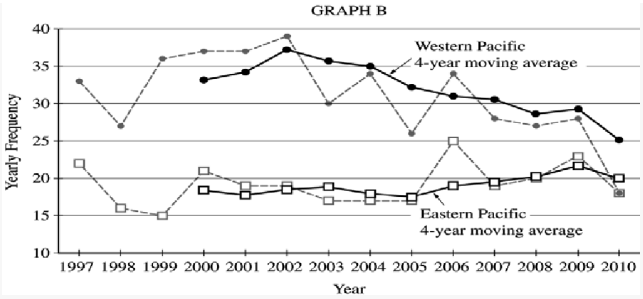
(a) The Western Pacific Ocean had more typhoons than the Eastern Pacific Ocean in all but one of these years. The average seems to have been about 31 typhoons per year in the Western Pacific Ocean, which is higher than the average of about 19 typhoons per year in the Eastern Pacific Ocean. The Western Pacific Ocean also saw more variability (in number of typhoons per year) than the Eastern Pacific Ocean; for example, the range of the frequencies for the Western Pacific is about 21 typhoons and only 10 typhoons for the Eastern Pacific.

(b) The Western Pacific Ocean had a decreasing trend in number of typhoons per year over this time period, especially from about 2001 through 2010. In contrast, the Eastern Pacific Ocean was fairly consistent in the number of typhoons per year over this time period, with a slight increasing trend in the later years from 2005 through 2010.

(c) The four-year moving average for the year 2010 in the Western Pacific Ocean is: $\frac{28+27+28+18}{4} = 25.25$.

The value is written in the lowest right empty box of the table.

(d)

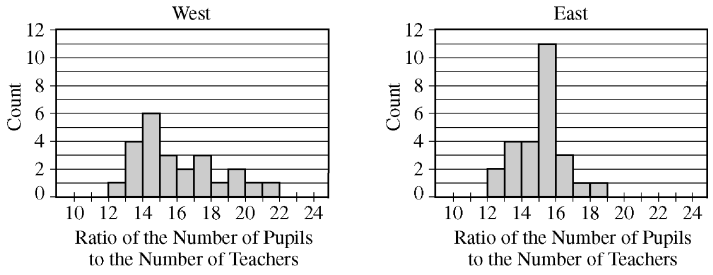


(e) (i) The overall trends across this time period were more apparent with the moving averages than with the original frequencies. The moving averages reduce variability, making more apparent the overall decreasing trend in number of typhoons in the Western Pacific Ocean and the slight increasing trend in the number of typhoons in the Eastern Pacific Ocean.

(ii) The year-to-year variability in number of typhoons is less apparent with the moving averages than with the original frequencies.

Problem 6. AP 2011 Form B №1

Records are kept by each state in the United States on the number of pupils enrolled in public schools and the number of teachers employed by public schools for each school year. From these records, the ratio of the number of pupils to the number of teachers (P-T ratio) can be calculated for each state. The histograms below show the P-T ratio for every state during the 2001–2002 school year. The histogram on the left displays the ratios for the 24 states that are west of the Mississippi River, and the histogram on the right displays the ratios for the 26 states that are east of the Mississippi River.



(a) Describe how you would use the histograms to estimate the median P-T ratio for each group (west and east) of states. Then use this procedure to estimate the median of the west group and the median of the east group.

(b) Write a few sentences comparing the distributions of P-T ratios for states in the two groups (west and east) during the 2001–2002 school year.

(c) Using your answers in parts (a) and (b), explain how you think the mean P-T ratio during the 2001–2002 school year will compare for the two groups (west and east).

Solution:

1. The median is the value with half of the P-T ratios at or below it and half of the values at or above it. For n observations in a group, use $\frac{n+1}{2}$ to find the position of the median in the ordered list of observations. For states west of the Mississippi ($n = 24$) the median falls between the 12th and 13th value in the ordered list, and both the 12th and 13th values fall in the interval 15–16. For states east of the Mississippi ($n = 26$) the median falls between the 13th and 14th value in the ordered list, and both of these values also fall in the interval 15–16. From the histogram, cumulative frequencies for the two groups are shown in the table below.

Interval	West	East
$12 - 13$	1	2
$13 - 14$	$1 + 4 = 5$	$2 + 4 = 6$
$14 - 15$	$1 + 4 + 6 = 11$	$2 + 4 + 4 = 10$
$15 - 16$	$1 + 4 + 6 + 3 = 14$	$2 + 4 + 4 + 11 = 21$

   Thus, the median P-T ratio for both groups is at least 15 students per teacher and at most 16 students per teacher.

2. The shapes of the two histograms are different. The histogram for states that are west of the Mississippi River is unimodal and skewed to the right, whereas the histogram for states that are east of the Mississippi River is unimodal and nearly symmetric. As noted in part (a), the medians of the two distributions are about the same, between 15 and 16 for both distributions. The histograms also show that there is more variability in the P-T ratios for states that are west of the Mississippi River. Although the greatest and least values for each group are not known, the range can be approximated. The range for the west is at most 22 - 12 = 10, and the range for the east is at most 19 -12 = 7.

3. The medians of the two distributions are about the same, as determined in part (a). The distribution of P-T ratios for states that are west of the Mississippi River is skewed to the right, indicating that the mean will probably be higher than the median. The rough symmetry for the east group indicates that the mean will be close to the median. Thus, the mean for the west group will probably be greater than the mean for the east group.

Problem 7. AP 2011 №1 b,c

A professional sports team evaluates potential players for a certain position based on two main characteristics, speed and strength.

(a) *Topic: Normal Distribution*

(b) Strength is measured by the amount of weight lifted, with more weight indicating more desirable (greater) strength. From previous strength data for all players in this position, the amount of weight lifted has a mean of 310 pounds and a standard deviation of 25 pounds, as shown in the table below.

	Mean	Standard Deviation
Amount of weight lifted	310 pounds	25 pounds

Calculate and interpret the z-score for a player in this position who can lift a weight of 370 pounds.

(c) The characteristics of speed and strength are considered to be of equal importance to the team in selecting a player for the position. Based on the information about the means and standard deviations of the speed and strength data for all players and the measurements listed in the table below for Players A and B, which player should the team select if the team can only select one of the two players? Justify your answer.

	Player A	Player B
Time to run 40 yards	4.42 seconds	4.57 seconds
Amount of weight lifted	370 pounds	375 pounds

Solution:

(b) The z-score for a player who can lift a weight of 370 pounds is z-score $z = \frac{370-310}{25} = 2.4$. The z-score indicates that the amount of weight the player can lift is 2.4 standard deviations above the mean for all previous players in this position, so this player is quite a good strong.

(c) Because the two variables – time to run 40 yards and amount of weight lifted – are recorded on different scales, it is important not only to compare the players' values but also to take into account the standard deviations of the distributions of the variables. One reasonable way to do this is with z-scores.

The z-scores for the 40-yard running times are as follows:

Player A: $z = \frac{4.42-4.60}{0.15} = -1.2$

Player B: $z = \frac{4.57-4.60}{0.15} = -0.2$

The z-scores for the amount of weight lifted are as follows:

Player A: $z = \frac{370-310}{25} = 2.4$

Player B: $z = \frac{375-310}{25} = 2.6$

The z-scores indicate that both players are faster than average in the 40-yard running time and both are well above average in the amount of weight lifted. Player A is better in running time, and Player B is better in weight lifting. But the z-scores also indicate that the difference in their weight

lifting (a difference of 0.2 standard deviation) is quite small compared with the difference in their running times (a difference of 1.0 standard deviation). Therefore, Player A is the better choice, because Player A is much faster than Player B and only slightly less strong.

Problem 8. AP 2011 №2 c

The table below shows the political party registration by gender of all 500 registered voters in Franklin Township.

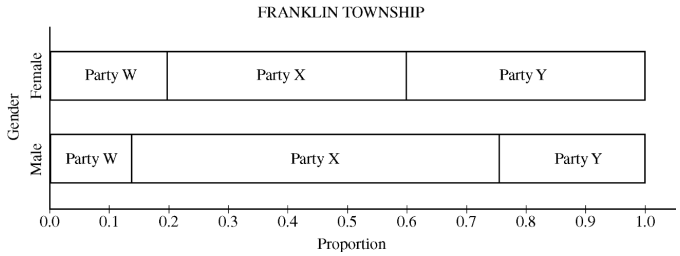
PARTY REGISTRATION – FRANKLIN TOWNSHIP

	Party W	Party X	Party Y	Total
Female	60	120	120	300
Male	28	124	48	200
Total	88	244	168	500

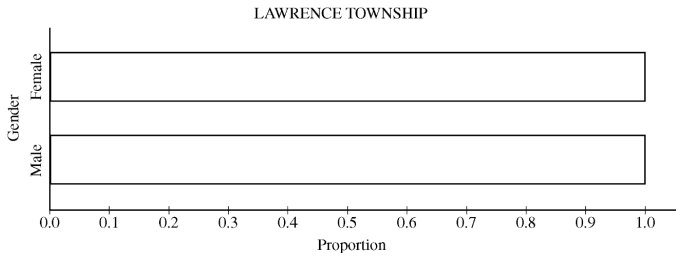
(a) *Topic: 2DRV*

(b) *Topic: 2DRV*

(c) One way to display the data in the table is to use a segmented bar graph. The following segmented bar graph, constructed from the data in the party registration – Franklin Township table, shows party-registration distributions for males and females in Franklin Township.



In Lawrence Township, the proportions of all registered voters for Parties W, X, and Y are the same as for Franklin Township, and party registration is independent of gender. Complete the graph below to show the distributions of party registration by gender in Lawrence Township.



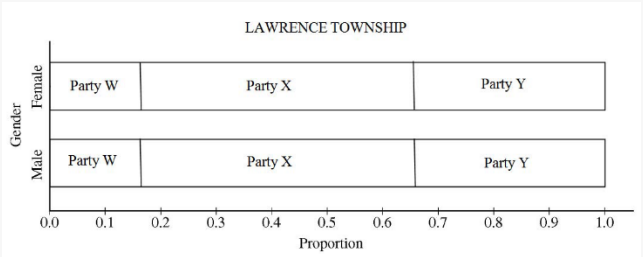
Solution:

Let W, X, Y denote belonging to parties W,X and Y correspondingly. Let M = Male

(c) The marginal proportions of voters registered for each of the three political parties (without regard to gender) are given below.

$P(W) = \frac{88}{500} = 0.176$, $P(X) = \frac{244}{500} = 0.488$, $P(Y) = \frac{168}{500} = 0.336$.

Because party registration is independent of gender in Lawrence Township, the proportions of males and females registered for each party must be identical to each other and also identical to the marginal proportion of voters registered for that party. Using the order Party W, Party X, and Party Y, the graph for Lawrence Township is displayed below.



Problem 9. AP 2010 №6

Hurricane damage amounts, in millions of dollars per acre, were estimated from insurance records for major hurricanes for the past three decades. A stratified random sample of five locations (based on categories of distance from the coast) was selected from each of three coastal regions in the southeastern United States. The three regions were Gulf Coast (Alabama, Louisiana, Mississippi), Florida, and Lower Atlantic (Georgia, South Carolina, North Carolina). Damage amounts in millions of dollars per acre, adjusted for inflation, are shown in the table below.

HURRICANE DAMAGE AMOUNTS IN MILLIONS OF DOLLARS PER ACRE

	Distance from Coast				
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles
Gulf Coast	24.7	21.0	12.0	7.3	1.7
Florida	35.1	31.7	20.7	6.4	3.0
Lower Atlantic	21.8	15.7	12.6	1.2	0.3

1. Sketch a graphical display that compares the hurricane damage amounts per acre for the three different coastal regions (Gulf Coast, Florida, and Lower Atlantic) and that also shows how the damage amounts vary with distance from the coast.

2. Describe differences and similarities in the hurricane damage amounts among the three regions. Because the distributions of hurricane damage amounts are often skewed, statisticians frequently use rank values to analyze such data.

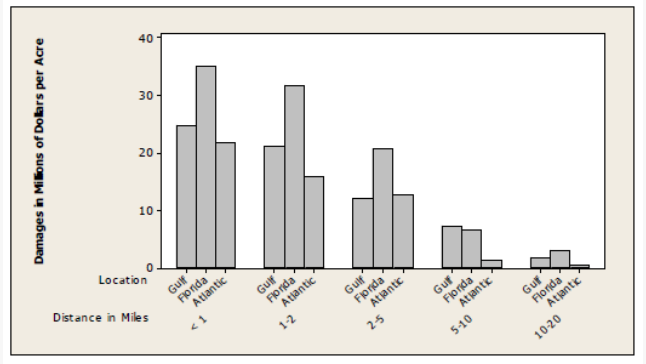
3. In the table below, the hurricane damage amounts have been replaced by the ranks 1, 2, or 3. For each of the distance categories, the highest damage amount is assigned a rank of 1 and the lowest damage amount is assigned a rank of 3. Determine the missing ranks for the 10-to-20-miles distance category and calculate the average rank for each of the three regions. Place the values in the table below.

ASSIGNED RANKS WITHIN DISTANCE CATEGORIES

	Distance from Coast					
	< 1 mile	1 to 2 miles	2 to 5 miles	5 to 10 miles	10 to 20 miles	Average Rank
Gulf Coast	2	2	3	1		
Florida	1	1	1	2		
Lower Atlantic	3	3	2	3		

Solution:

(a)



(b) In all three regions (Gulf Coast, Florida, Lower Atlantic) the hurricane damage amounts tend to decrease as distance from the coast increases. For almost all given distances from the coast, the Florida region has the largest damage amounts. Also, for any given distance, the Gulf Coast and Lower Atlantic regions have similar damage amounts but with the Lower Atlantic damage amounts generally smaller.

(c) For the "10 to 20 miles" distance category: The Florida region has the most damage (3.0 million dollars per acre) and so has rank 1. The region with the second-most damage is the Gulf Coast (1.7 million dollars), obtaining rank 2. The Lower Atlantic region has the least damage (0.3 million dollars) and so has rank 3. The last columns of the table should be filled in as follows:

	10 to 20 miles	Average Rank
Gulf Coast	2	2.0
Florida	1	1.2
Lower Atlantic	3	2.8

The average ranks are computed for: the five Gulf Coast damage ranks $\frac{2+2+3+1+2}{5} = 2.0$, the five Florida damage ranks $\frac{1+1+1+2+1}{5} = 1.2$ and the five Lower Atlantic damage ranks $\frac{3+3+2+3+3}{5} = 2.8$.

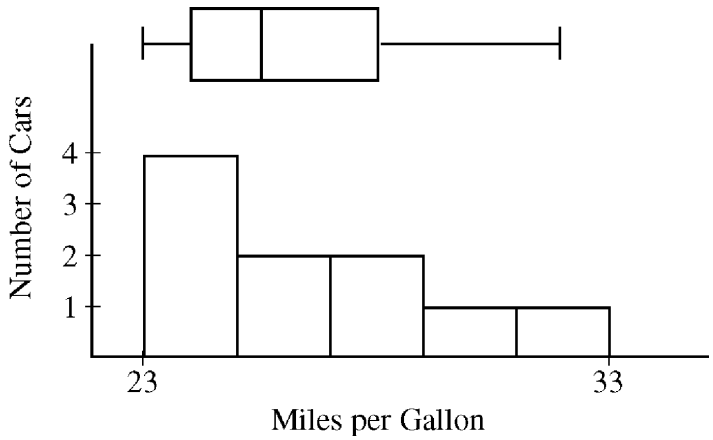
Problem 10. AP 2009 №2 a
A tire manufacturer designed a new tread pattern for its all-weather tires. Repeated tests were conducted on cars of approximately the same weight traveling at 60 miles per hour. The tests showed that the new tread pattern enables the cars to stop completely in an average distance of 125 feet with a standard deviation of 6.5 feet and that the stopping distances are approximately normally distributed.

(a) What is the 70th percentile of the distribution of stopping distances?

Solution:
Let X denote the stopping distance of a car with new tread tires where X is normally distributed with a mean of 125 feet and a standard deviation of 6.5 feet. The z-score corresponding to a cumulative probability of 70 percent is $z = 0.52$. Thus, the 70th percentile value can be computed as: $x = \mu + z\sigma = 125 + 0.52(6.5) = 128.4$ feet.

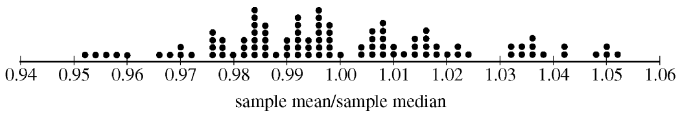
Problem 11. AP 2009 №6 b, c, d A consumer organization was concerned that an automobile manufacturer was misleading customers by overstating the average fuel efficiency (measured in miles per gallon, or mpg) of a particular car model. The model was advertised to get 27 mpg. To investigate, researchers selected a random sample of 10 cars of that model. Each car was then randomly assigned a different driver. Each car was driven for 5,000 miles, and the total fuel consumption was used to compute mpg for that car.

One condition for conducting a one-sample t-test in this situation is that the mpg measurements for the population of cars of this model should be normally distributed. However, the boxplot and histogram shown below indicate that the distribution of the 10 sample values is skewed to the right.



(b) One possible statistic that measures skewedness is the ratio $\frac{\text{sample mean}}{\text{sample median}}$. What values of that statistic (small, large, close to one) might indicate that the population distribution of mpg values is skewed to the right? Explain.

(c) Even though the mpg values in the sample were skewed to the right, it is still possible that the population distribution of mpg values is normally distributed and that the skewedness was due to sampling variability. To investigate, 100 samples, each of size 10, were taken from a normal distribution with the same mean and standard deviation as the original sample. For each of those 100 samples, the statistic $\frac{\text{sample mean}}{\text{sample median}}$ was calculated. A dotplot of the 100 simulated statistics is shown below.



In the original sample, the value of the statistic $\frac{\text{sample mean}}{\text{sample median}}$ was 1.03. Based on the value of 1.03 and the dotplot above, is it plausible that the original sample of 10 cars came from a normal population, or do the simulated results suggest the original population is really skewed to the right? Explain.

(d) The table below shows summary statistics for mpg measurements for the original sample of 10 cars.

Minimum	Q1	Median	Q2	Maximum
23	24	25.5	28	32

Choosing only from the summary statistics in the table, define a formula for a different statistic that measures skewedness. What values of that statistic might indicate that the distribution is skewed to the right? Explain.

Solution:

(b) If the distribution is right-skewed, one would expect the mean to be greater than the median. Therefore the ratio $\frac{\text{sample mean}}{\text{sample median}}$ should be large (at last greater than 1).

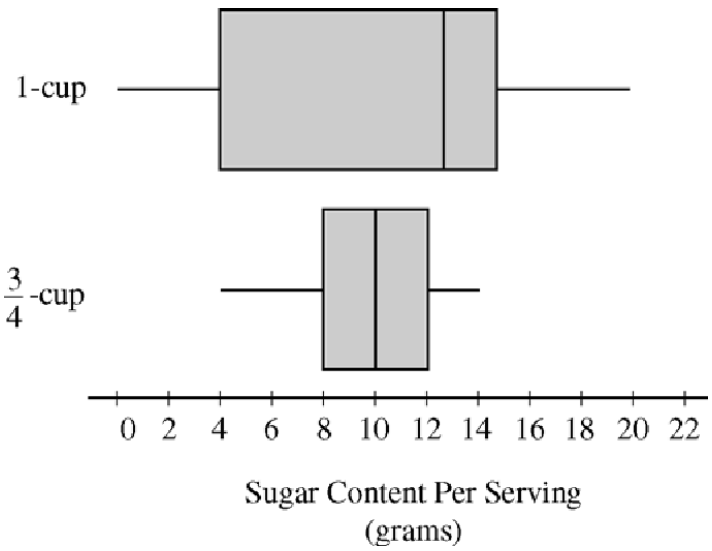
(c) Because we are testing for right-skewedness, the estimated p-value will be the proportion of the simulated statistics that are greater than or equal to the observed value of 1.03. The dotplot shows that 14 of the 100 values are more than 1.03. Because this simulated p-value (0.14) is larger than any reasonable significance level, we do not have convincing evidence that the original population is skewed to the right and conclude that it is plausible that the original sample came from a normal population.

(d) One possible statistic is $\frac{\text{maximum}-\text{median}}{\text{median}-\text{minimum}}$

If the distribution is right-skewed, one would expect the distance from the median to the maximum to be larger than the distance from the median to the minimum; thus the ratio should be greater than 1.

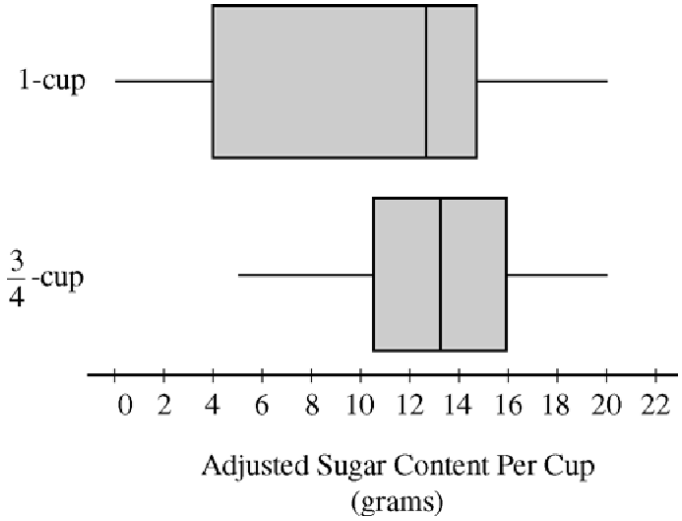
Problem 12. AP 2008 №1

To determine the amount of sugar in a typical serving of breakfast cereal, a student randomly selected 60 boxes of different types of cereal from the shelves of a large grocery store. The student noticed that the side panels of some of the cereal boxes showed sugar content based on one-cup servings, while others showed sugar content based on three-quarter-cup servings. Many of the cereal boxes with side panels that showed three-quarter-cup servings were ones that appealed to young children, and the student wondered whether there might be some difference in the sugar content of the cereals that showed different-size servings on their side panels. To investigate the question, the data were separated into two groups. One group consisted of 29 cereals that showed one-cup serving sizes; the other group consisted of 31 cereals that showed three-quarter-cup serving sizes. The box plots shown below display sugar content (in grams) per serving of cereals for each of the two serving sizes.



(a) Write a few sentences to compare the distributions of sugar content per serving for the two serving sizes of cereals.

After analyzing the boxplots on the preceding page, the student decided that instead of a comparison of sugar content per recommended serving, it might be more appropriate to compare sugar content for equal-size servings. To compare the amount of sugar in serving sizes of one cup each, the amount of sugar in each of the cereals showing three-quarter-cup servings on their side panels was multiplied by 4/3. The bottom boxplot shown below displays sugar content (in grams) per cup for those cereals that showed a serving size of three quarter-cup on their side panels.



(b) What new information about sugar content do the boxplots above provide?

(c) Based on the boxplots shown above on this page, how would you expect the mean amounts of sugar per cup to compare for the different recommended serving sizes? Explain.

Solution:

(a) The cereals that list a serving size of one cup have a median sugar amount larger than the median for the cereals that list a serving size of three-quarters of a cup. There is more variability (larger range and larger IQR) for the one-cup cereals. The shapes of the two distributions differ. The distribution of sugar content for three-quarter-cup cereals is reasonably symmetric: notice that the median is in the middle of the box. The distribution of sugar content for one-cup cereals is clearly skewed to the left (skewed toward the lower values): notice that the median is pulled to the right side of the central box closer to the third quartile.

(b) The distribution of sugar content in the cereals that list one-cup serving sizes remains the same as in part (a) because no transformations were applied to this distribution. There is a noticeable shift toward higher sugar content for the cereals that list three-quarter-cup servings after the transformation was applied to this distribution. The two types of cereals (one-cup and three-quarter-cup) now have similar medians, and the two distributions now show similar maximum values. In addition, the variability in the sugar content for cereals with a three-quarter-cup serving size increased by a factor of $4/3$ after the transformation was applied to the data in this distribution.

(c) Judging from the boxplots in part (b), we would expect the mean amounts of sugar per serving to be different. By the symmetry of the boxplot for the three-quarter-cup cereals, we would expect the mean and median to be similar. Because the boxplot for the one-cup cereals is skewed to the left, we

would expect the mean to be lower than the median. Thus, because both types of cereal have similar medians, we would expect the mean amount of sugar per cup for cereals with a one-cup serving size to be lower than the mean amount of sugar per cup for cereals with a three-quarter-cup serving size.

Problem 13. AP 2008 Form B №6 d

The nerves that supply sensation to the front portion of a person's foot run between the long bones of the foot.

Tight-fitting shoes can squeeze these nerves between the bones, causing pain when the nerves swell. This condition is called Morton's neuroma. Because most people have a dominant foot, muscular development is not the same in both feet. People who have Morton's neuroma may have the condition in only one foot or they may have it in both feet.

Investigators selected a random sample of 12 adult female patients with Morton's neuroma to study this disease further. The data below are measurements of nerve swelling as recorded by a physician. A value of 1.0 is considered "normal," and 2.0 is considered extreme swelling. The population distribution of the swelling measurements is approximately normal for adult females who have Morton's neuroma.

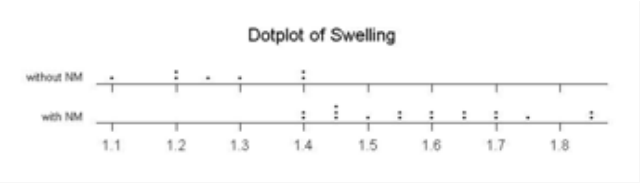
Dominant Foot	Swelling in Dominant Foot	Swelling in Nondominant Foot	Foot with Neuroma
Left	1.40	1.10	Left
Left	1.55	1.25	Left
Left	1.65	1.20	Left
Left	1.55	1.40	Both
Left	1.70	1.40	Left
Left	1.85	1.50	Both
Right	1.45	1.20	Right
Right	1.65	1.30	Right
Right	1.60	1.40	Right
Right	1.70	1.45	Both
Right	1.85	1.45	Both
Right	1.75	1.60	Both

(d) The nerve swelling measurement is used to indicate whether a foot has Morton's neuroma. Use the 24 measurements of nerve swelling to suggest a criterion for diagnosing Morton's neuroma. Justify your suggestion graphically.

Solution:

(d) There are more than one possible variants of the answer

Variant 1.



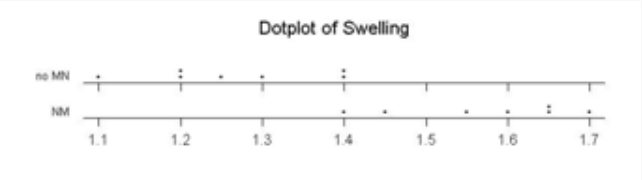
Separate the 24 swelling measurements into two groups—the 17 feet with MN and the 7 feet without MN.

Construct a plot that displays the two groups, such as stacked dotplots or a back-to-back stemplot. The plot

below suggests that a swelling measurement of about 1.4 or higher would be a reasonable criterion for

Morton's neuroma.

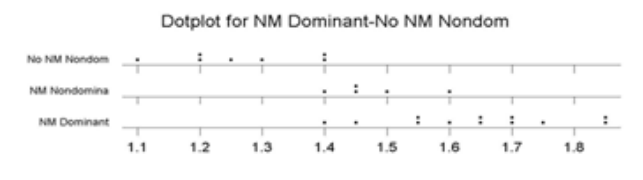
Variant 2.



Plot swelling measurements for only the seven individuals who do not have MN in both feet, plotting the

measurements for their MN feet and their non-MN feet. The plot below suggests that a swelling measurement of about 1.4 or higher would be a reasonable criterion for MN.

Variant 3.

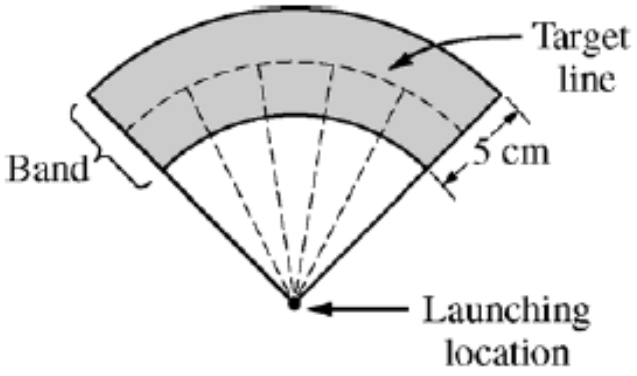


Plot the 12 measurements of MN in the dominant foot, the 5 measurements of MN in the nondominant foot, and the 7 measurements of no MN in the non-dominant foot. (There are no individuals in the sample who do not have MN in the dominant foot.) The plot below suggests that a swelling measurement of about 1.4 or higher would be a reasonable criterion for MN.

Problem 14. AP 2006 №1

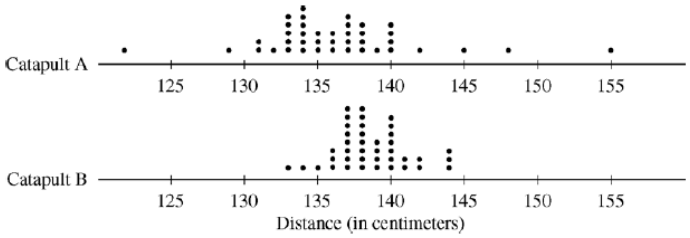
Two parents have each built a toy catapult for use in game at an elementary school fair. To play the game the students will attempt to launch Ping-Pong balls from the

catapults so that the balls land within a 5-centimeter band. A target line will be drawn through the middle of the band, as shown in the figure below. All points on the target lie are equidistant from the launching location.



If a ball lands within the shaded band, the student will win a prize.

The parents have constructed the two catapults according to slightly different plans. They want to test these catapults. Under identical conditions, the parents launch 40 Ping-Pong balls from each catapult and measure the distance that the ball travels before landing. Distances to the nearest centimeter are graphed in the dotplots below.



(a) Comment on any similarities and any differences in the two distributions of distances traveled by balls launched from catapult A and catapult B.

Solution:

1. Both distributions of distances are rather symmetric with high concentration in center and lighter tails.

   However the median of the distribution for catapult A($\approx$ 136 cm) is lower than the median of the distribution for catapult B ($\approx$ 138 cm). Also there is much more variability in the distances traveled by balls launched with catapult A than with catapult B. For catapult A there exist some distances to be called potential outliers while there are no such for catapult B.

Problem 15. AP 2005 №1
The goal of a nutritional study was to compare the caloric intake of adolescents (молодой человек, юноша) living in rural areas of the United States with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

The back-to-back box stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

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

Stem: tens
Leaf: ones

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of the daily caloric intake of ninth-grade students in the urban high school.

Solution:

1. The center of the distribution of daily caloric intake of ninth-grade students in the rural high school is higher than that of the urban school. Mean 40.45 c/kg and median 41 c/kg compared to mean 32.6c/kg and median 32 c/kg. There is also more variability in daily caloric intake for students in the rural school than in the urban school, namely range = 19, st.dev. = 6.04, IQR = 10 versus range = 16, st.dev. = 4.67, IQR = 7 correspondingly. The shapes of the two distributions of caloric intake are also different. The one for rural school is more uniform, symmetric, while for urban school appears to be skewed to the larger values (skewed to the right). estimate of the daily calories intake.

Problem 16. AP 2005 №2 c,d
Let the random variable X represent the number of telephone lines in use by the technical support center of a software manufacturer at noon each day.

X	0	1	2	3	4	5
$P(X)$	0.35	0.2	0.15	0.15	0.1	0.05

(c) The median of a random variable is defined as any value x such that $P(X \leqslant x) \geqslant 0.5$ and $P(X \geqslant x) \geqslant 0.5$. For the probability distribution shown in the table above, determine the median of X.

(d) In a sentence or two, comment on the relationship between the mean and the median relative to the shape of this distribution.

Solution:

(c) The median of X is equal to med $= 1$ because $P(X \leqslant 1) = 0.55 \geqslant 0.5 and P(X \geqslant 1) = 0.65 \geqslant 0.5$

(d) This distribution is clearly skewed to the right (towards the large values) since the mean 1.6 is greater than the median 1, which is typical for right-skewed distributions.

Problem 17. AP 2005 №6 b,c

Lead, found in some plants, is a neurotoxin that can be especially harmful to the developing brain and nervous system of children. Children frequently put their hands in their mouth after touching painted surfaces, and this is the most common type of exposure to lead.

A study was conducted to investigate whether there were differences in children's exposure to lead between suburban day-care centers and urban day-care centers in one large city. For this study, researchers used a random sample of 20 children in suburban day-care centers. Ten of these 20 children were randomly selected to play outside; the remaining 10 children played inside. All children had their hands wiped clean before beginning their assigned one-hour play period either outside or inside. After the play period ended, the amount if lead in micrograms (mcg) on each child's dominant hand was recorded.

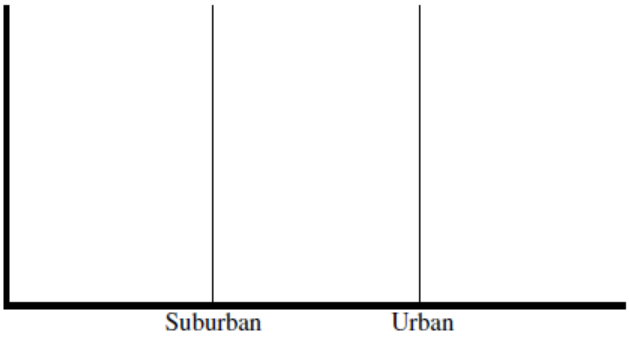
The mean amount of lead on the dominant hand for the children playing inside was 3.75 mcg, and the mean amount of lead for the children playing inside was 5.65 mcg. A 95 percent confidence interval for the difference in the mean amount of lead after one hour inside versus one hour outside was calculated to be (-2.46, -1.34).

A random sample of 18 children in urban day-care centers in the same large city was selected. For this sample, the same process was used, including randomly assigning children to play inside or outside. The data for the amount (in mcg) of lead in each child's dominant hand are shown in the table below.

Urban day-care centers									
Inside	6	5	4	4	4.5	5	4.5	3	5
Outside	15	25	18	14	20	13	11	22	20

(b) On the figure below,

- Using the vertical axis for the mean amount of lead, plot the mean for the amounts of lead on the dominant hand of children who played inside at the suburban day-care center and then plot the mean for the amounts of lead on the dominant hand of children who played inside at the urban day-care center.
- Connect these two points with a line segment.
- Plot the two means (suburban and urban) for the children who played outside at the two types of day-care centers.
- Connect these two points with a second line segment.



(c) From the study, what conclusions can be drawn about the impact of setting (inside, outside), environment (suburban, urban), and the relationship between the two on the amount of lead on the dominant hand of children after play in this city? Justify your answer.

Solution:

(b)

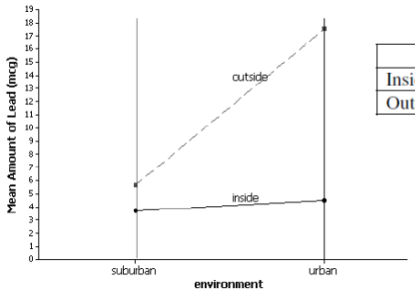


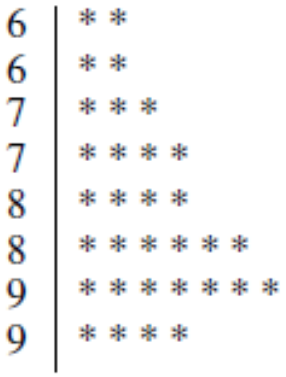
Table of Means

	Suburban	Urban
Inside	3.75	4.56
Outside	5.65	17.56

(c)
- *setting*: For both environments the mean amount of lead on the dominant hands of children who play outside is higher than mean amount of lead on the dominant hands of children who play inside. The graph clearly shows that the line connecting the two 'outside' means for the different environments is above the line that connects the "inside" means.

- *environment*: For both settings the amount of lead on the dominant hand of urban children is higher on average than the amount of lead on the dominant hand of suburban children. This can be justified by comparing means or interpreting the graph: both lines slant upward to the right, which indicates an increase from suburban to urban both for children who played inside and for children who played outside.
- *relationship*: Whether the children play inside or outside makes a bigger difference in the urban environment than in the suburban environment. The graph shows that the means for the urban environment are much farther apart than the means for the suburban environment.

Problem 18. AP 2005 Form B №1
The graph below displays the scores of 32 students on a recent exam. Scores on this exam ranged from 64 to 95 points.



(a) Describe the shape of this distribution.

(b) In order to motivate her students, the instructor of the class wants to report that, overall, the class's performance on the exam was high. Which summary statistic, the mea or the median, should the instructor use to report that overall exam performance was high? Explain.

(c) The midrange is defined as $\frac{\text{maximum}+\text{minimum}}{2}$. Compute this value using the data on the preceding page. Is the midrange considered a measure of center or a measure of spread? Explain.

Solution:

(a) The shape of this distribution is skewed towards the lower values (to the left)

(b) The instructor should use the median in order to achieve her goal, since the distribution is skewed to the lower values. Then, the mean will be pulled in that direction, and will be lower than the median.

(c) Midrange $= \frac{64+95}{2} = 79.5$ Midrange can be considered as a measure of center. The maximum provides information about the upper extreme value. The

minimum provides information about the lower extreme value. Averaging these two values, we are creating a value that reflects the halfway point between the two extremes. Such statistic is considered as a measure of center.

Problem 19. AP 2004 №1

A consumer advocate conducted a test of two popular gasoline additives, A and B. There are claims that the use of either of these additives will increase gasoline mileage in cars. A random sample of 30 cars was selected. Each car was filled with gasoline and the cars were run under the same driving conditions until the gas tanks were empty. The distance traveled was recorded for each car.

Additive A was randomly assigned to 15 of the cars and additive B was randomly assigned to the other 15 cars. The gas tank of each car was filled with gasoline and the assigned additive. The cars were again run under the same driving conditions until the tanks were empty. The distance traveled was recorded and the difference in the distance with the additive minus the distance without the additive for each car was calculated.

The following table summarizes the calculated differences. Note that negative values indicate less distance was traveled with the additive than without the additive.

Additive	Values below Q1	Q1	Median	Q3	Values above Q3
A	-10, -8, -2	1	3	4	5, 7, 9
B	-5, -3, -3	-2	1	25	35, 37, 40

(a) Display parallel boxplots (showing outliers, if any) of the differences of the two additives.

(b) Two ways that the effectiveness of a gasoline additive can be evaluated are by looking at either

- the proportion of cars that have increased gas mileage when the additive is used in those cars, or
- the mean increase in gas mileage when the additive is used in those cars

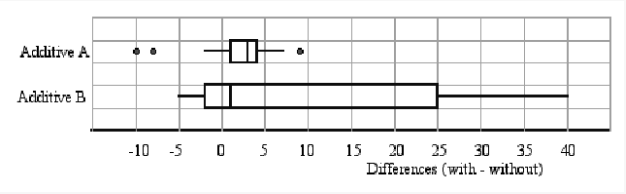
(i) Which additive, A or B, would you recommend if the goal is to increase gas mileage in the highest proportion of cars? Explain your choice.

(ii) Which additive, A or B, would you recommend if the goal is to have the highest mean increase in gas mileage? Explain your choice.

Solution:

1.

	Additive A	Additive B
IQR	$4-1=3$	$25-(-2)=27$
1.5*IQR	4.5	40.5
$Q_1-1.5\times$IQR	$1-4.5=-3.5$	$-2-40.5=-42.5$
$Q_3+1.5\times$IQR	$4+4.5=8.5$	$25+40.5=65.5$
	3 outliers	no outliers



2. (i) Additive A is better at increasing the mileage in the greatest number of cars. The mileage increased for at least seventy-five percent of the cars with additive A, whereas the mileage decreased for more than twenty-five percent of the cars with additive B.

   (ii) Additive B appears to produce a higher mean mileage gain than additive A. The boxplot for additive B clearly shows that the distribution of differences is skewed to the right, which will pull the average towards the larger values. The mean difference for additive B will be substantially greater than the median of 1. On the other hand, the distribution of differences for additive A has much less variability, as seen by comparing the lengths of the two boxes, and appears to be skewed to the left. The mean difference for additive A will be less than the median of 3.

Problem 20. AP 2002 Form B №5

At a school field day, 50 students and 50 faculty members each completed an obstacle course. Descriptive statistics for the completion times (in minutes) for the two groups are shown below.

	Students	Faculty Members
Mean	9.9	12.09
Median	9.25	11
Minimum	3.75	4.5
Maximum	16.5	25
Lower quartile	6.75	8.75
Upper quartile	13.75	15.75

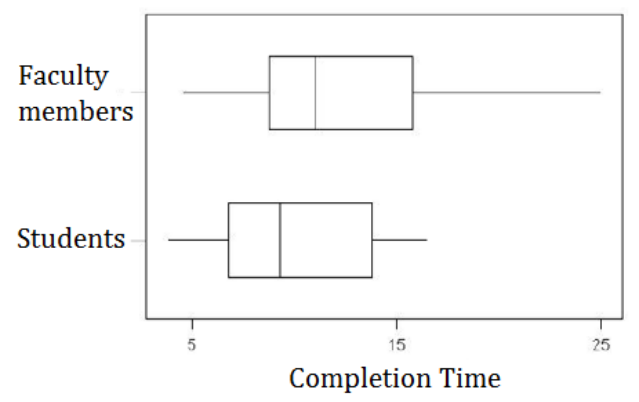
(a) Use the same scale to draw boxplots for the completion times for students and for faculty members.

(b) Write a few sentences comparing the variability of the two distributions.

(c) You have been asked to report on this event for the school newspaper. Write a few sentences describing students and faculty performances in this competition for the paper.

Solution:

(a)



The label of time is essential.

(b) The IQR of both distributions are similar, although the range for faculty members' completion times is much larger than that of students. Thus the spread is similar in the middle 50 percent of data, but the smallest 25 percent and largest 25 percent are more spread for the faculty members than for students.

(c) Although some faculty members completed the obstacle course quickly, generally students had shorter completion time. The students' completion time, ranging from 3.75 min to 16.5 min, was more consistent than the faculty time, which ranged from 4.5 min to 25 min. Both distributions tend to be skewed to the right, namely many students and faculty finished relatively quickly, but the other slower half of each group was spread.

Problem 21. AP 2001 №1
The summary statistics for the number of inches of rainfall in Los Angeles for 117 years, beginning in 1877, are shown below.

n	mean	median	trmean	stdev	se mean
117	14.941	13.070	14.416	6.747	0.624
min	max	Q1	Q3		
4.850	38.180	9.680	19.250		

(a) Describe a procedure that uses these summary statistics to determine whether there are outliers.

(b) Are there outliers in these data? Justify your answer based on the procedure that you described in part (a).

(c) The news media reported that in a particular year, there were only 10 inches of rainfall. Use the information provided to comment on this reported statement.

Solution:

(a) An outlier considered to be any value that is $1.5 \cdot IQR$ below the lower quartile or $1.5 \cdot IQR$ above the upper quartile. Thus, using this summary statistics, if:

- $\min < Q1 - 1.5 \cdot IQR \Rightarrow$ there is at least one outlier on the low side of these data
- $\max > Q3 + 1.5 \cdot IQR \Rightarrow$ there is at least one outlier on the high side of these data

Alternative answer: An outlier is any observation that is more than 2(or 3) standard deviations away from the mean.

(b) $IQR = Q3 - Q1 = 19.250 - 9.680 = 9.57$,
$1.5 \cdot IQR = 1.5 \cdot 9.57 = 14.355$
Bound for outlier on the low side: $Q1 - 1.5 \cdot IQR = 9.680 - 14.355 = \mathbf{-4.675}$
Bound for outlier on the high side: $Q3 + 1.5 \cdot IQR = 19.25 + 14.355 = \mathbf{33.605}$
There is at least one outlier on the high side since the maximum value 38.180, is greater than the upper bound 33.605.

(c) Lower quartile $Q1 = 9.68$ inches. That means more than 25% of the years had less than 10 inches of rain. Therefore, 10 inches of rain is not an unusual value.

Problem 22. AP 2000 №3.

Five hundred randomly selected middle aged men and five hundred randomly selected young adult men were rated on a scale from 1 to 10 on their physical flexibility, with 10 being the most flexible. Their rating appear in the frequency table below. For example, 17 middle-aged men has a flexibility rating of 1.

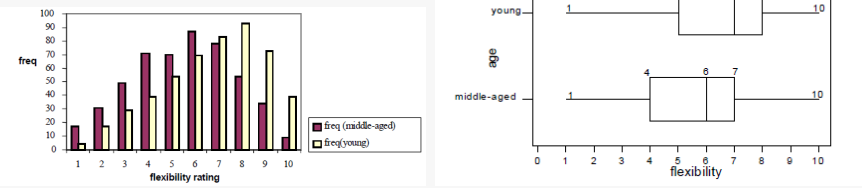
Physical Flexibility Rating	Frequency of Middle-Aged Men	Frequency of Young Adult Men
1	17	4
2	31	17
3	49	29
4	71	39
5	70	54
6	87	69
7	78	83
8	54	93
9	34	73
10	9	39

(a) Display the data graphically so that the flexibility of middle-aged men and young adult men can be easily compared.

(b) Based on an examination of your graphical display, write a few sentences comparing the flexibility of middle-aged men with the flexibility of young adult men.

Solution:

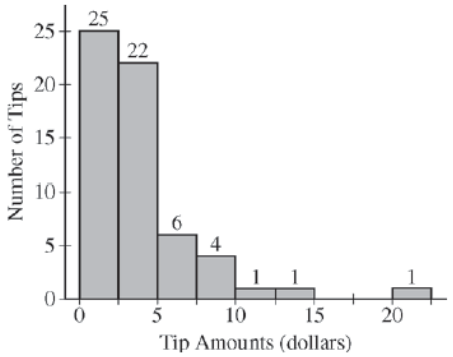
(a)



(b) The distribution of flexibility raiting for middle-aged men is centered around 5.5 while the distribution for young adult men is centered around 6.5, higher than the middle-aged men. There is quite a bit variability in both distributions. The distribution for middle-aged men is quite symmetric, while for young adult men is skewed to the left. Thus, there were more young men with high flexibility and fewer with low than for middle aged men.

Practice AP problems

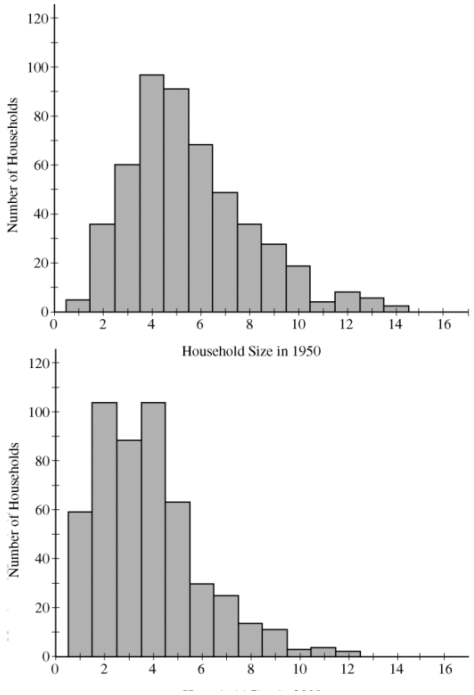
Problem 1. AP 2016 №1
Robin works as a server in a small restaurant, where she can earn a tip (extra money) from each customer she serves. The histogram below shows the distribution of her 60 tip amounts for one day of work.



(a) Write a few sentences to describe the distribution of tip amounts for the day shown.

(b) One of the tip amounts was \$8. If the \$8 tip had been \$18, what effect would the increase have had on the following statistics? Justify your answer. The mean. The median.

Problem 2. AP 2012 №3
Independent random samples of 500 households were taken from a large metropolitan area in the United States for the years 1950 and 2000. Histograms of household size (number of people in a household) for the years are shown below.

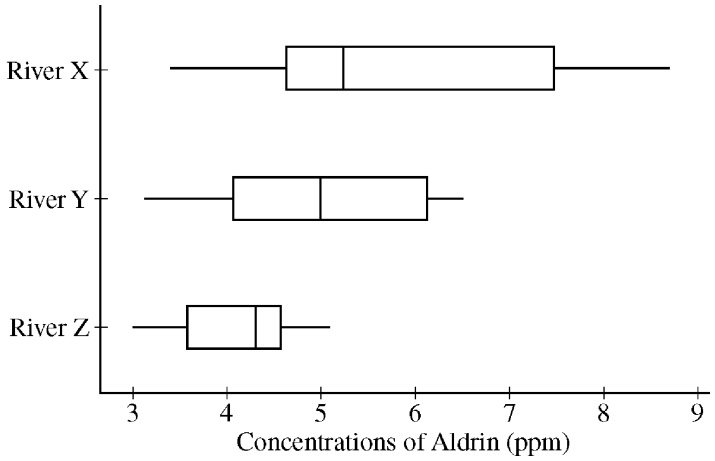


(a) Compare the distributions of household size in the metropolitan area for the years 1950 and 2000.

Problem 3. AP 2010 Form B №1

As a part of the United States Department of Agriculture's Super Dump cleanup efforts in the early 1990s, various sites in the country were targeted for cleanup. Three of the targeted sites – River X, River Y, and River Z – had become contaminated with pesticides because they were located near abandoned pesticide dump sites. Measurements of the concentration of aldrin (a commonly used pesticide) were taken at twenty randomly selected locations in each river near the dump sites.

The boxplots shown below display the five-number summaries for the concentrations, in parts per million (ppm) of aldrin, for the twenty locations that were sampled in each of the three rivers.



(a) Compare the distributions of the concentration of aldrin among the three rivers.

(b) The twenty concentrations of aldrin for River X are given below.

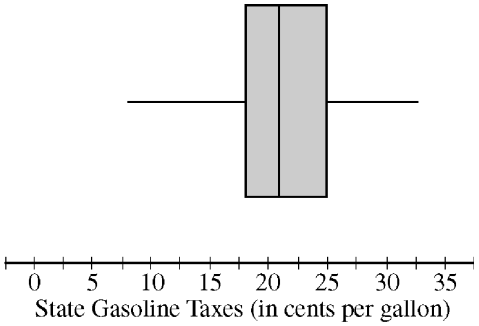
3.4 4.0 5.6 3.7 8.0 5.5 5.3 4.2 4.3 7.3
8.6 5.1 8.7 4.6 7.5 5.3 8.2 4.7 4.8 4.6

Construct a stemplot that displays the concentrations of aldrin for River X.

(c) Describe a characteristic of the distribution of aldrin concentrations in River X that can be seen in the stemplot but cannot be seen in the boxplot.

Problem 4. AP 2009 Form B №1

As gasoline prices have increased in recent years, many drivers have expressed concern about the taxes they pay on gasoline for their cars. In the United States, gasoline taxes are imposed by both the federal government and by individual states. The boxplot below shows the distribution of the state gasoline taxes, in cents per gallon, for all 50 states on January 1, 2006.



(a) Based on the boxplot, what are the approximate values of the median and the interquartile range of the distribution of state gasoline taxes, in cents per gallon? Mark and label the boxplot to indicate how you found the approximated values.

(b) The federal tax imposed on gasoline was 18.4 cents per gallon at the time the state taxes were in effect. The federal gasoline tax was added to the state gasoline tax for each state to create a new distribution of combined gasoline taxes. What are approximate values, in cents per gallon, of the median and Interquartile range of the new distribution of combined gasoline taxes? Justify your answer.

Problem 5. AP 2008 Form B №1 a

A certain state's education commissioner released a new report card for all the public schools in that state. This report card provides a new tool for comparing schools across the state. One of the key measures that can be computed from the report card is the student-to-teacher ratio, which is the number of students enrolled in a given school divided by the number of teachers at that school.

The data below give the student-to-teacher ratio at the 10 schools with the highest proportion of students meeting the state reading standards in the third grade and at the 10 schools with the lowest proportion of students meeting the state reading standards in the third grade.

Ratios in the 10 Schools with Highest Proportion of Students Meeting Standards

7	21	18	22	9	16	12	17	17	16

Ratios in the 10 Schools with Lowest Proportion of Students Meeting Standards

14	16	18	20	12	14	16	12	20	19

(a) Display a dotplot for each group to compare the distribution of student-to-teacher ratios in the top 10 schools with the distribution in the bottom 10 schools. Comment on the similarities and differences between the two distributions.

Problem 6. AP 2007 Form B №1

The Better Business Council of a large city has concluded that students in the city's schools are not learning enough about economics to function in the modern world. These findings were based on test results from a random sample of 20 twelfth-grade students who completed a 46-question multiple-choice test on basic economic concepts. The data set below shows the number of questions that each of the 20 students in the sample answered correctly.

12	16	18	17	18	33	41	44	38	35
19	36	19	13	43	8	16	14	10	9

(a) Display these data in a stemplot.

(b) Use your stemplot from part (a) to describe the main features of this score distribution.

(c) Why would it be misleading to report only a measure of center for this score distribution?

Problem 7. AP 2004 Form B №5 a

A researcher thins that modern Thai dogs may be descendants of golden jackals. A random sample of 16 animals was collected from each of the two populations. The length (in mm) of the mandible (jawbone) was measured for each animal. The lower quartile, median, and upper quartile for each sample are shown in the table below, along with all values below the lower quartile and all value above the upper quartile.

Sample	Values below Q1	Q1	Median	Q3	Values above Q3
Modern Thai dog	114, 116, 116, 120	121	125	128	129, 130, 130, 132
Golden jackal	104, 104, 105, 106	107	108	112	114, 122, 124, 125

(a) Display parallel boxplots of mandible lengths (showing outliers, if any) for the modern Thai dogs and the golden jackals. Based on the boxplots, write a few sentences comparing the distributions of mandible length for the two types of dogs.

Problem 8. AP 2003 №1

Since Hill Valley High School eliminates the use of bells between classes, teachers had noticed that more students seem to be arriving to class a few minutes late. One teacher decided to collect data to determine whether the students' and teachers' watches are displaying the correct time. At exactly 12:00 noon the teacher asked 9 randomly selected students and 9 randomly selected teachers to record the time on their watches to the nearest half minute. The ordered data showing minutes after 12:00 as positive values as minutes before 12:00 as negative values are shown in the table below.

Students	-4.5	-3.0	-0.5	0	0	0.5	0.5	1.5	5.0
Teachers	-2.0	-1.5	-1.5	-1.0	-1.0	-0.5	0	0	0.5

(a) Construct parallel boxplots using these data.

(b) Based on the boxplots in part (a), which of the two groups, students or teachers, tends to have watch tmes that are closer to the true time? Explain your choice.

Problem 9. AP 2001 №6 a

The statistics department at a large university is trying to determine if it is possible to predict whether an applicant will successfully complete the Ph.D. program or will leave before completing the program. The department is considering whether GPA (grade point average) in undergraduate statistics and mathematics courses (a measure of performance) and mean number of credit hours per semester (a measure of workload) would be helpful measures. To gather data, a random sample of 20 entering students from the 5 years is taken. The data are given below.

Successfully completed Ph.D. program

Student	A	B	C	D	E	F	G	H	I	J	K	L	M
GPA	3.8	3.5	4.0	3.9	2.9	3.5	3.5	4.0	3.9	3.0	3.4	3.7	3.6
Credit hours	12.7	13.1	12.5	13.0	15.0	14.7	14.5	12.0	13.1	15.3	14.6	12.5	14.0

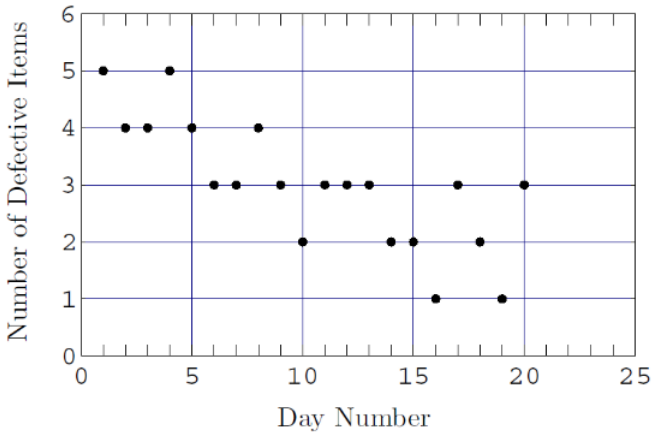
Did not complete Ph.D. program

Student	N	O	P	Q	R	S	T
GPA	3.6	2.9	3.1	3.5	3.9	3.6	3.3
Credit hours	11.1	14.5	14.0	10.9	11.5	12.1	12.0

(a) Use an appropriate graphical display to compare the GPA's for the two groups. Write a few sentences commenting on your display.

Problem 10. AP 1998 №2

A plot of the number of defective items produced during 20 consecutive days at a factory is shown below.



(a) Draw a histogram that shows the frequencies of the number of defective items.

(b) Give one fact that is obvious from the histogram but is not obvious from the scatterplot.

(c) Give one fact that is obvious from the scatterplot but is not obvious from the histogram.

Answers

Problem 1.

(a) Median between 2.5 and 5. Right-skewed. Exists a gap, possibly an outlier. Describe the range and concentration between 0 and 5.

(b) Increase by 17 cents ($10/60). Not change, both values are above median.

Problem 2.

(a) Larger in 1950. Compare the proportion of small and large households. Compare the median (from graph 5 in 1950 vs 3 or 4 in 2000). Compare variability (does not differ much). Both are skewed to the right.

Problem 3.

(a) Compare medians (X-highest, Z-lowest), compare variability(X-the most), compare shapes (skewed to the right X, symmetric Y, slightly skewed to the left Z.

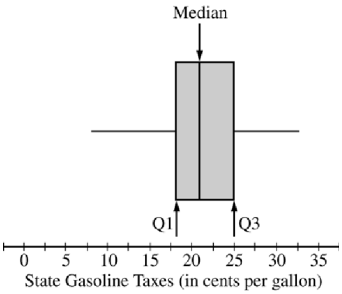
(b)

3	47
4	0236678
5	13356
6	
7	35
8	0267

(c) a gap

Problem 4.

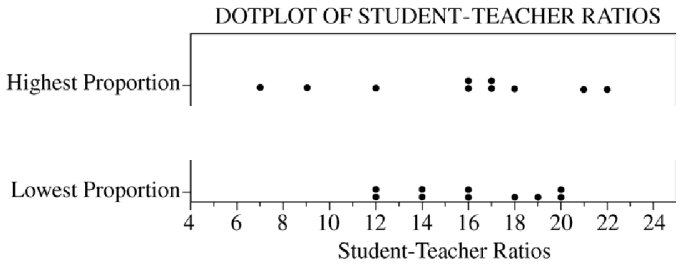
(a) Med = 21 cents per gallon, IQR = 7s cents per gallon.



(b) Med = 39.4 cents per gallon, IQR = 7 cents per gallon (the same).

Problem 5.

(a)



Similarity: The two are centered in approximately the same place. Difference: the distribution for the schools with the lowest proportions is less variable.

Problem 6.

(a)

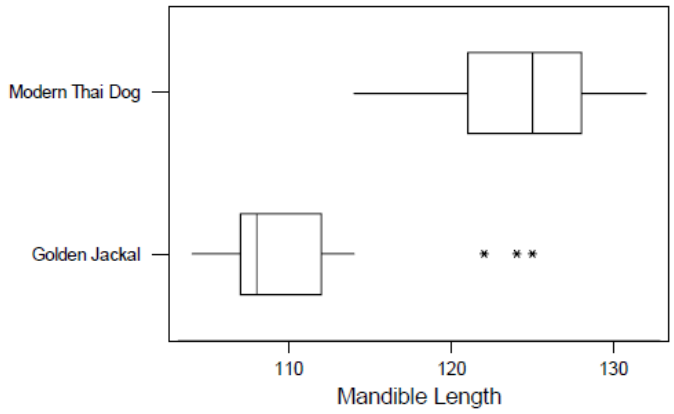
0| 89
1| 02346678899
2|
3| 3568
4| 134

(b) Scores cluster in two groups.

(c) Will be misleading. Center might fall between two clusters where there is no data.

Problem 7.

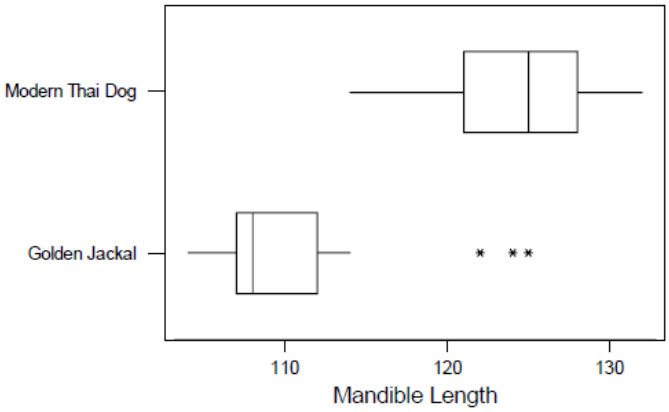
(a)



Not similar. For Thai dogs is approximately symmetric, typical value around 125. For Golden jackals skewed to the right, typical value around 108 (much smaller), with outliers. The variability does not differ too much.

Problem 8.

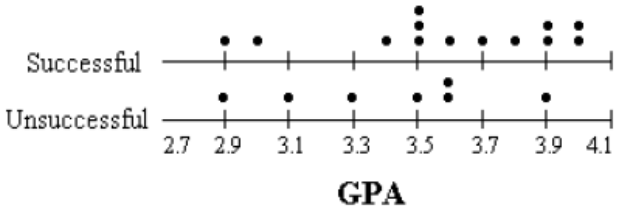
(a)



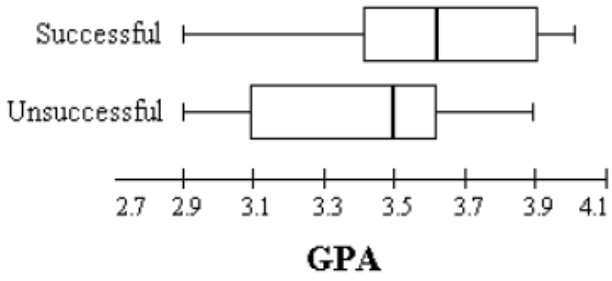
(b) The teacher's. Slow, but less variable.

Problem 9.

(a)



or



Although minimum is the same, there are more GPAs at high part for successful group.

Problem 10.

(a) histogram

(b) shape

(c) relationship