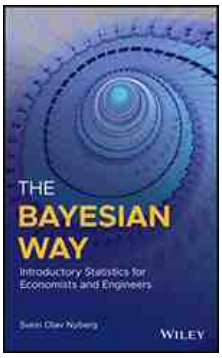


The Bayesian Way: Introductory Statistics for Economists and Engineers

by Svein Olav Nyberg

Chapter 2: Data

Overview

By *data* we mean a collection of a given type of values. The values are most commonly numbers, but can be anything we have received in response to our queries or measurements. Non-numerical values are called *categorical* data, which simply means information about membership of a category. One example of this is if our query is about preferences for a political election; the data would then be the names of a political party like *Democrat*, *Republican*, *Libertarian*, or *Green* in the USA, but also the categories *Don't know*, *Others*, and *Blank*. With categorical data, numbers come into play only when we are counting the number of hits in the different categories. This is as opposed to *numerical data*, which is the most common form of data, where the data are themselves numbers. Examples of such data are the *times* for a 60 yards dash, where the data are the times, and not the runners themselves or their names. Or the waist *circumference* of diabetic teenagers, where again the data are simply the number of inches in each measurement.

The *population* is the total of all possible *values* – including the ones that are not measured. We have two models of this: the rather concrete *urn* model, and the more abstract *process* model. We start with the urn model.

The urn model of a population is a finite set, an "urn" that contains little notes with values written on them. In an election, the population is the political preferences of the voters, and not the voters themselves. In an urn model of the population of the 2016 US election, the population would be the 139 million individual votes, and have values "Green", "Libertarian", "Republican", or "Democrat". So if, for instance, there were 4 042 291 votes for the libertarian party, the population contains 4 042 291 values "Libertarian". If you are looking at diabetic teenagers' waist circumferences, the population is the total set of waist measurements that could have been collected. So the population might contain a million "36.0 inches", and none of "20.0 inches". What matters to us are the values, and how many there are of each.

The *process* model of a population differs from the urn model in the same way that dice differ from a deck of cards. We abstract away the number of instances of each value, and look instead at the *proportions* for each value. For the US election mentioned above, four million then becomes 3.2%. For the value "17" on a D_{32} die, the proportion 3.125% is all we have, as there is no fixed number of dice tosses. When we later in this book will be talking about *sampling from a probability distribution*, we are referring to the process model.

The *sample* consists of the data we have actually collected. In an urn model, we justify sampling by appealing to cost: the sample will usually be a lot smaller than the population, as illustrated in Figure 2.1. So if we can draw sufficiently reliable conclusions by sampling a thousand values rather than doing an exhaustive measurement of several millions, then we should be sampling. In a presidential election, polls are often conducted by asking a few thousand voters for their preference. The pollsters then draw a fairly reliable conclusion about the political preferences of the entire population.

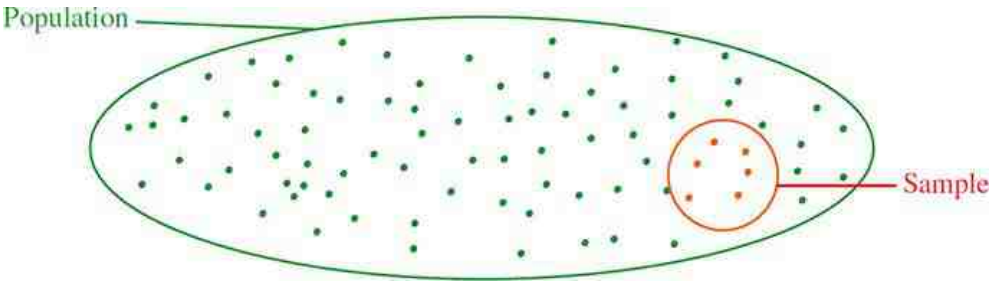
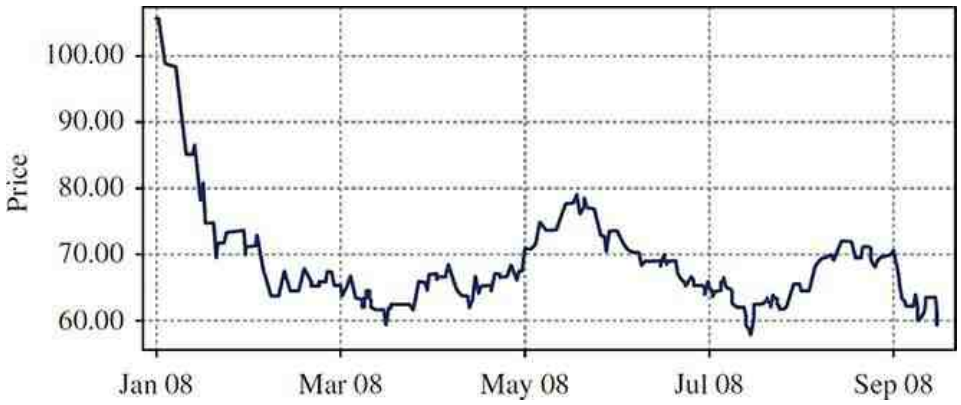


Figure 2.1: *Population* and *sample*

We will need to *index* our data. The most common way of indexing is enumeration, x_1, x_2, …, x_n, but other indexes like *time* and *location* might at times be more expedient. If you are looking at stock prices, like the ones in Figure 2.2 from the Oslo Stock Exchange (OSE), then it is better to write the price of an Orkla stock, at 12:00 on the 16th of September 2008, as $x_{2008.09.16.12:00}$, than to enumerate it as x_{3127} if it was your 3127th observation. But if no special factors come into play, enumeration x_1, x_2, …, x_n is the default choice.



Source: OSE 2008

Figure 2.2: Orkla stock price

2.1 Tables and Diagrams

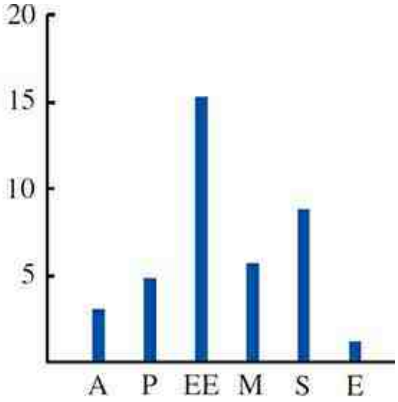
We often compress our observations into groups of equal value, noting only the number of observations for the value. We express this in a frequency table, counting how many there are of each kind, and a bar chart.

Example 2.1.1

Nathan counts the different books on Suzie's science bookshelf. He makes the frequency table and bar chart in Figure 2.3 from his measurements.

Value v_k (subject)	Frequency a_k
A: Astronomy	3
P: Physics	5
EE: Electrical Engineering	16
M: Mathematics	6
S: Statistics	9
E: Economics	1
Sum, n	40

(a) Table.



(b) Bar chart.

Figure 2.3: The books on Suzie's bookshelf

We have two basic types of data: numerical data and categorical data. When the data values are numbers, the horizontal axis becomes a value axis, whereas the vertical axis marks (relative) frequency.

Example 2.1.2

We asked 150 households how many TVs they owned. Our data are collected in Figure 2.4.

Value v_k (TVs)	Frequency a_k
1	33
2	39
3	42
4	36
Sum	150

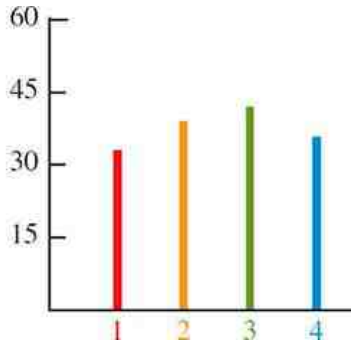


Figure 2.4: Number of TVs in a household

We are frequently more interested in the relative frequencies (proportions) $p_k = a_k/\Sigma a_k$ than in the absolute frequencies a_k themselves. When polling agencies report probable voter distribution for the next election, most of us are more interested in hearing that Jill Stein got 1% of the polled votes than in knowing that exactly 18 of the 1800 respondents said they would vote for Stein.

Example 2.1.3

(Continuation of Example 2.1.2) We find the proportions of how many households own how many TVs by normalizing the frequency table. That is, by dividing the category frequency by the total frequency, as shown in Figure 2.5.

Value v_k (TVs)	Relative frequency p_k
1	33/150 = 0.22 = 22%
2	39/150 = 0.26 = 26%
3	42/150 = 0.28 = 28%
4	36/150 = 0.24 = 24%
Sum	1

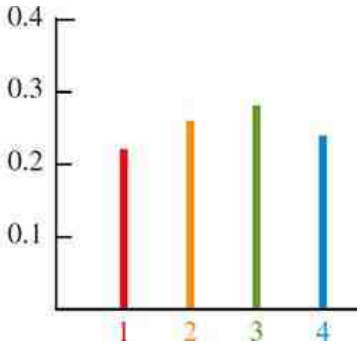
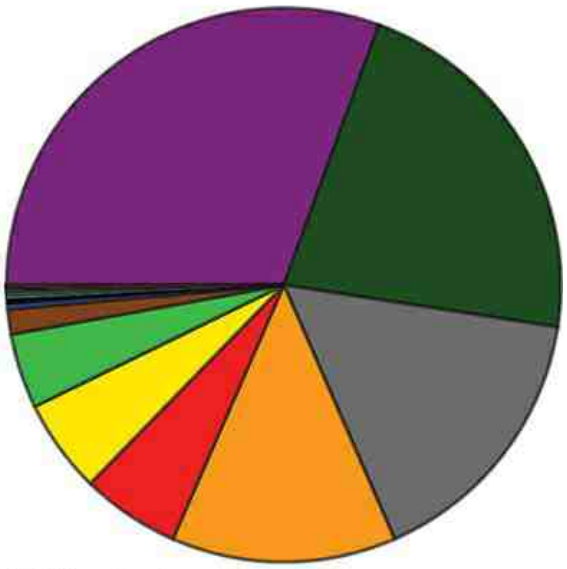


Figure 2.5: Relative frequency of number of TVs in a household
Example 2.1.4

We have data from adherents.com about the adherents of the largest religions. We display the data in Figures 2.6 and 2.7.

Christianity	2 100 000 000
Islam	1 500 000 000
None	1 100 000 000
Hindu	900 000 000
Traditional Chinese	394 000 000
Buddhism	376 000 000
Tribal religion	300 000 000
Traditional African	100 000 000
Sikh	23 000 000
Juche	19 000 000
Spirits	15 000 000
Judaism	14 000 000
Baha'i	7 000 000
Jainism	4 200 000
Shinto	4 000 000
Cao Dai	4 000 000
Zoroaster	2 600 000
Tenrikyo	2 000 000
Traditional European	1 000 000

(a) Colour codes.



(b) The pie chart.

Figure 2.6: Pie charts are good for visualizing proportions

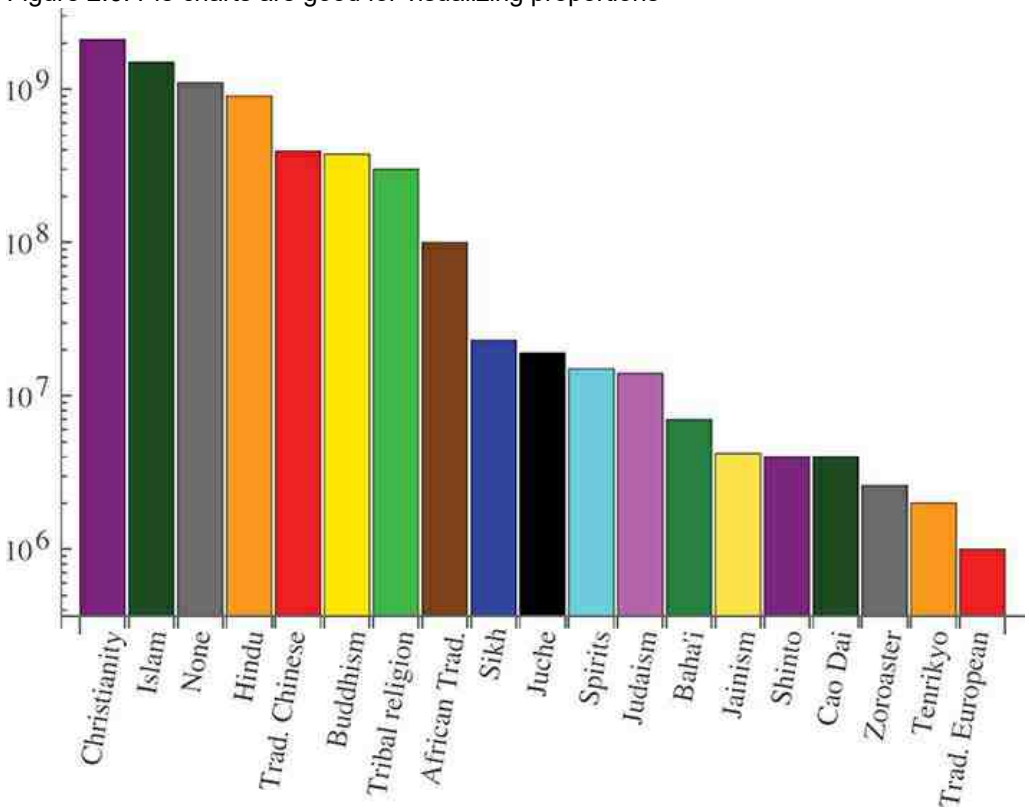


Figure 2.7: Bar chart with logarithmic vertical axis

In this chart, the relative size of each religion is very visible, since the size of the pie slices is proportional to the number of

adherents of each religion. But if we are interested making the smaller religious groups visible, we must abandon proportional representation. We choose a method that maintains the ordering by size, but not proportionally, by taking the logarithm of the number of adherents. The largest religions remain largest, but a factor of 10 is now only 1 unit, and a factor of 100 is 2 units higher.

So a logarithmic vertical axis brings forth the smaller religions in this chart.

2.1.1 Cumulative Data

In many applications, it is more useful to know how many of the values fall within a certain interval, or over or below a given value, than it is to know how many have a certain exact value. It is, for instance, of more interest to know how large a proportion of drivers have a blood alcohol content above $0.5‰$, than it is to know how many of them have precisely $0.73‰$. The most common way of stating these numbers is the *cumulative* frequency: the number who are at or *below* a given threshold value.

Example 2.1.5

(Continuation of Example 2.1.2) We asked 150 households how many TVs they had. We want to know how many households had three TVs *or fewer*, and the same for the other possible values. In Table 2.1, we expand the frequency table with an extra column for the cumulative frequency, and then form the table with the cumulative frequency alone.

Table 2.1: Cumulative frequency table for number of TVs in household

Value	Frequency	Cumulative frequency
1	33	**33**
2	39	$39 + 33 = \mathbf{72}$
3	42	$42 + 72 = \mathbf{114}$
4	36	$36 + 114 = \mathbf{150}$

Value	Cumulative frequency
1	33
2	72
3	114
4	150

The cumulative frequency chart is related to its table in the same way that the bar chart is related to the regular frequency table. In Figure 2.14, we illustrate how we construct the cumulative frequency diagram 2.8c from the frequency chart 2.8a:

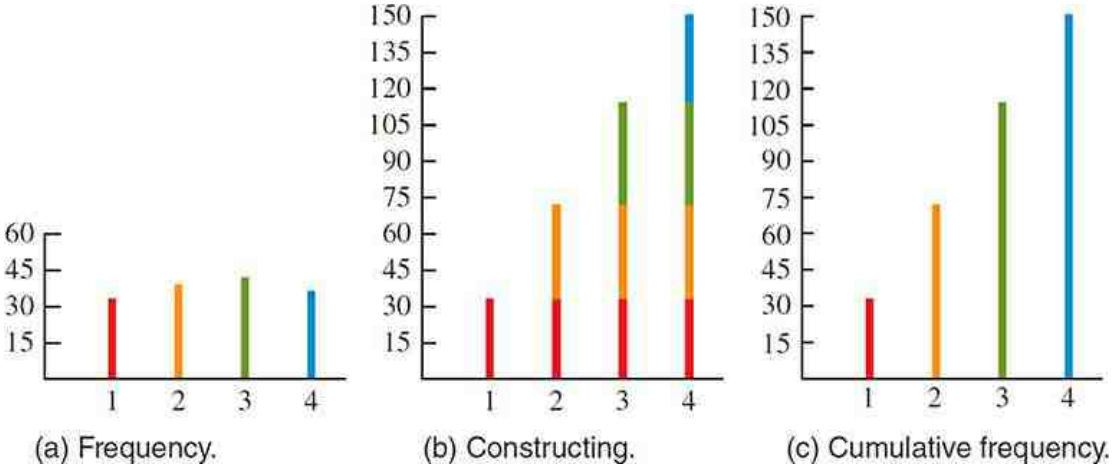


Figure 2.8: Construction of the cumulative frequency diagram

2.2 Measure of Location: Mode

It is frequently expedient to say what is a "typical value" for a given set of data. This is often known as a *measure of location*. The main measures of location are the *mean* $\bar{x}$ and the *median* $\tilde{x}$. The median is literally in the middle, with half of the observations on either side, whereas the mean is weighted by distance, like the center of gravity of a group of equally heavy objects.

But there is one more measure of what is "typical", and it's particularly useful for categorical data. You can't perform arithmetic on categorical data, or even sort by size, so there is no mean or median. But the *mode* can be defined regardless of the nature of the data. It is simply the most frequently occurring value. In the religions example, the mode is "Christianity", since it has the highest number of adherents. In the TV survey, the mode is three, since the most common number of TVs in a household is three.

2.3 Proportion Based Measures: Median and Percentile

For proportion based measures, we order our observations by giving them new indexes by rising order of magnitude: $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$.

Example 2.3.1

Let $x_1 = 9$, $x_2 = -1$, $x_3 = 7$, $x_4 = 5$, $x_5 = 2$. Switching to ordering indexes, we get $x_{(1)} = -1$, $x_{(2)} = 2$, $x_{(3)} = 5$, $x_{(4)} = 7$, $x_{(5)} = 9$.

The median is the "observation in the middle": there are as many observations above it as below it. We calculate the median $\tilde{x}$ of the observations $x_{(1)}$, …, $x_{(n)}$ like this: If n is odd, there is one observation in the middle, which is the median. If n is even, there are two observations in the middle; the median is then the middle value between these two.

Definition 2.3.2

The median $\tilde{x}$ of the observations $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$ is given by

$$\tilde{x} = \begin{cases} x_{([n+1]/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases}$$

Example 2.3.3

Your observations are 2, 200, 10, 78, 5. What is the median?

Answer: We order the five observations by value (see Figure 2.9): $x_{(1)} = 2$, $x_{(2)} = 5$, $x_{(3)} = 10$, $x_{(4)} = 78$, $x_{(5)} = 200$. Then

$$\tilde{x} = x_{([5+1]/2)} = x_{(3)} = \underline{\underline{10}}.$$

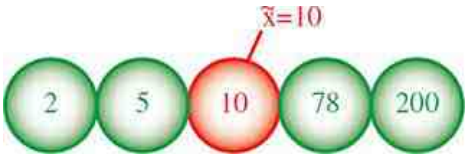


Figure 2.9: Odd numbers: the median is the middle observation

Example 2.3.4

Your observations are 3, 6, −1, 7, 6, 5. What is the median?

Answer: We order the six observations by value (see Figure 2.10): $x_{(1)} = -1$, $x_{(2)} = 3$, $x_{(3)} = 5$, $x_{(4)} = 6$, $x_{(5)} = 6$, $x_{(6)} = 7$. Then

$$\tilde{x} = \frac{1}{2}\left(x_{(6/2)} + x_{(6/2+1)}\right) = \frac{1}{2}(x_{(3)} + x_{(4)}) = \frac{1}{2}(5+6) = \underline{\underline{5.5}}.$$

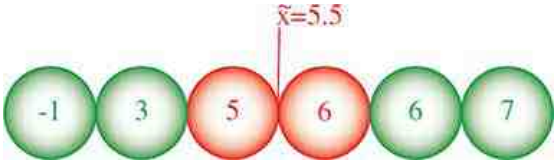


Figure 2.10: Even numbers: the median is the average of the two middle observations

The formula remains the same if we have grouped data having the same value. We just need to remember to count each value as many times as it was observed, and find the middle of the sorted observations.

Example 2.3.5

What is the median number of TVs in the households from Examples 2.1.2, 2.1.3, and 2.1.5?

To answer this calls for a revisit of the cumulative table, as shown in Table 2.2).

Table 2.2: Using the cumulative table to find the median

Value	Cumulative frequency	Observations, ordered
1	33	$x_{(1)}$ to $x_{(33)}$ have 1 TV.
2	72	$x_{(34)}$ to $x_{(72)}$ have 2 TVs.
3	114	$x_{(73)}$ to $x_{(114)}$ have 3 TVs.
4	150	$x_{(115)}$ to $x_{(150)}$ have 4 TVs.

There are 150 measurements in total, an even number. The median is then the average of $x_{(75)}$ and $x_{(76)}$. We see from the ordered table that $x_{(75)} = 3$ and $x_{(76)} = 3$, so $\tilde{x} = (3+3)/2 = \underline{3}$.

2.3.1 Measure of Proportion: Percentile

The median $\tilde{x}$ divides the data so that at least 50% of them are smaller than or equal to $\tilde{x}$, and at least 50% are larger than or equal to $\tilde{x}$. The generalization of this is the *p*th *percentile*, a value that divides the data so that at least *p*% of them are smaller than or equal to it, and at least (100 − *p*)% of them are larger than or equal to it. There are, however, several different versions of the percentile, so we must choose one standard among the many possible.

We will follow the National Institute of Standards and Technology, NIST. The reasoning leading up to this standard goes like this: The median is the most important percentile, the 50th percentile. For an even number of observations $\tilde{x}$ is an average. But let us refine that perspective by allowing non-integer ordering indexes: If we have $n = 4$ observations, $\tilde{x} = x_{((50/100)\times(n+1))} = x_{((50/100)\times(4+1))} = x_{(2.5)}$. This puts $x_{(2.5)}$ right in the middle of $x_{(2)}$ and $x_{(3)}$, giving us $x_{(2.5)} = 0.5x_{(2)} + 0.5x_{(3)}$ by linear interpolation. We then define the *p*th percentile as $x_{((p/100)\times(n+1))}$, and calculate it by linear interpolation where necessary. We will mostly get $x_{(\kappa)}$ for a non-integer κ. For instance $x_{(3.7)}$. We calculate $x_{(3.7)}$ by linear interpolation, like this: $x_{(3.7)} = (1 - 0.7)x_3 + 0.7x_4$. The general formula and method are as follows.

Definition 2.3.6

The p*th percentile for an ordered list* $\{x_{(1)}, \ldots, x_{(n)}\}$ *is*

$$P_p = x_{(\kappa)} = x_{(h)} + d \times (x_{(h+1)} - x_{(h)}),$$

where $\kappa = \frac{p}{100} \times (n+1)$, *a number with integer part h and decimal part d, meaning* $\kappa = h.d$. *If* $\kappa < 1$ or $\kappa > n$, *use respectively* $\kappa = 1$ or $\kappa = n$ *instead*.

Mathematica: Quantile$[list, \frac{p}{100}, \{\{0,1\},\{0,1\}\}]$

Excel: PERCENTILE.EXC(*marked cells*, $\frac{p}{100}$)

Method 2.3.7

How to calculate the *p*th percentile P_p, in detailed steps in the table below (with illustration in Figure 2.11):

Method	Example
0. Find the data and the desired percentile.	Data: $x_1 = 5.5, x_2 = 7.42, x_3 = -14.7,$ $x_4 = 22.8, x_5 = 1.12, x_6 = 5.02, x_7 = 1.$ Find the 70th percentile for these data.
1. Identify p and n.	$p = 70$ and $n = 7$.
2. Order the data by rising value.	$x_{(1)} = -14.7, x_{(2)} = 1, x_{(3)} = 1.12,$ $x_{(4)} = 5.02, x_{(5)} = 5.5, x_{(6)} = 7.42,$ $x_{(7)} = 22.8.$
3. $\kappa = \frac{p}{100} \times (n+1)$, with integer part h and decimal part d.	$\kappa = \frac{70}{100} \times (7+1) = 5.6$, so $h = 5$ and $d = 0.6$.
4. $P_p = x_{(\kappa)}$ $= x_{(h)} + d \times (x_{(h+1)} - x_{(h)})$.	$P_{70} = x_{(5.6)}$ $= x_{(5)} + 0.6 \times (x_{(6)} - x_{(5)})$ $= 5.5 + 0.6 \times (7.42 - 5.5) = \underline{6.652}$.

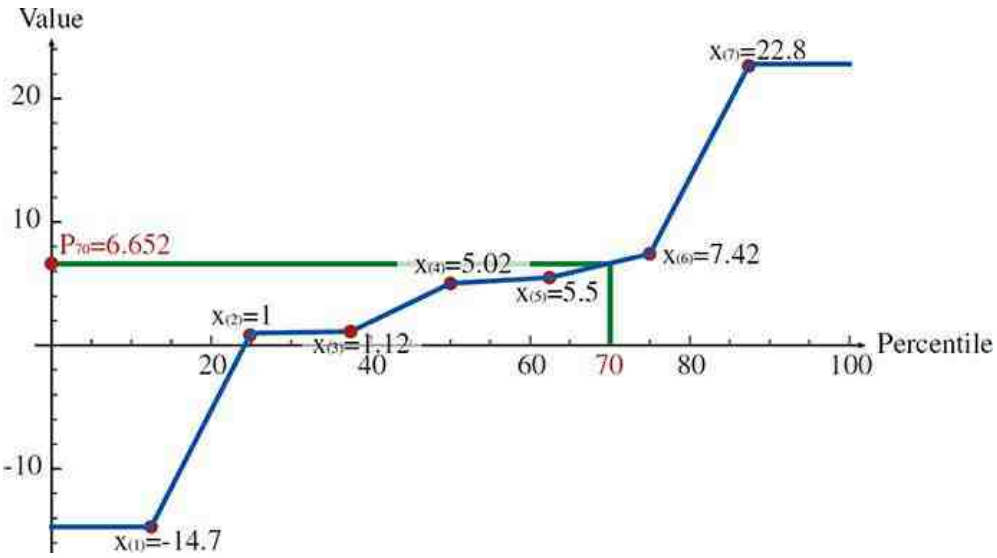


Figure 2.11: Percentile

Example 2.3.8

The median $\tilde{x} = P_{50}$ *is one of three important percentiles known as the quartiles*: $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$. *The interquartile range* $Q_3 - Q_1$ *is often used as a measure of how spread out the data are.*

2.3.2 Measure of Location: Mean

The *mean* is a popular measure, and is the value we obtain if we obtain is we divide the total evenly between the objects we have measured. If, for instance, Peter, Paul, Ewan, and Tom earn respectively 1, 2, 4, and 5 ounces of gold per month, they would each have earned 3 ounces of gold per month if it had been divided evenly, since $\frac{1+2+4+5}{4} = 3$. So 3 is the *mean* of the numbers 1, 2, 4, and 5.

Definition 2.3.9

Given values $x_1, \ldots, x_n$, *the sum is*

$$\Sigma_x = \sum_{k=1}^{n} x_k.$$

We use this for the following formal definition of the mean.

Definition 2.3.10

The **mean** $\bar{x}$ *of the observations* $x_1, x_2, \ldots, x_n$ *is defined as*

$$\bar{x} = \frac{\Sigma_x}{n}.$$

Example 2.3.11

"Ye olde milky bar" wants to know the mean number of milk shakes sold per day, for a given week. The number of milk shakes sold on the different weekdays are $x_1 = 163$, $x_2 = 178$, $x_3 = 167$, $x_4 = 191$, $x_5 = 175$. The mean number of units sold per day is then the total divided by the number of days:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{163 + 178 + 167 + 191 + 175}{5} = 174.8.$$

If we have many measurements of equal value, it may make sense to group the data by value. We definere v_k to be the measured *values*, and a_k to be the number of observations with value v_k. Then

(2.1) $$\bar{x} = \frac{\Sigma_x}{n} \quad \text{where } \Sigma_x = \sum_k a_k v_k \quad \text{and} \quad n = \sum_k a_k.$$

Example 2.3.12

We want to find the mean number of TVs in Example 2.1.2. We use formula 2.1 on the data from the table, and get

$$\bar{x} = \frac{33 \times 1 + 39 \times 2 + 42 \times 3 + 36 \times 4}{33 + 39 + 42 + 36} = 2.54.$$

We get a related formula if we use the *relative* frequency p_k rather than the frequency a_k:

(2.2) $$\bar{x} = \sum_k p_k \times v_k, \qquad \text{where } p_k = \frac{a_k}{\sum a_k}.$$

The relative frequency formula is particularly useful when there is no canonical base unit to break the measurements down into, as in the following example:

Example 2.3.13

In the hydro power plant *Vanna* they measure the water flow through the turbines at millions of cubic meter per hour. We have the following data from 14:00 to 15:00 on a given day: For the first 17 minutes the flow was 0.83 million m^3/h. From 14:17 to 14:45, it was 1.13 million m^3/h, after which it fell to 0.98 million m^3/h for the last quarter hour. What was the mean water flow at Vanna during that hour?

So how large a proportion of the hour did each period take? The first period lasted 17 minutes, so $p_1 = \frac{17}{60}$, and $v_1 = 0.83$. The second period lasted 28 minutes, so $p_2 = \frac{28}{60}$ and $v_2 = 1.13$. Finally, the third period lasted for 15 minutes, so $p_3 = \frac{15}{60}$ and $v_3 = 0.98$. The mean flow is then

$$\bar{x} = p_1 v_1 + p_2 v_2 + p_3 v_3 = \frac{17}{60} \times 0.83 + \frac{28}{60} \times 1.13 + \frac{15}{60} \times 0.98 = \underline{\underline{1.0075}}.$$

2.4 Measures of Spread: Variance and Standard Deviation

Variance and standard deviation are measures of how spread out the data are around the mean. The standard deviation is the square root of the variance, so we always calculate the variance first. There are two types of variance, the *sample* variance and the *population* variance. We write σ^2_x for the population variance, and s^2_x for the sample variance, meaning σ_x and s_x are the respective standard deviations. We recall that a *population* is all the possible values in our scope, whereas the *sample* is our actual measurements.

Definition 2.4.1

For values x_1, …, x_n, the sum of the squared deviations is

$$SS_x = \sum_{k=1}^{n} (x_k - \bar{x})^2.$$

Given a finite population, the *population variance* is simply the average squared deviation for the population.

Definition 2.4.2

The population variance *of a population whose values are x_1, x_2, …, x_n, is*

$$\sigma_x^2 = \frac{SS_x}{n}.$$

The quantity $\sigma_x = \sqrt{\sigma_x^2}$ *is the* population standard deviation.

However, we rarely measure entire populations, so the population variance is mostly a creature of theory. What we tend to do instead, is to measure a smaller *sample* from the population, and approximate the variance by calculations from that sample. The best approximation to the population variance of the total, is the *sample variance*. The sample variance is similar to the population variance, but to compensate for how the sample mean deviates from the population mean, we divide by $n - 1$ instead of n.

Look in Section 16.1 for a more detailed exposition of why we divide by $n - 1$ rather than by n.

Definition 2.4.3

The sample variance *calculated from the values x_1, x_2, …, x_n, is*

$$s_x^2 = \frac{SS_x}{n-1}.$$

The sample standard deviation *is* $s_x = \sqrt{s_x^2}$.

The most useful formula for SS_x is the following one, using the sum of squares.

Definition 2.4.4

The square sum *of the values x_1, x_2, …, x_n, is*

$$\Sigma_{x^2} = \sum_{k=1}^{n} x_k^2.$$

As we add more data, we just need to add the new values x_i to Σ_x, and x^2_i to our new friend Σ_{x^2} for book-keeping. This way, we don't need to recalculate all the $(x_k - \bar{x})^2$ against a $\bar{x}$ that will of course change value as new data arrive.

Rule 2.4.5

$$SS_x = \Sigma_{x^2} - n \cdot \bar{x}^2 = \Sigma_{x^2} - \frac{\Sigma_x^2}{n}.$$

Example 2.4.6

"Ye olde milky bar" wants to find the variance and standard deviation of the number of milk shakes sold per day. They use one week's measurements. They want to

1. find the variance and standard deviation of the number of servings within the week itself, i.e. the week itself is the population;
2. estimate the variance and standard deviation of the number of servings over a larger time period, i.e. the week itself is a sample from the greater time period.

We recall our data, x_1 = 163, x_2 = 178, x_3 = 167, x_4 = 191, x_5 = 175, and remember from example 2.3.11 that $\bar{x} = 174.8$. We then get

1. $SS_x = (163^2 + 178^2 + 167^2 + 191^2 + 175^2) - 5 \cdot 174.8^2 = 472.8$;
2. *population variance and standard deviation for the week itself:*
   - $\sigma_x^2 = \frac{SS_x}{5} = \frac{472.8}{5} = 94.56$
   - $\sigma_x = \sqrt{94.56} = 9.7$;
3. *Sample variance and standard deviation for the week as an approximation to a larger time frame:*
   - $s_x^2 = \frac{SS_x}{5-1} = \frac{472.8}{5-1} = 118.2$
   - $s_x = \sqrt{118.2} = 10.9$.

We have the following simplified formula when the data are given with relative frequencies.

Rule 2.4.7

$$\sigma_x^2 = \left(\sum_k p_k \times v_k^2\right) - \bar{x}^2.$$

We revisit the continuous measurements of the water flow at the *Vanna* hydro power plant in the following example.

Example 2.4.8

We will now find the variance and standard deviation of the water flow at *Vanna*. From Example 2.3.13, we have $p_1 = \frac{17}{60}$ and v_1 = 0.83, and then $p_2 = \frac{28}{60}$ and v_2 = 1.13, and finally $p_3 = \frac{15}{60}$ and v_3 = 0.98. Mean flow was $\bar{x} = 1.0075$. So

- $\sigma_x^2 = \left(\frac{17}{60} \times 0.83^2 + \frac{28}{60} \times 1.13^2 + \frac{15}{60} \times 0.98^2\right) - 1.0075^2 = 0.016\,1187$;
- $\sigma_x = \sqrt{0.016\,1187} = 0.126\,96$.

2.5 Grouped Data

We often group data into larger groups of values, either for expediency, or because the data already is present in such groups. One example is clothes sizes, where we are typically presented with values of the kind "49–51 cm" rather than a precise, single value.

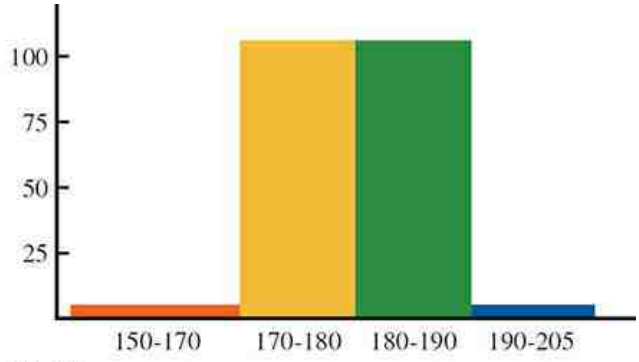
But the accuracy of measurements also naturally groups the data into such groups. If we measure the height of recruits for a military batallion, our accuracy will be in the order of 1 cm, so a measurement of "175 cm" really means all heights from 174.5 to 175.5 cm, and so on. Or it could be deliberately divided into even coarser groups, since a single cm matters little. We then display our data in a *histogram*. Notice that histograms differ from bar charts in that they mark entire intervals packed back to front, rather than individual values. Do also note that with the histogram, we use area instead of height.

Example 2.5.1

As an example, here are the heights of recruits from Indre Istindfjord in 1959, grouped into groups of uneven width for the purpose of illustration in Figure 2.12.

Height in cm	Number of recruits
150–170	88
170–180	1032
180–190	811
190–205	69
Sum	2000

(a) Grouped frequency table.



(b) Histogram.

Figure 2.12: Indre Istindfjord, 1959

How do we treat grouped data? We have the following two options.

1. *Representation*

   We treat the data in the interval between x and y as discretely represented by the interval midpoint $\frac{x+y}{2}$, and calculate as if the midpoint was the actual value. With this option, we use the formulas we have already established for point data. When we ignore measure inaccuracy, and for instance just record the height of all men between 173.5 and 174.5 cm as "174 cm", this is what we actually do, and it works fine as long as the intervals are not too wide. See Figure 2.13.

2. *Continuous*

   We consider the data in the interval between x and y to be evenly distributed over the interval. At the beginning of the interval, we will therefore include none of the data, but will then progress along a straight line until the end of the interval, where all the data in the interval are included.

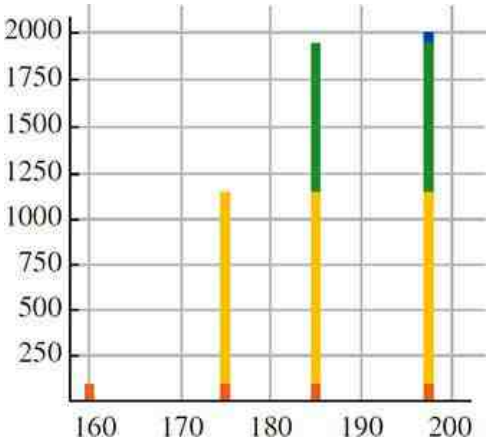
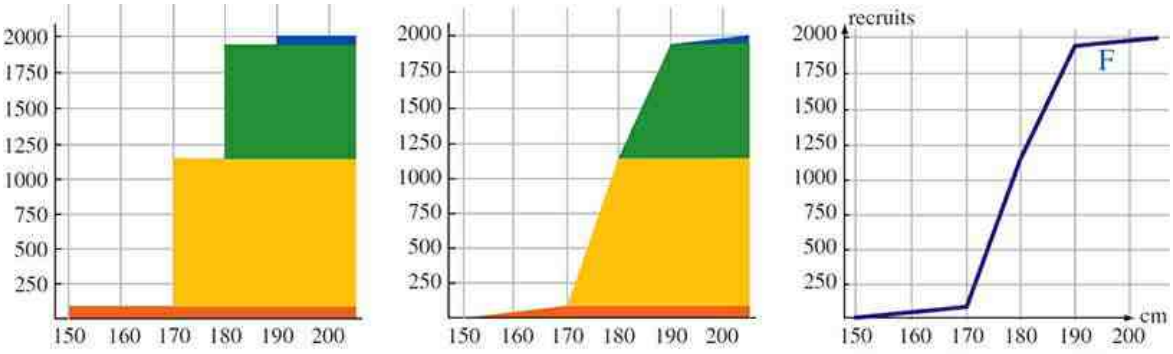


Figure 2.13: Cumulative bar chart when data are treated by interval midpoint

We will explore the continuous alternative. The cumulative graph in Figure 2.14 provides the information we need, and is also easy to set up directly from the values at the ends of the interval.



(a) Cumulative chart by itself. (b) Evenly distributed. (c) Cumulative graph $F(x)$ by itself.

Figure 2.14: Cumulative bar chart when data are treated as evenly distributed over interval

Example 2.5.2

We want to know the number of recruits at 183 cm or shorter, so we draw the cumulative graph for the data in Figure 2.15, and read (red line) that there is slightly in excess of 1400 recruits at 183 cm or less.

Height in cm	(Cumulative) frequency
Up to 150	0
Up to 170	88
Up to 180	1120
Up to 190	1931
Up to 205	2000

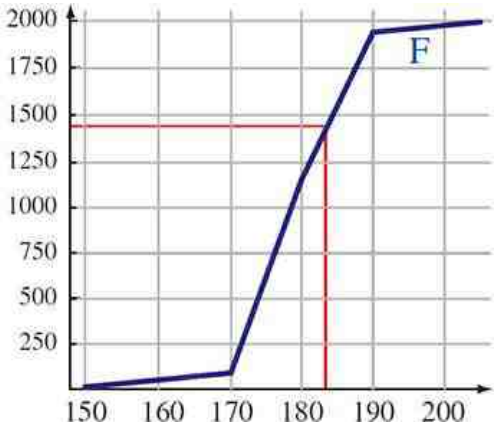


Figure 2.15: Finding percentiles when data are treated as evenly distributed over their respective intervals

We can study *proportions* in the same way.

Example 2.5.3

We want to find the proportion of recruits between 172 and 178 cm. We may do this in two ways, both illustrated in Figure 2.16. We either find the area under the graph of histogram 2.16a between 172 and 178, divided by the total area below the graph. Or: we simply read the difference between the values of the cumulative graph 2.16b at 172 and at 178.

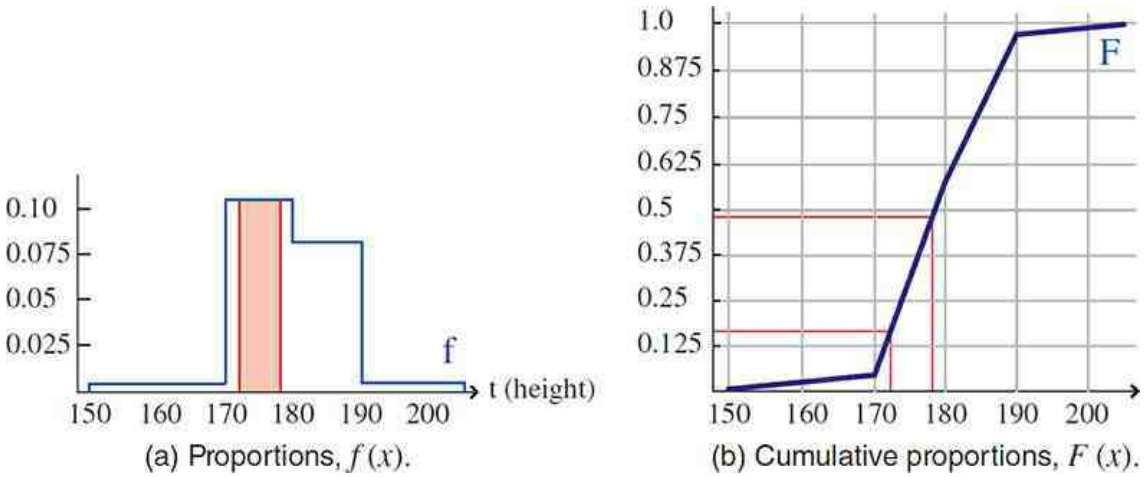


Figure 2.16: Finding proportions when data are treated as evenly distributed over interval

The cumulative approach is of course also just an approximation, giving us half and quarter recruits. But it enables us to deal with large data sets, and is therefore often preferable to exact calculations even where these are possible. They are also an early model of another type of proportion: *probabilities*. For even though observed recruits never will be evenly smeared over an interval, the *probabilites* for the different heights will be distributed with a density given by a function not too dissimilar from these continuous distributions of data.

2.5.1 Measures of Location and Spread for Grouped Data

Question: What were the mean, and the sample variance and standard deviation of the height of the recruits from Indre Istindfjord in 1959?

We will answer this question both by representation and by the continuous model.

Example 2.5.4

Answered through *representation*: the first group, from 150 to 170, has $a_1 = 88$ data points, and is represented by the midtpoint $v_1 = 160$. For the second group, $v_2 = 175$ and $a_2 = 1032$. The third group has $v_3 = 185$ and $a_3 = 811$. Finally, for the fourth group, $v_4 = 197.5$ and $a_4 = 69$. The total number of recruits is $n = a_1 + a_2 + a_3 + a_4 = 2000$. This gives an average height of

- $\Sigma_x = 88 \times 160 + 1032 \times 175 + 811 \times 185 + 69 \times 197.5 = 358\,343$
- $\bar{x} = \frac{\Sigma_x}{n} = \frac{358\,343}{1999} = 179.171.$

We calculate (sample) variance and standard deviation as follows:

- $\Sigma_x^2 = 88 \times 160^2 + 1032 \times 175^2 + 811 \times 185^2 + 69 \times 197.5^2 = 64\,305\,706$
- $SS_x = \Sigma_x^2 - \Sigma_x^2/n = 64\,305\,706 - \frac{358\,343^2}{2000} = 101\,032$
- $s_x^2 = \frac{SS_x}{n-1} = \frac{101\,032}{1999} = 50.54$
- $s_x = \sqrt{50.5413} = 7.11.$

For the *continuous* model, we need to modify the formulas used to calculate for variances. The modifications are needed in order to take into account that the data already are spread out by virtue of being evenly distributed over the interval. It turns out that the needed correction is a factor of $\frac{1}{12}$ times the width of the interval, like this:

> Let interval k be $I_k = (l_k, u_k)$. The interval *midpoint* is $v_k = (l_k + u_k)/2$ and the interval *width* is $b_k = u_k - l_k$. With this notation, the modified formulas may be stated as the following rule.

Rule 2.5.5

For numerical data grouped into intervals where interval k has midpoint v_k and width b_k, we get that

(2.3) $$\Sigma_x = \sum_k a_k \times v_k$$

(2.4) $$\Sigma_{x^2} = \sum_k a_k \times \left(v_k^2 + \frac{1}{12} \times b_k^2\right).$$

Example 2.5.6

We return our attention to the example of the recruits: Σ_x and $\bar{x}$ are the same for both models, so Σ_x = 358 343 and $\bar{x} = 179.171$

$$\begin{aligned}\Sigma_{x^2} &= 88 \times \left(160^2 + \tfrac{1}{12} \times 20^2\right) + 1032 \times \left(175^2 + \tfrac{1}{12} \times 10^2\right) \\ &\quad + 811 \times \left(185^2 + \tfrac{1}{12} \times 10^2\right) + 69 \times \left(197.5^2 + \tfrac{1}{12} \times 15^2\right) \\ &= 64\,325\,291\end{aligned}$$

$$SS_x = \Sigma_{x^2} - \frac{\Sigma_x^2}{n} = 64\,325\,291 - \frac{358\,343^2}{2000} = 120\,438$$

$$s_x^2 = \frac{SS_x}{n-1} = \frac{120\,438}{1999} = 60.25$$

$$s_x = \sqrt{60.25} = 7.76.$$

2.5.2 Median and Percentile for Grouped Data

Definition 2.5.7

For grouped data we define the percentile P_p as the t value at which $F(t) = \frac{p}{100}$. The median is P_{50}.

Method 2.5.8

1. Set up a table of cumulative relative frequencies. In the relative frequency column, for interval k, you will have $F(u_k)$ (which is equal to $F(l_{k+1})$).
2. Locate the group k for which $F(l_k) \le \frac{p}{100} \le F(u_k)$.
3. $$P_p = l_k + \frac{\frac{p}{100} - F(l_k)}{F(u_k) - F(l_k)} \times (u_k - l_k)$$.

Example 2.5.9

Find the 37.4th percentile for the Indre Istindfjord recruits in 1959.

1. We do this through the cumulative table (Table 2.3), and the diagram shown in Figure 2.17.
2. $\frac{p}{100} = 0.374$, which is between $F(l_2) = 0.044$ and $F(u_2) = 0.56$, so $k = 2$.
3. Since $l_2 = 170$ and $u_2 = 180$, then

$$P_{37.4} = l_2 + \frac{\frac{p}{100} - F(l_2)}{F(u_2) - F(l_2)} \times (u_2 - l_2)$$
$$= 170 + \frac{0.374 - 0.044}{0.56 - 0.044} \times (180 - 170) = 176.395.$$

Table 2.3: Cumulative frequency table and cumulative relative frequency table

k	Height in group k	Cumulative frequency, $\tilde{F}(u_k)$	Relative cumulative frequency, $F(u_k)$
0	Up to 150	0	0
1	150 to 170	88	0.044
2	170 to 180	1120	0.56
3	180 to 190	1931	0.9655
4	190 to 205	2000	1

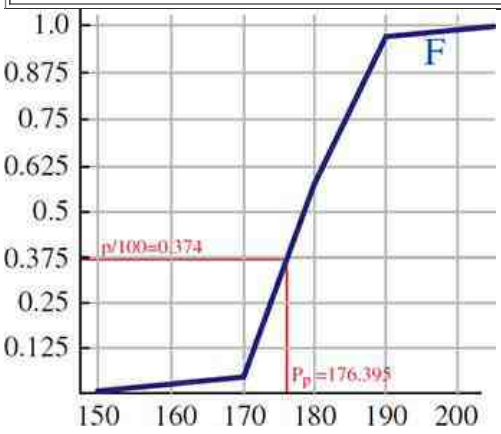


Figure 2.17: Finding the percentile through a cumulative graph

2.6 Exercises

1. **Review:** Read the chapter. ?
   a. Explain in your own words what Σ_x, Σ_{x^2}, SS_x, $\bar{x}$, and $\tilde{x}$ are. Why do we have different formulas for these quantities?
   b. What is the difference between *population* and *sample*, and how are they connected?
2. **Measures of location:** Find all the three measures of location (mode, median, mean), and decide which is the most suitable one for the situation. ?
   a. The members of Femund Fishers' Union are located as follows:
      i. 24 are from Drevsjø in Engerdal, which has postal code 2443
      ii. 6 are from Ålesund, which has postal code 6020
      iii. 19 are from Røros, which has postal code 7374
      iv. 1 are from Bodø, which has postal code 8092
   b. A class of 50 enginering students had the following income distribution 3 months after graduating:
      i. 5 were unemployed, so their income was 0.
      ii. 41 had yearly salaries (in NOK) of respectively 340 000, 341 000, 342 000 … in increments of 1000 up to 380 000.
      iii. The last four earned respectively 613 000, 727 000, 958 000 and 70 000 000.
   c. For a class of 100 economists, the wage distribution was as follows 3 months after graduation:
      i. 51 earned nothing.

ii. Of the 49 remaining, 24 earned 312 000, while 25 earned 478 000.

3. The position number $\omega(letter)$ tells us where in the alphabet the given *letter* is located. For instance: ω("b") = 2. Find the mean, median and population standard deviation for the vowel position numbers in the English alphabet.

4. For the data sets below:

- Calculate the median and the interquartile range.
- Find the mean and the sample standard deviation.

a. $\{-1, -3, 4\}$

b. $\{-0.2, 9.6, -0.1, 11.1, 1.3, -0.2, 11.1, -0.8, 0.4\}$

c. {60, 66, 70, 103, 138, 34}

d. {0.971 49, 0.659 64, 0.345 81, 0.515 90, 0.928 81}

5. We have written the data sets below as pairs of lists: $v = \{v_1, \ldots, v_n\}$ and $a = \{a_1, \ldots, a_n\}$, meaning a_1 observations of value v_1 etc.

- Set up a frequency table and bar chart.
- Set up a cumulative frequency table and a cumulative bar chart.
- Calculate the median and the interquartile range.
- Calculate the mean and the sample standard deviation.
- Mark these measures on the horizontal axes of your charts.

a. $v = \{5.6, 5.8, 5.1, 6.4, 5.2, 6.3, 5.0, 5.8, 6.0\}$ and $a =$

$\{3, 2, 1, 2, 9, 8, 7, 4, 8\}$

b. $v = \{2, 3, 5, 7, 11\}$ and $a = \{3, 4, 2, 6, 4\}$

c. $v = \{0.620\,362, 0.230\,49, 0.375\,471, 0.035\,230\,2, 0.562\,372, 0.485\,507\}$ and $a = \{109, 130, 73, 61, 9, 74\}$

6. You are in charge of a joint purchase of retro sports jackets for local FC Bayern Munich supporter club. The sizes correspond to chest measurements, and are (in cm): S=87-94, M=94-102, L=102-110, XL=110-121, XXL=121-133, 3XL=133-145. A few of the members are interested in making orders, and the total is 23S, 161M, 93L, 211XL, 131XXL and 42 3XL. Use the formulas for grouped data for calculations on the chest measurements in your local FCB supporter club.

a. Create a table and a cumulative table, and draw the histogram and the cumulative graph.

b. For each interval, find the interval limits l_k and u_k, and calculate the widths b_k and the average value v_k.

c. Find the median and the two quartiles.

d. Find mean and the sample variance.

7. We have written the data sets below as pairs of lists, $a = \{a_1, \ldots, a_n\}$ and $I = \{(l_1, u_1), \ldots, (l_n, u_n)\}$, meaning a_1 observations in interval I_1 etc.

- Make the cumulative table and graph.
- Calculate median and interquartile range.
- Find the mean and the sample standard deviation.
- Mark these measures on the horizontal axes of your charts.

a. $I = \{(12, 24)\}$ and $a = \{100\}$

b. $I = \{(0, 30), (30, 60), (60, 90)\}$ and $a = \{48, 96, 48\}$

c. $I = \{(0, 0.07), (0.07, 0.14), (0.14, 0.28), (0.28, 0.56), (0.56, 1.00)\}$ and $a = \{24, 38, 48, 66, 74\}$

8. A randomized survey among the supporters of the Scottish football team Heart of Midlothian FC yielded the following numbers in the different age groups: 0-12 years: 1, 13-18: 9, 19-34: 41, 35-50: 58, 51-64: 33, 65-80: 2. Use the formulas for grouped data in the following calculations:

a. Create a table and a cumulative table, and draw the histogram and the cumulative graph.

b. For each interval: Find the limits l_k and u_k, and the width b_k and the average value v_k. (Remember you are dealing with stated age here.)

c. Find the median and the two quartiles.

d. Find the mean and the sample standard deviation.

9. **The Big Gummy Worm Project:** This is a practical exercise, where you need

   - a big bag of gummy worms or equivalent
   - a measuring tool (e.g., a ruler or measuring tape)

   For each gummy worm,

   - stretch the gummy worm over the measuring tool until it snaps
   - write down the colour, and the length at which it snapped

   When you are done measuring and eating gummy worms, gather the data into tables: one joint table, and one per colour or colour group (depending on how many tables your teacher has told you to make). Remember that these are grouped data. For each group:

   - Create the frequency table
   - Create the cumulative frequency table
   - Draw a diagram for each table. (bar chart or histogram; which one do you think is suitable?)
   - Draw a cumulative diagram for each table. (cumulative bar chart or cumulative graph; which one do you think is suitable?)
   - Calculate the median and the mean for each group, and mark on the horizontal axis of the diagrams.
   - Calculate the population standard deviation for whole bag, and the sample standard deviation for each colour group. Explain the difference and the connection.

Answers

1. **Review:** Read the chapter itself until you have found your answers.
2. **Measures of location:**

   a.

      i. Mode = 2443 (Drevsjø; central location near Femunden, and has the most members)

      ii. Median = 6020 (Ålesund; is in the list, but with few members, and far away from Femunden)

      iii. Mean = 4859 (Postal code Nedenes, a small place down south far away from Femunden, not in list)

      iv. Mode is best.

   b.

      i. Mode = 0 NOK

      ii. Median = 359 500 NOK

      iii. Mean = 1.74 NOK

      iv. Median is the most representative.

   c.

      i. Mode = 0 NOK

      ii. Median = 0 NOK

      iii. Mean = 194 000 NOK

      iv. Mean is the most representative.

3. The position number … The vowels are *a, e, i, o, u*, and their positions are 1, 5, 9, 15, 21. $\Sigma_x = 51$, so the mean = 10.2, the median = 9, and

since $\Sigma_{x^2} = 773$, then SS_x = 252.8, which means the population standard deviation is σ_x = 15.9.

4. To find the answers in Mathematica/Wolfram Alpha, write Median[list],…, Mean[list], StandardDeviation[list]. The interquartile range is a bit heftier, since there are about 20 different versions of it, and you have to specify which:

   Quantile[theList, 3/4, {{1, −1}, {0, 1}}] − Quantile[theList, 1/4, {{1, −1}, {0, 1}}].

   For the exercises, we get:

   a. $\tilde{x} = -1$, $Q_3 - Q_1 = 7$, $\bar{x} = 0$, $s_x = \sqrt{13}$

   b. $\tilde{y} = \underline{0.4}$, $Q_3 - Q_1 = 10.55$, $\bar{y} = 3.577\,78$, $s_y = 5.314\,55$

   c. $\tilde{z} = \underline{68}$, $Q_3 - Q_1 = 58.25$, $\bar{z} = 78.5$, $s_z = 36.582\,8$

   d. $\tilde{w} = \underline{0.659\,64}$, $Q_3 - Q_1 = 0.519\,295$, $\bar{w} = 0.684\,33$, $s_w = 0.267\,304$.

5. The numbers are:

   a. $\tilde{y} = 5.8$, $Q_3 - Q_1 = 0.8$, $\bar{y} = 5.675$, $s_y = 0.498\,428$

   b. $\tilde{z} = 7$, $Q_3 - Q_1 = 4$, $\bar{z} = 6$, $s_z = 3.265\,99$

   c. $\tilde{w} = 0.375\,471$, $Q_3 - Q_1 = 0.331\,882$, $\bar{w} = 0.368\,707$, $s_w = 0.196\,231$.

6. You are in charge of …

   a. …

   b. $b_S = 7$, $b_M = 8$, $b_L = 8$, $b_{XL} = 11$, $b_{XXL} = 12$, $b_{3XL} = 12$, and $v_S = 90.5$, $v_M = 98$, $v_L = 106$, $v_{XL} = 115.5$, $v_{XXL} = 127$, $v_{3XL} = 139$

   c. $\tilde{x} = 112.789$, $Q_1 = 101.068$, and $Q_3 = 121.71$

   d. $\bar{x} = 112.803$, and $s_x^2 = 171.877$.

7. We give the values only – no diagrams or tables:

   a. $\tilde{x} = 18$, $Q_3 - Q_1 = 6$, $\bar{x} = 18$, $s_x = 3.481\,55$

   b. $\tilde{x} = 45$, $Q_3 - Q_1 = 30$, $\bar{x} = 45$, $s_x = 22.972\,8$

   c. $\tilde{x} = 0.343\,636$, $Q_3 - Q_1 = 0.486\,920$, $\bar{x} = 0.401\,4$, $s_x = 0.288\,17$.

8. A randomized survey …

   a. …

   b. Intervals: (0,13), (13,19), (19,35), (35,51), (51,65), (65,81) (If you are 12.95 years old, your stated age is 12.) Finding b_k and v_k is straightforward.

   c. $\tilde{x} = 40.793\,1$ and $Q_1 = 29.146\,3$ and $Q_3 = 50.724\,1$

   d. $\bar{x} = 40.357\,6$ and s_x =14.324 8.

   The solution here will depend on your specific jelly worms.