Applying Statistics in Behavioural Research

Applying Statistics in Behavioural Research

Jules L. Ellis

Boom

Contents

Preface

The primary focus of this book is: reporting about behavioural data. The book tries to teach this in an efficient way, by using the same structure of a 'basic report' and a 'short report' for many different analyses. Although that is not very exciting for the teacher, it saves students time because they need less time to discover the structure of what is being learnt.

The second focus of this book is: understanding the relation between data and statistics. This is the reason why, in the first three parts of the book, students are required to perform most computations themselves, using only a simple calculator rather than statistical software. This is also the reason why there are special chapters about visualising, in which the student learns how the features of a data plot influence the various statistics.

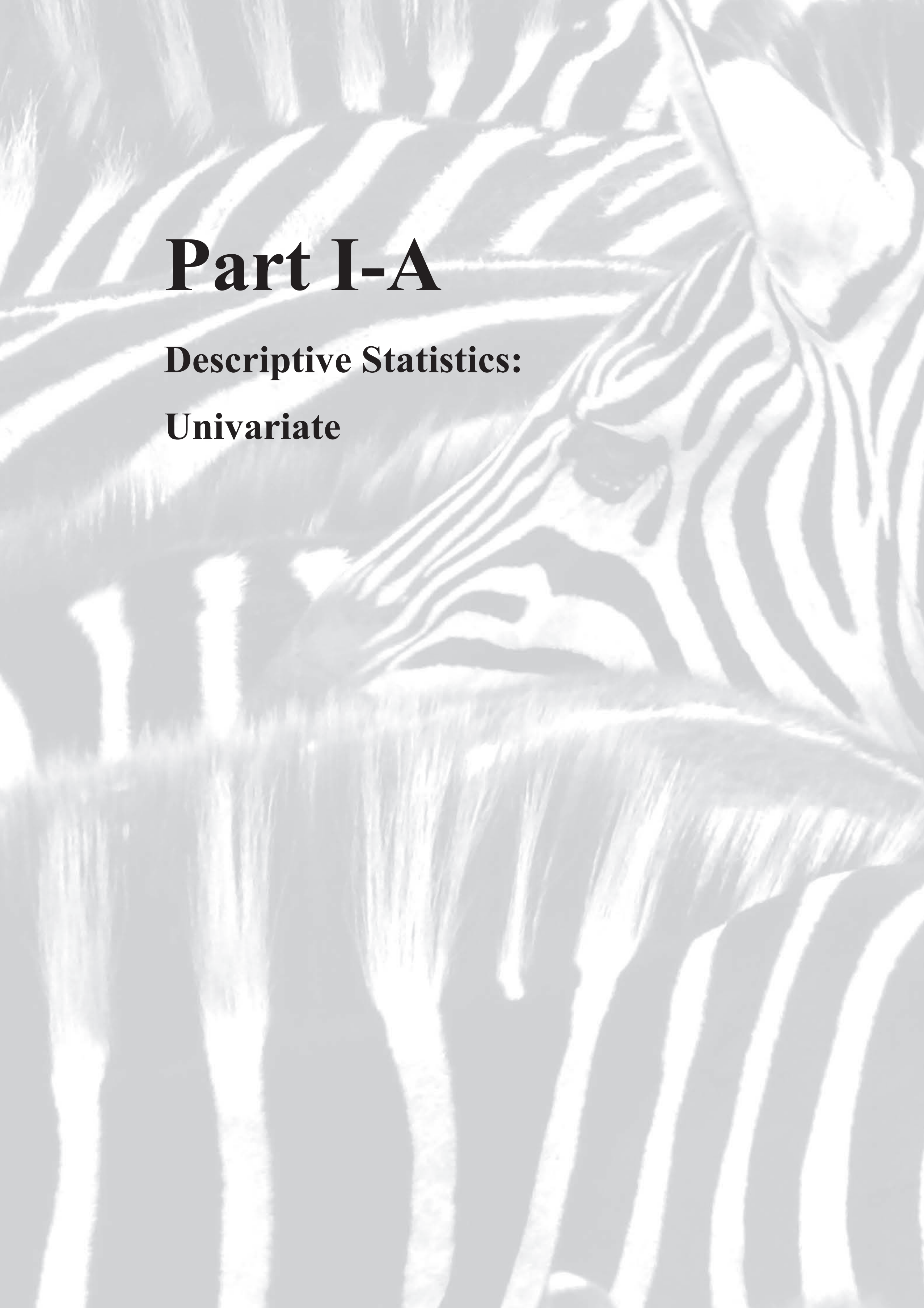
Reading the book is not enough. Students should spend most of their time to the exercises. Actually completing the exercises, and not avoiding them, is maybe the most challenging part of learning to apply statistics. If you are a student, you might easily find other urgent activities. But trust me, there is only one way to learn statistics: Make the exercises.

The book's website www.applyingstatisticsinbehaviouralresearch.com contains further material that may help you with learning. First, the website has several apps that automatically create new exercises and that correct your answers. Second, the website offers self tests that may be helpful with preparing for exams. Third, the website provides outcomes of selected exercises.

In the future, the website may contain try-out versions of new material for the book. Some topics that I plan to include are: factor analysis, item response theory, pitfalls in intuitive statistical thinking, and probability.

The present book is based on my Dutch series of introductory statistics books. Some chapters are almost literal translations. This is true for the chapters on the NHST, power, and Simpson's paradox. Most chapters are more losely based on the Dutch version, however. In any case, many students and assistants have made valuable suggestions throughout the years.

Jules L. Ellis,
Nijmegen



Part I-A

Descriptive Statistics:

Univariate

1 Introduction of the examples

What is statistics? According to the American Statistical Association (2016),

> 'Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (Davidian, M. and Louis, T. A., 10.1126/science.1218685).'

So, for learning about statistics then, we need data. Several data examples will be considered repeatedly throughout this book. These examples come from different behavioural research areas. In this chapter, the examples are introduced.

1.1 Example 1: Clinical psychology (depression prevention)

Reference of the study. This example is based on Veltman, Ruiter, and Hosman (1996).

Goal of the study. The study has evaluated a training program aimed to prevent depression.

Subjects. Participants were individuals aged between 15 and 19 from the Dutch cities of Nijmegen and Arnhem who were believed to be at risk of developing a depression.

Intervention. Some participants were given a training to reduce automatic negative thoughts (ANTs). An example of an ANT is the thought 'what's wrong with me' if one cannot stop the thought from recurring again and again. Automatic negative thoughts are an important cause of depression according to 'Beck's cognitive theory of depression' (Moilanen, 1993). The training therefore was hoped to reduce the risk of depression. The goal of the study was to evaluate whether this was actually achieved. The participants who did not receive the training served as a **control** group. The training will henceforth be called the **treatment**.

Measures. The degree of depression was assessed by the Beck Depression Inventory (BDI), which contains 21 multiple choice questions. In each question, the participant has to pick one out of four possible alternatives. An example of one question is this:

- ☐ I do not feel sad
- ☐ I feel sad
- ☐ I am sad all the time and I can't snap out of it
- ☐ I am so sad or unhappy that I can't stand it

The answer of the participant is scored as 0, 1, 2, or 3. For example, if the participant has picked the third alternative ('I am sad all the time and I can't snap out of it'), then the participant's score for that question would be 2. After completing the questionnaire, the scores of the participant are summed, yielding the participant's total score.
A similar questionnaire was used to measure the amount of automatic negative thoughts of the participant. The frequency of such thoughts was assessed by the Automatic Thoughts Questionnaire (ATQ) (Hollon & Kendall, 1980), which has 30 questions. An example of one of those questions is:

	not at all	sometimes	moderately often	often	all the time
My life is a mess	1	2	3	4	5

Each question is scored by the indicated number. The scores are summed per participant.

Repeated measures. Both the BDI and the ATQ were administered three times to each participant: before the treatment period, after the treatment period and after a follow-up. Only the first two measurements are displayed below. The batch of third measurements is available in the file *prevention.sav*.

Data. The data matrix is given in the table below (Table 1-1). The variables are:

Group = indicates whether the participant had the treatment (1 = treatment, 2 = control)

Bdi1 = the total score on BDI before the treatment period (BDI pretest)

Bdi2 = the total score on BDI after the treatment period (BDI posttest)

Atq1 = the total score on ATQ before the treatment period (ATQ pretest)

Atq2 = the total score on ATG after the treatment period (ATQ posttest).

Table 1-1

Group	*Bdi1*	*Bdi2*	*Atq1*	*Atq2*
1	22	4	80	40
1	13	3	57	38
1	15	4	89	47
1	19		70	53
1	23	8	79	56
1	12	3	68	38
1	15	7	63	55
1	18	22	77	104
1	14	3	70	47
1	12	1	91	32
1	16	7	50	40
1	21	3	69	40
1	23	11	117	80
1	11	7	72	56
1	16	21	71	73
1	20	33	95	95
1	13	6	50	47
1	16	18	103	94
1	14	10	52	57
1	15	14	78	70
1	16	8	99	73
1	12	7	68	48
1	11	1	80	35
1	15	6	47	49
1	11	6	58	45
1	11	6	55	51
1	14	14	93	73
1	12	10	54	44
1	12	9	66	49
1	14	11	59	54
1	21	13	67	45
1	17	27	67	72
1	12	14	55	79
1	12	8	70	61
1	21	19	110	91
1	21	9	66	70
1	13	3	56	39
1	25	4	92	36
1	14	7	61	45
1	14	6	84	65

1	19	19	103	108
1	22	14	99	78
1	19	22	97	90
1	20	14	91	69
1	19	25	85	87
1	11	5	57	57
1	12	15	63	68
1	10	19	65	77
1	19	16	94	90
2	11	5	69	52
2	11	5	58	52
2	13	0	43	33
2	12	5	60	51
2	22	17	58	61
2	12	3	44	32
2	10	13	53	53
2	10	5	36	43
2	12	5	47	39
2	11	9	67	57
2	18	3	79	56
2	14	15	61	58
2	10	10	58	58
2	16	0	66	84
2	15	16	55	66
2	10	13	46	59
2	16	15	89	92
2	11	4	58	31
2	13	12	50	45
2	16	13	82	62
2	11	0	47	38
2	22	22	87	100
2	16	14	64	58
2	13	9	67	62
2	15	11	84	50
2	10	6	70	62
2	19	8	99	83
2	11	12	64	59
2	14	12	85	81
2	11	5	38	41
2	12	13	69	63
2	20	8	88	80
2	10	10	50	51

2	12	6	58	55
2	15	9	64	68
2	13	11	58	67
2	15	8	53	61
2	18	26	72	72
2	16	12	43	66
2	11	7	87	41
2	20	16	70	59
2	10	16	35	38

1.2 Example 2: Education (arithmetic lesson)

Reference of the study. This example is based on Ellis and Van de Veerdonk (2016).

Goal of the study. This study has evaluated two different methods for teaching long division to students in intermediate vocational education.

Subjects. Participants were students of a school in the South-west of the Netherlands who were supposed to refresh their long-division skills.

Intervention. All participants followed a classroom lesson of one hour. Some classes were taught the classical long division algorithm. Other classes were taught the newer chunking method.

Measures. The proficiency at the long division was assessed before and after the lesson by two different though similar tests with 17 items each. The number of correct answers was converted into a grade between 1 and 10.

Data. The data matrix is given in the table below (Table 1-2). The variables are

> Algorithm = indicates whether the student learned the chunking or the classical long division
>
> GradeBefore = grade before the lesson (arithmetic pretest)
>
> GradeAfter = grade after the lesson (arithmetic posttest)

Table 1-2

Person	*Algorithm*	*GradeBefore*	*GradeAfter*
1	chunk	5.8	7.4
2	chunk	6.3	5.8
3	chunk	8.4	7.9
4	chunk	5.2	6.8
5	chunk	6.8	6.8
6	chunk	8.4	7.9
7	chunk	6.8	6.8
8	chunk	6.8	8.9
9	chunk	7.9	7.4
10	chunk	3.1	1.5
11	chunk	7.4	4.7
12	chunk	8.9	5.8
13	chunk	8.4	6.3
14	chunk	8.9	7.4
15	chunk	5.2	6.3
16	chunk	4.7	5.2
17	chunk	4.7	5.2
18	chunk	5.2	6.8
19	classical	4.7	5.2
20	classical	5.2	7.9
21	classical	7.9	4.7
22	classical	8.4	6.8
23	classical	6.8	8.4
24	classical	7.4	6.3
25	classical	8.4	4.7
26	classical	6.3	6.3
27	classical	5.2	5.2
28	classical	7.4	4.7
29	classical	7.9	5.8

1.3 Example 3: Human-computer interaction (mind reading)

Reference of the study. This example is based on the study of Ryan et al. (2011).

Goal of the study. This study evaluated whether a predictive spelling program can improve the performance of a brain-computer interface (BCI) in typing sentences. Some diseases cause patients to become completely paralysed (locked-in syndrome), which makes communication extremely difficult. The BCI2000 software uses electrophysiological brain responses (EEG) to provide alphanumeric character selection by which patients can communicate. It is based on the P300, which is a positive part of the EEG response that occurs approximately 300 ms after presentation of a meaningful stimulus.

Subjects. Participants were 24 able-bodied undergraduate university students.

Intervention. All participants completed two sessions on separate days. In each session they had to write a sentence of 58 characters with their EEG, using the BCI2000 software. In one session the session (PS), the BCI2000 software was extended with a predictive spelling application. In the other session (NS), the BCI2000 software was used without predictive spelling. The order of the sessions was counterbalanced. This means that half of the subjects had the sessions in the order PS-NS, and the other half of the subjects had the sessions in the order NS-PS, and this order was assigned randomly to the subjects.

Measures. In this example, we will use some but not all of the measures of Table 1 and 2 of Ryan et al. (2011). The first measure is accuracy, which is the number of correct selections divided by the total number of selections in the session. The second measure is the time needed to complete the sentence, measured in minutes.

Data. The data matrix is given in the table below (Table 1-3). The variables are

PS Acc = accuracy in the PS condition

NS Acc = accuracy in the NS condition

PS Comp = time (minutes) to complete the sentence in the PS condition

NS Comp = time (minutes) to complete the sentence in the NS condition

Table 1-3 *Part of data of Ryan et al. (2011). Reprinted by permission of the publisher (Taylor & Francis Ltd, http://www.tandfonline.com).*

Subject	*PS Acc*	*NS Acc*	*PS Comp*	*NS Comp*
1	96.88	95.31	7.80	12.70
2	88.89	87.50	9.00	17.90
3	70.00	88.16	24.00	22.70
4	79.59	89.86	10.92	17.15
5	91.89	92.65	11.00	23.58
6	87.18	95.31	8.67	15.90
7	91.67	100.00	8.90	14.40
8	81.13	87.50	14.47	16.10
9	80.95	70.83	10.40	23.90
10	82.35	98.33	10.10	11.90

11	80.00	91.18	19.15	23.70
12	77.59	82.50	20.20	27.90
13	82.22	93.94	11.25	16.40
14	94.29	77.17	8.75	25.20
15	91.18	95.31	9.25	17.50
16	94.29	85.25	10.40	18.20
17	72.50	77.23	17.75	35.25
18	91.89	100.00	7.30	11.50
19	100.00	100.00	7.65	14.40
20	96.88	91.18	8.70	18.60
21	86.67	91.43	13.40	24.45
22	57.58	83.67	16.95	29.30
23	67.07	96.77	22.45	21.60
24	94.44	84.15	9.80	24.50

1.4 Example 4: Criminology (reconviction)

Reference of the study. This example is based on Killias, Gilliéron, Villard, and Poglia (2010).

Goal of the study. To compare the effect of imprisonment versus community service on re-offending.

Subjects. The participants were 141 defendants who had been sentenced to a short custodial sentence in the Swiss Canton of Vaud between 1993 and 1995. However, some subjects were excluded from the analysis because they had died, emigrated, had been removed from the programme, and for various other reasons.

Intervention. Subjects were randomly assigned to either community service or prison. Studies like this, in which real sentences are randomly assigned, are extremely exceptional.

Measures. It was recorded whether the subject has had at least one new conviction after 11 years.

Data. The counts are presented in Table 1 of Killias et al. (2010), and in the text pertaining to that table. In the custodial group, 22 out of 38 subjects had a new conviction after 11 years. In the community service group, 41 out of 78 subjects had a new conviction after 11 years.

1.5 Example 5: Developmental psychology (bullying)

Reference of the study. Unknown. The source of these data were developmental psychologists of the Radboud University.

Goal of the study. To investigate the effect of bullying on social isolation.

Subjects. 42 schoolchildren.

Measures. The children were questioned as to one another's social positions within the group. On the basis of this, the children were categorised as either bully, victim or non-involved. Each child responded to a questionnaire in order to measure social isolation. A high score on this variable means that the child was experiencing a lot of social isolation.

Data.

Table 1-4

Position	*Social isolation*
non-involved	1.75
non-involved	0.25
non-involved	0.25
non-involved	1.25
non-involved	0.75
non-involved	0.5
non-involved	0.25
non-involved	1
non-involved	0.75
non-involved	1
non-involved	1.25
non-involved	0.25
non-involved	1.5
non-involved	0.5
non-involved	0.25
non-involved	0.25
non-involved	0.25
non-involved	0.5
non-involved	1.75
non-involved	0.25
bully	0.75
bully	0.25
bully	0.75
bully	0.5

bully	0.5
bully	0.5
bully	0.75
bully	0.25
bully	0.25
victim	1
victim	2
victim	0.75
victim	3.25
victim	0.5
victim	2.25
victim	3
victim	1
victim	1.75
victim	1.75
victim	0.25
victim	0.25
victim	1.75

1.6 Example 6: Social psychology (food consumption)

Reference of the study. This example is based on Van de Veer, Van Herpen, and Van Trijp (2015).

Goal of the study. To investigate the effect of outer appearance focus on food consumption.

Subjects. The study does not specify how the participants were initially sampled. Participants who did not follow the instructions and participants who indicated a strong dislike for M&Ms were excluded from the study. Also, a participant who was an outlier in mean M&M consumption was excluded. There remained 107 subjects.

Interventions. In the appearance focus condition, a mirror was positioned so that participants could view their face and upper body part. In a previous study it was found that this increases attention to outward appearance of the body. In the condition of no appearance focus, the mirrors were turned back to front. Participants were served a 300-ml milkshake that contained either 534 or 215 kcal. So, in total there were two focus conditions (appearance focus vs no appearance focus) and two conditions as to the milkshake's caloric content (low vs high). The design was crossed, which means that all 2 × 2 = 4 combinations of focus and caloric content were used. Participants were assigned randomly to one of these four conditions. The experimenters also manipulated the milkshake's label, but this will not be used in the data below.

Measures. Next, each participant was asked to watch and evaluate video fragments individually for 15 min. A bowl of M&Ms was placed within reach of each participant. The participants were free to eat the M&Ms. After the participant had left, the experimenter weighed the bowl and calculated the amount the participant had consumed.

Simulated data. The data matrix below (Table 1-5) is simulated so that the means and standard deviations are approximately equal to the means and standard deviations in Table 1 of Van de Veer et al. (2015). The sample size per cell is slightly different as to the actual data, but is taken to be equal here. The following variables are present:

> Focus = indicates whether the subject was either in an appearance-focus condition or not
>
> Calories = indicates whether the participant got a milkshake with low or high caloric content
>
> Intake = amount of M&Ms (g) consumed by the participant

Table 1-5

Focus	*Calories*	*Intake*
no appearance focus	low	44
no appearance focus	low	23
no appearance focus	low	28
no appearance focus	low	1
no appearance focus	low	1
no appearance focus	low	55
no appearance focus	low	1
no appearance focus	low	52
no appearance focus	low	61
no appearance focus	low	44
no appearance focus	low	1
no appearance focus	low	9
no appearance focus	low	70
no appearance focus	low	59
no appearance focus	low	2
no appearance focus	low	98
no appearance focus	high	24
no appearance focus	high	13

no appearance focus	high	17
no appearance focus	high	1
no appearance focus	high	7
no appearance focus	high	28
no appearance focus	high	1
no appearance focus	high	27
no appearance focus	high	33
no appearance focus	high	25
no appearance focus	high	1
no appearance focus	high	9
no appearance focus	high	36
no appearance focus	high	31
no appearance focus	high	10
no appearance focus	high	60
appearance focus	low	30
appearance focus	low	6
appearance focus	low	16
appearance focus	low	2
appearance focus	low	2
appearance focus	low	41
appearance focus	low	2
appearance focus	low	39
appearance focus	low	48
appearance focus	low	30
appearance focus	low	1
appearance focus	low	1
appearance focus	low	56
appearance focus	low	42
appearance focus	low	1
appearance focus	low	90
appearance focus	high	83
appearance focus	high	17
appearance focus	high	32
appearance focus	high	57
appearance focus	high	1
appearance focus	high	38
appearance focus	high	27
appearance focus	high	12
appearance focus	high	41
appearance focus	high	42
appearance focus	high	6
appearance focus	high	12

appearance focus	high	64
appearance focus	high	64
appearance focus	high	38
appearance focus	high	95

1.7 Example 7: Sociology (spirituality)

Reference of the study. This example is based on MacDonald et al. (2015).

Goal of the study. To develop a measurable, quantitative construct of spirituality that is valid across cultures.

Subjects. 4004 university students of eight countries (Canada, India, Japan, Korea, Poland, Slovakia, Uganda, US).

Measures. The information was obtained by questionnaires. This included questions about demographic variables (country, sex, age and religious affiliation) besides the Expressions of Spirituality Inventory-Revised (ESI-R), which is a self-report questionnaire with 32 items about religion, based on earlier studies. Several other variables were measured too, but will not be used here. Based on further analyses, responses from the ESI-R were combined into five subtotal scores per subject, called ‘dimensions’ or ‘subscales’.

Data. The data are available at *http://datahub.io/dataset/spirituality-across-cultures-and-languages-data*. The following variables are present in the data file esimain.sav:

Table 1-6

Variable	*Description*	*Type*
country	Country from where data came	Demographic
sex	Sex	Demographic
age	Age in years	Demographic
relaff	Religious Affiliation	Demographic
esi1cos	Item 1 of ESI-R	ESI-R item
esi2epd	Item 2 of ESI-R	ESI-R item
esi3ewb	Item 3 of ESI-R	ESI-R item
...	...	
esi31val	Item 31 of ESI-R	ESI-R item
esi32val	Item 32 of ESI-R	ESI-R item
esircos	Cognitive Orientation toward Spirituality (COS)	Subscale (dimension)
esirepd	Experiential/Phenomenological Dimension (EPD)	Subscale (dimension)
esirewb	Existential Well-Being (EWB)	Subscale (dimension)

esirpar	Paranormal Beliefs (PAR)	Subscale (dimension)
esirrel	Religiousness (REL)	Subscale (dimension)

The dataset is a large one, so only part of it is displayed here.

Table 1-7

country	*sex*	*age*	*relaff*	*esi1cos*	...	*esi32 val*	*esir cos*	*esir epd*	*esir ewb*	*esir par*	*esir rel*
Canada	female	21	Islam	3	...	3	15	16	15	16	16
Canada	female	19	Christian	4	...	2	18	10	17	13	18
Canada	female	21	Christian	2	...	4	11	10	20	19	10
Canada	female	20	Christian	3	...	3	17	11	13	23	12
Canada	female	18	Christian	4	...	4	19	12	20	18	20
...	...	...	...	...	...	...	...	...	...	...	...

In cases where we need an example with a small dataset, we will use only the COS items for a random sample of 20 persons, as given below (Table 1-8).

Table 1-8

country	*sex*	*age*	*relaff*	*esi1 cos*	*esi6 cos*	*esi11 cos*	*esi16 cos*	*esi21 cos*	*esi26 cos*
Canada	female	19	Christian	4	3	3	3	3	4
Canada	female	20	Other Religion	2	1	1	1	3	2
Canada	female	19	Christian	3	3	2	2	3	3
Canada	Male	19	Christian	3	3	3	2	4	3
Canada	Male	26	No Religion	1	3	1	2	2	1
USA	female	20	Christian	3	2	3	3	4	3
USA	female	20	Christian	2	3	2	2	3	3
USA	female	20	No Religion	1	3	2	3	2	3
USA	female	51	No Religion	4	4	4	4	4	4
India	female	23	Buddhist	2	3	2	1	2	3
India	Male	21	Hindu	2	1	0	2	0	2
India	Male	24	Hindu	1	1	0	1	1	1
Uganda	female	19	Christian	3	4	4	4	4	4
Uganda	Male	27	Christian	1	1	2	3	1	1
Poland	female	20	Christian	3	3	3	1	3	3
Japan	female	21	No Religion	3	3	3	2	2	0
Korea	female	19	Other Religion	2	0	0	2	1	0
Korea	female	45	Other Religion	4	4	3	3	3	3
Korea	female	23	Buddhist	3	4	3	4	4	4
Korea	Male	33	No Religion	2	1	1	2	2	3

1.8 Example 8: Ethology (the gaze of dogs)

Reference of the study. This example is based on Somppi et al. (2016).

Goal of the study. To determine how gaze fixation in domestic dogs is influenced by emotional facial expressions.

Subjects. 31 domestic dogs of different breeds, sex and age. Both pet dogs (n = 23) and kennel dogs (n = 8) were used.

Interventions. Prior to the experiment, the dogs were trained to lie still in front of a monitor and to put their jaws on a chin rest. During the experiment, the dogs were shown pictures of dogs and humans with different facial expressions. The facial expressions were either Neutral, Pleasant, or Threatening. With each expression, 10 pictures of dog faces and 10 pictures of human faces were shown. The order of the pictures was randomised. Each picture was shown 1500 ms.

Measures. The eye movements of the dogs were recorded with a contact-free eye-tracker. The raw gaze data were further processed to identify fixation points. The authors defined various Areas of Interest (AOI) in the viewed faces, such as eyes and mouth. Next, they computed how long the dog fixated its gaze on each AOI. This was compared with AOI's areal size.

Data. The data are available in the Supporting Information files of the article of Somppi et al. (2016). The data below (in Table 1-9) were obtained from their S1_dataset.xlsx file. This dataset contains a fixation score for the inner face, which was computed, as explained in the article, by the formula

$$\text{FixationScore\_Inner} = \frac{\text{TotalFixationTime\_Inner}}{\text{TotalFixationTime\_Face}} - \frac{\text{AOISize\_Inner}}{\text{AOISize\_Face}}$$

A positive score indicates that the dog fixated more upon the inner region of the face than one would expect considering the inner region's size if the gazing pattern was random. The dataset was reduced by the following transformations: (1) only the fixation scores for the inner face were selected; (2) aggregate per dog and stimulus category; (3) restructured with dog as case identifier; (4) rounded values to facilitate presentation. The resulting data are given in Table 1-9.

Table 1-9 *Fixation scores for the inner face (N = Neutral, P = Pleasant, T = Threatening).*

Dog	*Gender*	*Breed*	*Age*	*Dog N*	*Dog P*	*Dog T*	*Human N*	*Human P*	*Human T*
1	bitch	Australian kelpie	4.5	.0989	-.0166	.0667	.1558	.0600	-.1538
2	male	beagle	4.0	.0766	-.1297	.1246	.0712	.0295	-.1786
3	male	beagle	4.0	-.0386	.1175	.0900	-.0408	-.1103	-.1496
4	male	beagle	4.0	.2865	.1157	.2933	.1441	.2464	.2341
5	bitch	beagle	4.0	.0654	-.0757	.2467	.2006	.1791	.1766
6	male	beagle	4.0	-.2031	-.0527	.2381	-.0329	.1891	.0595
7	male	beagle	4.0	.0190	.0891	.0588	-.2224	.0048	.0241
8	male	beagle	4.0	.1365	.1845	.1257	.3111	.3596	.0061
9	bitch	beagle	4.0	.2953	.1040	.2287	.0054	.3835	-.0670
10	bitch	Swedish vallhund	5.0	.2186	.2228	-.0173	.1875	.0968	.0510
11	bitch	border collie	5.0	-.0422	-.0965	.1599	.0802	.0727	-.1453
12	bitch	great Pyrenees	6.0	.0286	.3148	.1075	-.1154	.0271	-.0342
13	bitch	Beauce shepherd	3.5	.2589	.1163	.1139	.0358	.2926	.1593
14	bitch	hovawart	5.5	.2359	.0839	.1005	.0666	.0953	.0286
15	bitch	rough collie	2.0	.1310	.1500	.0919	-.0050	.1785	.1911
16	male	lagotto romagnolo	3.5	.0277	.0875	.2726	.2188	.0653	-.0471
17	bitch	Beauce shepherd	1.5	.0065	.0279	.0863	.2988	.2642	.0399
18	bitch	mixed	1.0	.3314	.3637	.2885	.4436	.4375	.0772
19	male	Welsh corgi	6.5	-.0062	.0415	.2726	.1262	-.0363	.0614
20	male	border collie	6.5	-.0136	.1786	.2047	.0655	-.0925	-.0439
21	male	hovawart	5.0	.0966	-.0440	.0815	.1900	-.0411	-.0258
22	bitch	border collie	8.0	.1008	.0516	.1369	.0784	-.0534	-.0151
23	male	smooth collie	2.5	.2435	.2727	.3108	.2894	.1701	.0315
24	bitch	mixed	7.0	-.1118	.1199	.3025	.0121	-.0172	.0196
25	bitch	rough collie	1.5	.0662	-.0999	.1897	-.2031	.0043	-.0880
26	bitch	Beauce shepherd	3.0	.0410	-.0751	.1211	-.1593	-.1303	-.2191
27	bitch	border collie	5.0	.0487	.2554	.2493	.0642	.0276	.2758
28	male	smooth collie	6.5	.1566	.1720	.3996	.2005	.0970	.1532
29	bitch	border collie	2.0	.1548	.1585	.2696	.2301	.1190	.0240
30	male	Finnish lapphund	8.5	.1688	.0279	.1075	.0141	.1114	-.0298
31	bitch	border collie	8.0	.2381	.1053	.2122	.0376	-.0580	.0255

2 Basic report of one variable

2.1 Learning goals

In this chapter you will learn to make a basic report of one variable. A basic report contains the important statistical characteristics of the variable. When you publish a research article in a scientific journal, you will usually select parts of the information from the basic report. After studying this chapter, you should be able to make assignments like the example below.

Example Clinical Psychology (depression prevention)
Write a basic report about BDI before the treatment

2.2 Definition of a basic report of one variable

A **basic report** of one variable contains the following elements:

- design, degree of control, and name of the analysis;
- frequency distribution, *N*, and histogram;
- five-number summary, outliers, and modified boxplot;
- mean and standard deviation;
- indication of normality.

This will be explained in the rest of this chapter. We will make the above example assignment.

2.3 Design

The design specifies the variable that you report about. It contains:

- the name of the variable;
- the measurement level of the variable (qualitative or quantitative);
- the possible outcomes or levels of the variable, if known.

If the variable does not have a name yet, you should assign a name to it. Choose a specific name, such as ‘IQ’ or ‘Age’, not a generic name such as ‘score’, ‘value’, or ‘number’. Furthermore, if only a part of the data is being used, this should preferably be stated in the design.

Explanation: Name of the variable

In statistics we use systematically collected information. This information is often arranged in the form of a table, called the **data matrix**. Each row of the data matrix is called a **case** or a **subject**. Each column is called a **variable**. The collection of all subjects in the data matrix is the **sample**. The crosspoint of a row and a column is called a **cell**. In each cell of the data matrix you find a piece of information, which is called a **score** or an **outcome**.

In this chapter we report about one variable, but most research contains many variables. In the basic report, you need to identify the reported variable by giving it a unique name. It should be perfectly clear for the reader which variable you mean. A name such as 'score' is not clear enough. However, in statistic books, outside of basic reports, such generic names may be appropriate.

Explanation: Measurement level of the variable

A qualitative variable is a variable that has a small number of possible outcomes. The outcomes are called categories or levels. It is not required that the categories are somehow ordered. The categories are usually identified by either a piece of text, the category name, or by an integer number. A quantitative variable is a variable with numerical outcomes that indicate how much of a certain property is present.

For example, if you investigate Dutch, German, and French people, then the variable Nationality is a qualitative variable with levels Dutch, German, and French. However, in a data matrix these three nationalities may be coded as 1, 2, and 3. If these numbers are being used, they do not represent a quantity; French is not more Nationality than Dutch. In contrast, the variable Body weight is a quantitative variable. A higher score indicates that you have more of something.

A **binary** variable is a variable with two possible levels. The levels are usually coded as 0 and 1, or as 1 and 2. A binary variable can usually be viewed as both a qualitative and a quantitative variable. For example, suppose the variable is Marital status, with the levels 'Married' and 'Not married', coded as 1 and 0. On the one hand, this is only a small number of categories, so the variable is qualitative. On the other hand, you can rename the variable as Number of partners married to, which is a quantity – which incidentally is restricted to either 0 or 1 in most countries.

Another distinction is that a variable can be discrete or continuous. A **discrete** variable is a variable that has only integer numbers as possible outcomes. For example, if you measure Age in whole years, then Age is a discrete variable. A **continuous** variable is a variable of which the possible outcomes are an interval of real numbers. If you measure Age with infinite precision, then Age is a continuous variable. Measuring with infinite precision is of course impossible, and therefore no actual variable is really continuous. However, in practice the term 'continuous' is used also for variables that are measured with high precision in comparison to the observed range of scores. For example, if you measure Age in whole years that run from 9 to 99, you may choose to treat the variable as if it is continuous. This can be a matter of subjective judgement, and it is not required to make a choice between discrete and continuous in this phase of the basic report.

Many textbooks distinguish the following measurement levels: **nominal**, **ordinal**, **interval**, and **ratio**. A nominal variable is the same as a qualitative variable. Another name is categorical variable. An ordinal variable is a variable that provides a ranking of subjects. Interval and ratio variables are both forms of quantitative variables, but the distinction is not very important for this book.

Explanation: Possible levels or outcomes of the variable
For a qualitative variable, make a list of all categories and write this in the design. For a quantitative variable, specify the smallest and largest possible value. In both cases, describe all **possible** outcomes, not only the observed outcomes. So you can specify the design before you collect the data.

Example Clinical Psychology (depression prevention)
Variable = Bdi1 (quantitative, range 0 – 63)

2.4 Degree of control

In the degree of control you specify whether the variable was passively observed or actively manipulated.

Explanation: Manipulated and observed
The terms 'manipulated' and 'observed' are used here in meanings that are slightly different from what they mean in everyday life. A variable is **manipulated** if the investigator has created or changed that aspect of reality. A variable is **observed** if it describes a part of the reality that is not changed by the investigator. For example, if you investigate the effect of drinking coffee on the exam scores of students, and you give some people coffee and others water, then Drink {coffee, water} is manipulated and Exam score is observed.

Another example: Suppose that you have made a questionnaire with the question whether a person is a woman or a man. You have (presumably) not changed the gender of the person, so this variable is observed. I know: you created the questionnaire, and you changed reality by giving it to people, and you did not see with your own eyes the bodily manifestations of gender, and maybe some persons changed their gender, and you change the data matrix when you record the answers. Nevertheless, the variable is observed because you, the investigator, did not change the underlying reality.

Example Clinical Psychology (depression prevention)
Bdi1 is observed

2.5 Name of the analysis

The analysis that we are going to do may be called an **exploratory data analysis** (EDA), a term coined by Tukey (1977). We consider only **univariate** analyses here, meaning that we analyse one variable at a time. After having obtained the data, one should always start with EDA as a preparation for further analyses, in order to see whether there is nothing wrong or strange with the data. EDA is also needed if the reader of an article wants to know how the sample compares to a known population. Another reason for EDA – and maybe the most important one – is to generate hypotheses.

2.6 Frequency distribution

A frequency distribution is a table that contains the frequencies or counts of possible scores or intervals of possible scores. Use the following rules:

> Rule 1. Use intervals for quantitative, continuous variables. For qualitative variables and for discrete variables with a small number of possible scores, do not use intervals, but count each possible score.
>
> Rule 2. The intervals should have the same width.
>
> Rule 3. If possible, the width of the intervals should be chosen in such a way that there are at least seven intervals that are not empty (i.e. have a count larger than 0).
>
> Rule 4. If possible, the width of the intervals should be chosen in such a way that at least one interval has frequency larger than five.
>
> Rule 5. Also list empty intervals if they are surrounded by non-empty intervals.
>
> Rule 6. Try to let the intervals start and end with easy numbers. For example, an interval 1 – 5 is easier to digest for most readers than the interval 1.83 – 4.27.

Explanation: Choice of intervals

The rules stated above are only rules of thumb. Exceptions are possible and sometimes very well defensible, especially if the rules conflict with each other. However, in many cases it is wise to adhere to these rules.

The choice of the interval width is partially a subjective matter. What you want is that the frequency distribution is a summary that shows the reader clearly and easily where the scores of the variable are concentrated. If you choose very small intervals, then you get many intervals, but most of them will be empty or have at the most one observation. That is hardly a summary, it is more like a copy of the data. On the other hand, if you choose very wide intervals, then you will get only one or two intervals with large frequencies. That is not very informative, as you have lost most of the details. The best choice is somewhere between very small and very large intervals... It is comparable to summarising a book of 100 pages. A summary of 99 pages is useless, and a summary of two lines is useless as well.

If the sample becomes larger and larger, you can take smaller intervals while still increasing the expected count in each interval. On the basis of this, it is possible to construct formulas for a good interval width. You might be tempted to think that this solves the subjectivity, but then you are mistaken. There are many possible formulas for this, from which you would have to choose, which again is subjective.

History. According to Miller (2014), the term 'frequency distribution' was first used by Pearson (1895, p. 412).

Example Clinical Psychology (depression prevention)

Table 2-1

BDI before treatment	*Frequency*
10	8
11	13
12	13
13	7
14	8
15	8
16	9
17	1
18	3
19	6
20	4
21	4
22	4
23	2
24	0
25	1

2.7 Number of observations

The number of observations is the total count of all scores in the variable. Missing observations are not counted. The number of observations is denoted by N.

This is probably the easiest statistic of your life, but nevertheless there is one error that is frequently made: forgetting to mention N. Count on it, people want to know how large your sample was before they believe anything else.

Example Clinical Psychology (depression prevention)
$N = 91$
Explanation: This is the sum of the elements in the column Frequency.

2.8 Histogram

A histogram is a chart of the frequency distribution. Each frequency is shown as a bar. You are free to choose the orientation of the bars as horizontal or vertical, but vertical (in columns) is preferred. A histogram should be readable independently of the rest of the report. Therefore, it should contain appropriate labels on the axes: One axis should have the name of the variable, one axis should be labelled 'Frequency' or 'Count', and intervals should have a label that either indicates their boundary or their midpoint. Indicate the boundaries if the variable is discrete with multiple scores per interval. It is allowed to omit some of the interval labels in the middle if space so demands, provided that the pattern remains clear.

Explanation: Relation between frequency distribution and histogram
Whether the interval width is good, is, unfortunately, best seen after you've finished the histogram. If you feel that it has too few or too many intervals, make a new frequency distribution and a new histogram. The important thing is that you end up with a good summary.

Explanation: What to do with scores on the boundary of an interval
If the variable is continuous, then the probability is zero that a score falls exactly on a boundary. So it should not happen – or very rarely. Therefore, if the variable is continuous, you may label the intervals with their midpoints or with their boundaries, without explicating to which interval the boundary belongs. For example, with Age measured on a continuous scale, you may describe the intervals as 20 – 25, 25 – 30, 30 – 35, ... or as 22.5, 27.5, 32.5, ... without stating what happens with someone of exactly 25 years. Such a person will in fact not exist because 25 means exactly 25 years, and that moment has passed within a billionth of a billionth of a second.

If the variable is discrete, then in many cases you would count each possible score and not create intervals. However, if intervals are created, you must describe both the largest and the smallest scores that can be included in the interval. For example, with Age measured in years, you may describe intervals as 20 – 24, 25 – 29, 30 – 34, ...

Explanation: Histogram or bar chart
Outside this book, the term 'histogram' is commonly used only if the bars indicate counts of intervals, and the term 'bar chart' is used if the bars indicate counts of single scores. However, if you make the intervals small, you get single scores, and therefore I use histogram for both versions.

History. According to David (1995), the term ‘histogram’ was first used by Pearson (1895, p. 399). Ioannidis (2003) states that the idea of a bar chart dates back to at least 1786, in a book of the political economist Playfair.

Example Clinical Psychology (depression prevention)

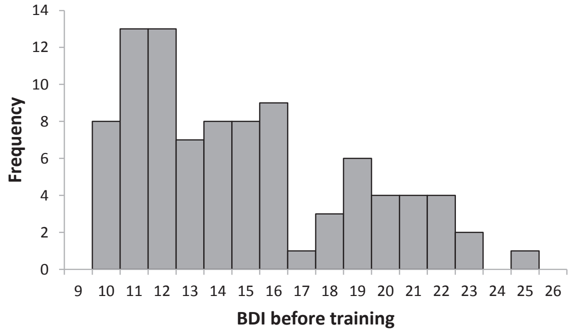


Figure 2-1

Explanation: It might be convenient to create a frequency distribution and a histogram at the same time. A simple method would be the following: Look at column Bdi1 in the data matrix. Find the lowest and the highest score, and use these to draw the horizontal axis, including the labels. Next, start with row 1, and go over the rows one by one. For each score, write an ‘x’ above that score in the chart.

14																
13		x	x													
12		x	x													
11		x	x													
10		x	x													
9		x	x				x									
8	x	x	x		x	x	x									
7	x	x	x	x	x	x	x									
6	x	x	x	x	x	x	x			x						
5	x	x	x	x	x	x	x			x						
4	x	x	x	x	x	x	x			x	x	x	x			
3	x	x	x	x	x	x	x		x	x	x	x	x			
2	x	x	x	x	x	x	x		x	x	x	x	x	x		
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Figure 2-2

Now, you can count the number of 'x' in each column and make the frequency table. Do not forget to complete the chart by adding the axis names 'Frequency' and 'BDI before training'. And maybe you want to make the chart artistically more gratifying by colouring the columns, but that is less essential.

Example Education (arithmetic lesson)

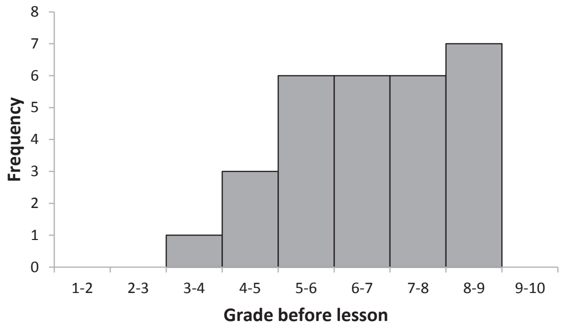


Figure 2-3

Example Sociology (spirituality)

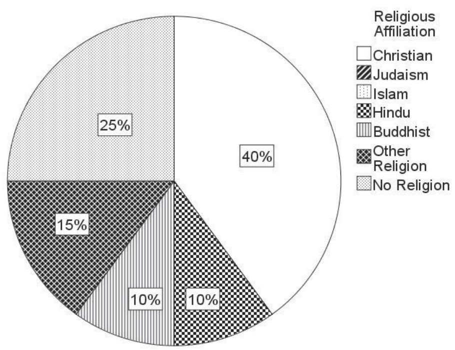


Figure 2-4

Explanation: This is a qualitative variable, and a histogram or bar chart would falsely suggest that there is some order in the religions. A pie chart like this might be a better option.

2.9 Five-number summary

The five-number summary lists, in this order: minimum, first quartile, median, third quartile, maximum. These statistics are obtained by first sorting the scores from small to large. The **minimum** is the smallest score in the sample. The **maximum** is the largest score in the sample. The **median** is the middle score (if N is odd), or the average of the two middle scores (if N is even). The **first quartile** is the median of the scores below the median. The **third quartile** is the median of the scores above the median. If the median is a datum, include it in both halves (this is debatable; see the explanation below). The following notation is often used for the five numbers: Min, Q_1, Md or Q_2, Q_3, Max.

Explanation: The definition of quartiles
The intended meaning of the quartiles is best described for continuous variables:

- The first quartile is the number that divides the 25% lowest scores from the 75% highest scores.
- The median is the number that divides the 50% lowest scores from the 50% highest scores.
- The third quartile is the number that divides the 75% lowest scores from the 25% highest scores.

For discrete variables, however, there is no universally agreed definition. The reason is that there are usually either no numbers with this property, or many numbers with this property, but seldom precisely one number with this property.

For example, take these scores:

11, 12, 13, 14

The number 12.5 divides the scores in two halves, which is what the median should do. However, the same is true for 12.6 and 12.9. Logically, every number larger than 12 and smaller than 13 may equally well be called ‘a median’. It is a convention that we pick only 12.5, but there is no rigorous math behind this convention.

Now, consider these scores:

11, 12, 13, 14, 15

The number 13 is clearly the middle score. But can you say that it divides the data into two halves of 50%? If you say that the lowest half is 11, 12 then this is 40%, not 50%; but if you say the lowest half is 11, 12, 13 then this is 60%, not 50%. It can never be 50%. Perhaps we should say that there is no median? It is a convention that we pick the middle number, here 13.

Even if we accept this convention for the median, the quartiles are still ill-defined, because one can debate whether the median itself should be included in both halves or not. In the last example, if we include 13 in the lowest half, then the first quartile is 12. If we do not include 13 in the lowest half, then the first quartile is 11.5.

History. The term ‘quartile’ was first used by Galton (1882, p. 245), according to David (1995). Galton was clearly aware that their *definition* had to be for infinite groups (p. 245, 247). The term ‘median’ was first used by Cournot in 1843, according to David (1998). The term ‘five number summary’ is probably due to Tukey (1977).

Example Clinical Psychology (depression prevention)

Table 2-2 *Five-number summary of BDI pretest.*

Min	Q_1	*Md*	Q_3	*Max*
10	12	14	18	25

Explanation: For sorting the scores, it may be convenient to use the frequency table or the histogram as a basis. If there are no intervals, they show the scores already in the right order. Compute the cumulative frequencies by adding the frequencies. So the cumulative frequency of score x is the number of subjects with a score less than or equal to x. Next, convert the cumulative frequencies into percentages by dividing them by N. The score with cumulative percentage 25% or just above that is the first quartile; the smallest score with cumulative percentage 50% or more is the median; the first score with cumulative percentage 75% or more is the third quartile.

Table 2-3

BDI before treatment	*Frequency*	*Cumulative Frequency*	*Cumulative Percentage*	
10	8	8	8.8%	*Min* = 10
11	13	21	23.1%	
12	13	34	37.4%	$Q_1 = 12$
13	7	41	45.1%	
14	8	49	53.8%	*Md* = 14
15	8	57	62.6%	
16	9	66	72.5%	
17	1	67	73.6%	
18	3	70	76.9%	$Q_3 = 18$
19	6	76	83.5%	
20	4	80	87.9%	
21	4	84	92.3%	
22	4	88	96.7%	
23	2	90	98.9%	
24	0	90	98.9%	
25	1	91	100.0%	*Max* = 25

The result is the same as with sorting: There are 91 scores. If you sort them from small (left) to large (right) then there are 45 scores to the left of the median, and 45 scores to the right the median, and the median is score number 46. From the cumulative frequencies you can infer that (a) the smallest 41 scores are all 13 or less, and (b) the smallest 49 scores are all 14 or less; so score number 46 must be one of the scores 14.

Example Education (arithmetic lesson)

Table 2-4

Min	Q_1	*Md*	Q_3	*Max*
3.1	5.2	6.8	7.9	8.9

Explanation: The histogram of this variable used intervals, so it does not contain sufficient information to determine the quartiles precisely. We have to sort the raw scores.

3.1	4.7	4.7	4.7	5.2	5.2	5.2	<u>5.2</u>	5.2	5.8	6.3	6.3	6.8	6.8	**6.8**
6.8	7.4	7.4	7.4	7.9	7.9	<u>7.9</u>	8.4	8.4	8.4	8.4	8.4	8.9	8.9	

The middle score (6.8) is made bold; this is the median. Now consider the scores to the left of it, including the bold score itself. Their middle score (5.2) is underlined; this is the first quartile. Similarly, to the right of it, the third quartile (7.9) is underlined. However, if you do not include the bold score in the upper half, then the third quartile would be (7.9 + 8.4) / 2 = 8.15. Both outcomes, Q_3 = 7.9 and Q_3 = 8.15, are correct. Indeed, both Excel and SPSS can report both 7.9 and 8.15 as the third quartile, depending on which version you ask for.

2.10 Outliers

In the basic report you have to list which scores are outliers. Outliers are scores that are extremely large or extremely small in comparison to other scores. The following rule is often applied. First compute the **interquartile range** (*IQR*):

$$IQR = Q_3 - Q_1$$

Next, compute these bounds:

$$\text{lower bound} = Q_1 - 1.5IQR$$

$$\text{upper bound} = Q_3 + 1.5IQR$$

Scores between the lower bound and upper bound are not outliers. Scores at or outside these bounds are **outliers**. That is, every score smaller than or equal to the lower bound is an outlier; and every score larger than or equal to the upper bound is an outlier; every other score is not an outlier. Do not remove outliers.

Explanation. Interquartile range
Although the interquartile range is not necessarily part of the basic report, it is quite important. It can be viewed as a measure of the amount of variability in the scores. It is called a measure of '*dispersion*'. In a histogram, it indicates how wide the middle 50% of the histogram (counting only the area in the columns) is.

Explanation: Bounds for outliers
There are many different rules for identifying outliers, and this is just one of them. Indeed, maybe it would be better if we said 'look at the histogram and see whether there are scores that are strangely large or small'.

Explanation: Do not remove outliers
It is a breach of data integrity if you remove outliers for the sole reason of being an outlier. It is even worse if you have the additional reason that your results will become better and if you do not describe this in your report; in that case removal of outliers comes close to academic fraud.

However, it is recommended to investigate outliers further and be sure that they are not caused by errors. And if an outlier turns out to be an error, then the score may be improved or removed because it is an error (not because it is an outlier). For example, if your variable is Year of birth, and some persons have recorded a date more than 200 years in the past, say in the year '202', you may notice them first as outliers, and then recognise they must be errors. However, if the outlier is a person aged 115, you cannot blindly dismiss it. There have been people of 115 years old, so it is possible. But it is rather unlikely, so you must verify that this age is not a typo. If the person is really 115, do not remove him or her. There are valid statistical techniques, such as Winsorising, that adjust outliers (Keselman et al., 2008), but one should be aware that these are only valid if all subsequent statistical analyses are adjusted too. Consult a statistician if you apply this.

History. According to David (1995), the term 'interquartile range' was first used by Galton (1882, p. 245). Miller (2014) states that the term 'outlier' became a statistical term in the early 20th century. The origin of the factor 1.5 is obscure. Its corresponding bound is called 'the inner fence', while the more extreme bound with factor 3 is called 'the outer fence' (Dawson, 2011).

Example Clinical Psychology (depression prevention)
IQR = 6 low bound for outliers = 3 high bound for outliers = 27 There are no outliers

2.11 Modified boxplot

A boxplot is a chart of the five-number summary. A modified boxplot is a chart of the five-number summary and outliers. A boxplot is made as follows: on a number line (axis), draw a line from *Min* to *Max*, and draw a box from Q_1 to Md and a box from Md to Q_3 .For a **modified** boxplot, you need the largest score in the data that is not an outlier, and the smallest score in the data that is not an outlier. Denote these scores as *Max** and *Min**. Please note that these are generally not equal to the lower bound and upper bound of the previous section, because *Min** and *Max** should be actually existing scores in the data. On a number line (axis), draw a line from *Min** to *Max**, and draw a box from Q_1 to Md and a box from Md to Q_3 . Next, plot the outliers individually as stars.

History. The term 'box plot' was first used by Tukey in 1970, according to David (1995). See McGill, Tukey and Larsen (1978) for modified versions of boxplots, and also Benjamini (1988).

Example Clinical Psychology (depression prevention)

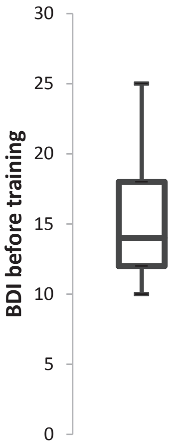


Figure 2-5

Comment: Consider what the modified boxplot would be if the maximum score 25 was changed into 30. This would not change the quartiles, so the bound for high outliers would remain 27. Then 30 is an outlier, and it is indicated by a separate symbol such as a star. Note that the vertical line (the whisker) does not stop at the bound 27, but at the lowest actual score below that, which is 23.

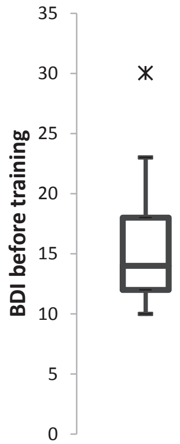


Figure 2-6

2.12 Mean and standard deviation

The **mean** indicates the middle point of the scores. The **standard deviation** indicates how large the distances between scores of different subjects are, on average. It also indicates how large the distances between individual scores and the mean are, on average. In other words, it is a measure of variation.

It is best to calculate both mean and standard deviation with the statistical functions of a calculator or with a statistical program. However, one should be aware that there are **two different standard deviations**, and you should pick the right button. One standard deviation is for the case that you have only data from a sample, which is always the case in empirical research. The other standard deviation is for the case that you have all data from a population. This happens mostly in theoretical exercises. The sample standard deviation uses $N-1$ and is usually denoted as S or SD or s. The population standard deviation uses N and is usually denoted by the Greek letter σ. The mean is denoted as $\bar{X}$ or M or m.

Explanation: Why a standard deviation?
The standard deviation describes the variability of scores. This is not described by the mean. For example, consider the grades of two classes on an exam:

Class A: 4, 4, 4, 5, 6, 6, 6

Class B: 1, 1, 1, 5, 9, 9, 9

Both classes have mean 5. But the differences are much larger in class B. This is reflected in the standard deviation. Class A has $S = 1$ and class B has $S = 4$.

There are many jokes where statisticians arrive at stupid conclusion because they consider only the mean. However, in reality the opposite is true. A statistician would never look at only the mean. Statistics is all about variation.

Explanation: What is a variance?
The **variance** of a variable is the square of the standard deviation. So, if the standard deviation is 4, then the variance is 16. It is not necessary to list the variance in a basic report. However, in some statistical discussions it is customary to use the variance rather than the standard deviation. We will encounter such discussions later on in the book. In that case, remember that variance and standard deviation are similar concepts. They both measure the degree of variation in the scores, but on different scales. Compare this: If you want to measure the size of circles, you can use the diameter or you can use the surface of the circles. These are different numbers, but they are logically related, and both may be used as a measure of size.

History. The concept of a mean dates back to Pythagoras, as described by Miller (2014). The term 'standard deviation' was first used by Pearson (1894, p. 80) and the term 'variance' was coined by Fisher (1918a p. 399), both according to David (1995). However the concept itself must already have been known much earlier by Gauss.

Example Clinical Psychology (depression prevention)	
$m =$	14.88
$s =$	3.89
Example Education (arithmetic lesson)	
$m =$	6.71
$s =$	1.55

2.13 Indication of normality

In this section of the report you indicate whether the variable has approximately a normal distribution. This means:

- The histogram is approximately symmetric.
- There is a small percentage of outliers.

You may infer this from either the histogram or the modified boxplot. For small values of *N*, it is best to say that the sample is too small to judge normality. In this book we will set the bound for that at $N < 20$, but be aware that that is quite arbitrarily.

Explanation: The normal distribution
The normal definition is rigorously defined for continuous variables in infinite populations. We will not use this definition, but in case you want to know:

> **Definition**. A variable has a **standard normal distribution** if its histogram in the population is proportional with $e^{-0.5x^2}$. A variable Y has a **normal distribution** if it can be obtained as $Y = a + bX$, where X has a standard normal distribution and a and b are real numbers with $b > 0$.

The histogram of a normal distribution is neatly bell-shaped, symmetric, with light tails, as shown in Figure 2-7. However, you will never encounter it in your data, because you will have only finite samples and not infinite populations. Moreover, in many cases the sample size is too small to assess reliably whether the shape is bell-shaped.

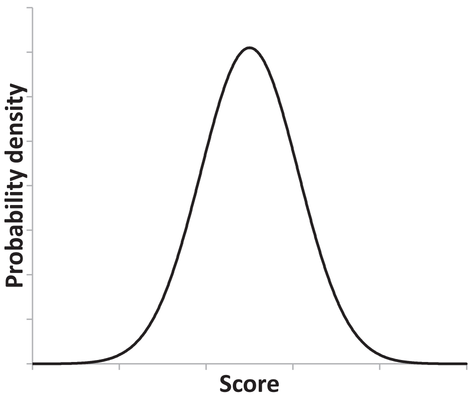


Figure 2-7

Some important statistical procedures (notably, the *t*-tests) assume that the variable is normally distributed. However, these procedures can still be good if *N* is not too small and the variable is approximately symmetric with a small percentage of outliers. Therefore, we will call this ‘approximately normal’. One should be aware, however, that other authors may use that term less loosely.

Explanation: How to judge approximate normality

There are no generally approved rules for evaluating whether a variable is normally distributed. It depends in part on the follow-up statistical procedures that you want to apply. Here, we will assume that you want to judge whether the variable may be used in a *t*-test. In that case, it is often good enough if the variable is not very skewed and that there are not many outliers, and that the outliers are limited in size. You can judge this from the modified boxplot. If the variable is normally distributed, the modified boxplot looks like Figure 2-8, with about 1% outliers. Note that in samples from a normal distribution, about 30% of the samples will have outliers (Dawson, 2011). Thus, it is quite normal to have outliers, even under a normal distribution. Normality is suspect only if the number or magnitude of outliers is unusually large.

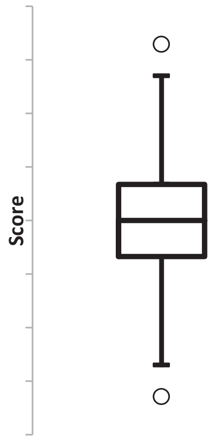


Figure 2-8

Example Clinical Psychology (depression prevention)
The histogram of Bdi1 (BDI before treatment) is not approximately normal

2.14 Exercises

Exercise 2.1

Using the data of Example 1: Clinical Psychology (depression prevention), write a basic report for the variable Bdi2 (BDI after training).

Exercise 2.2

Using the data of Example 2: Education (arithmetic lesson), make a basic report for the variable GradeAfterLesson.

Exercise 2.3
Using the data of Example 3: Human-computer interaction (mind reading), make a basic report for accuracy in the predictive spelling condition.

Exercise 2.4
Using the data of Example 5: Developmental psychology (bullying), make a basic report for social isolation in the non-involved group.

Exercise 2.5
Using the data of Example 6: Social psychology (food consumption), make a basic report for the amount of M&M's consumed, all four groups combined.

Exercise 2.6
Using the small data set of Example 7: Sociology (spirituality), make a basic report for the first item of the COS-scale of the ESI-R.

Exercise 2.7
Using the data of Example 8: Ethology (the gaze of dogs), make a basic report for the fixation score on inner faces when the dogs watched a neutral dog.

Exercise 2.8
People with manic-depression have intervals in which they are manic, alternating with periods of depression. It has been suggested that mania increases creativity. Weisberg (1994) investigated this for the composer Robert Schumann. Based on his diaries, earlier investigators had concluded that Schumann had manic-depression, and that there were many years in which he had a single dominant mood – either mainly depressed or mainly hypomanic. Other investigators had separately determined how many compositions he made in each year, and what their quality was. Table 2-5 is reproduced from Weisberg (1994, p. 363).

a. Create histogram, five-number summary, and modified boxplot of the number of compositions in Schumann's years with hypomania.
b. Do the same for the years with depression.
c. Describe the most important similarities and dissimilarities between the years with depression and the years with hypomania. This should be based on the boxplots. The description has to consist of proper English sentences, literate and precise enough to be part of a scientific article.

Table 2-5 *Reproduced from Weisberg (1994, p. 363), © 1994 American Psychological Society reprinted by permission of SAGE Publications, Inc.*

Year		*Number of compositions*
		Depression
1830		1
1831		1
1839		4
1842		3
1844		0
1847		5
1848		5
	Average	2.7
		Hypomania
1829		1
1832		4
1840		25
1843		2
1849		28
1851		16
	Average	12.7

Exercise 2.9

a. Make a histogram and determine the five-number summary for the following sequence of numbers.

1	3	5	1	3	1
14	1	5	18	12	3

b. How certain are you of the value of the third quartile? Are there other values that could reasonably be reported as the outcome?

c. Are there any outliers?

d. Draw an arrow that points to the position of the median in the histogram. Do the same for the first and the third quartile. Explain why in some cases a quartile is not defined exactly.

e. Is the distribution skewed or symmetric? How can you infer this from the histogram? And how can you infer this from the boxplot?

Exercise 2.10

The following data are a subset of data obtained in an investigation of conversion disorders conducted by Elsbeth Nauta and Karin Roelofs. Conversion disorders are characterised by neurological symptoms such as numbness and paralysis without any clear organic cause. It is supposed that psychological stress can play a part in the development of these disorders. In this investigation, people with conversion disorders were compared with people having an affective disorder like fear or depression. This latter group will be called the control group. The theory is that conversion disorders are a consequence of negative emotional experiences at a young age. All participants in this research answered a questionnaire about their youth. On the basis of their answers, the following variables were computed for each person: Emotional neglect, emotional abuse, physical abuse, sexual harassment, sexual abuse. These variables will be called *trauma variables*. Figure 2-9 displays the boxplot for each of the trauma variables in each of the groups. The variables are sorted per group in the same order as they are displayed in the legend. All scores are between 0 and 12. The conversion group contained 26 persons, and the control group contained 16 persons. All of these persons had scores on all five trauma variables.

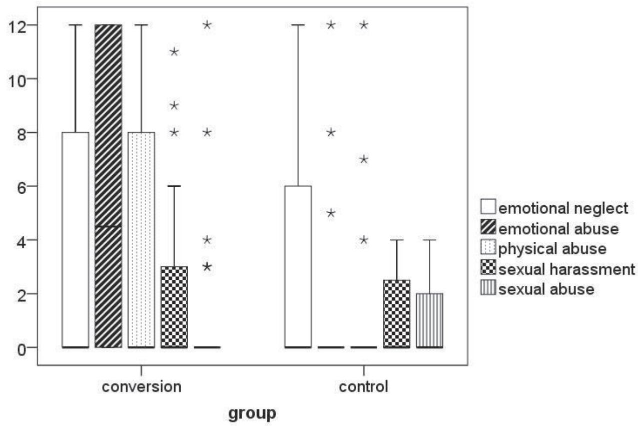


Figure 2-9

a. Explain why there is no box displayed for emotional abuse and physical abuse in the control group. Beware that the chart was produced with a statistical program.
b. On which trauma variables is the difference between the conversion group and the control group most pronounced? Describe the main differences between the two groups.
c. To what extent do these data give any reason to assume that emotional abuse will lead to the development of a conversion disorder?
d. Same question for emotional neglect.
e. Same question for sexual abuse.
f. To what extent may it be concluded that persons with a traumatic experience in their youth are more likely to develop a psychological disorder?

Exercise 2.11

The data of this exercise are obtained by Manon Boeschoten, Chiel Luytelaar, Kim Stael and Marielle Gorissen, using the gating paradigm. If people get to hear a click in an experiment, then on average across a large number of trials, there is a small peak in the central EEG after about 50 ms. The size of this peak is called the *P50 amplitude*. If two clicks occur shortly after each other, then normal persons have a smaller P50 after the second click. One theory is that this reflects a filter process in the brain. This filter process would be less effective in persons with schizophrenia, and as a result their P50 after the second click would be approximately of the same size as it was after the first click.

The data of Table 2-6 are obtained from such an experiment. The variable Diagnosis contains the diagnosis of the patient, as assessed by a psychiatrist. The variables Stim1 and Stim2 contain the P50 after the first and second click, respectively, both averaged across about 60 trials per person. The variable Difference contains the differences Stim2 – Stim1.

a. Make a frequency table and a histogram for the variable Difference in the group with schizophrenia.
b. Create a five-number summary for schizophrenic and normal persons separately, and a modified boxplot for schizophrenic and normal persons separately. Draw the boxplots side-by-side (next to each other).
c. Discuss whether the data confirm the theory. Here, and everywhere in this book, the word discuss means that you have to write down the pro's and con's. The discussion should have the precision and nuance that can be expected from a scientist. So, a simple yes or no will certainly not suffice. Use all outcomes of the five-number summaries.
d. Make a basic report for schizophrenic and normal persons separately.
e. Does the group with the largest mean also have the largest median? Does the group with the largest IQR also have the largest standard deviation? Is this an exception, or something that happens often but not always, or a logical necessity?

Table 2-6

Diagnosis	*Sex*	*Stim1*	*Stim2*	*Difference*	*Rdif*	*Pdif*	*Ndif*
schizophrenia	man	8.04	2.24	-5.80			
schizophrenia	man	2.07	3.74	1.67			
schizophrenia	man	4.63	1.59	-3.04			
schizophrenia	man	1.22	1.56	.34			
schizophrenia	man	3.40	2.26	-1.14			
schizophrenia	man	1.41	2.86	1.45			
schizophrenia	man	4.60	2.94	-1.66			
schizophrenia	man	2.37	2.47	.10			

schizophrenia	man	4.48	8.94	4.46			
schizophrenia	man	7.89	6.18	-1.71			
schizophrenia	man	2.10	1.20	-.90			
schizophrenia	woman	1.79	2.52	.73			
schizophrenia	woman	5.69	2.96	-2.73			
depression	man	11.89	5.27	-6.62			
depression	man	7.14	8.87	1.73			
depression	man	5.32	4.37	-.95			
depression	man	5.54	9.72	4.18			
depression	man	11.25	8.11	-3.14			
depression	woman	3.51	4.21	.70			
depression	woman	4.25	4.03	-.22			
depression	woman	3.07	1.69	-1.38			
depression	woman	1.79	2.63	.84			
depression	woman	22.79	11.09	-11.70			
depression	woman	10.50	7.52	-2.98			
depression	woman	7.19	6.79	-.40			
depression	woman	13.14	3.48	-9.66			
depression	woman	3.77	1.56	-2.21			
normal	man	2.03	1.50	-.53	?	?	?
normal	man	10.89	10.79	-.10	?	?	?
normal	man	9.59	5.23	-4.36	?	?	?
normal	man	3.38	1.21	-2.17	?	?	?
normal	man	9.87	2.96	-6.91	?	?	?
normal	man	9.69	5.33	-4.36	36.96	34.78	-.3912
normal	man	8.41	3.79	-4.62	26.09	23.91	-.7091
normal	man	8.75	3.24	-5.51	13.04	10.87	-1.233
normal	man	3.54	1.72	-1.82	69.57	67.39	.4507
normal	man	10.84	6.88	-3.96	43.48	41.30	-.2197
normal	man	8.49	4.63	-3.86	52.17	50.00	.0000
normal	woman	2.90	2.54	-.36	91.30	89.13	1.2335
normal	woman	1.67	1.30	-.37	86.96	84.78	1.0272
normal	woman	7.89	2.74	-5.15	17.39	15.22	-1.027
normal	woman	7.70	2.02	-5.68	8.70	6.52	-1.512
normal	woman	8.98	8.54	-.44	82.61	80.43	.8573
normal	woman	4.27	7.35	3.08	100.00	97.83	2.0191
normal	woman	7.31	2.87	-4.44	30.43	28.26	-.5751
normal	woman	7.39	4.40	-2.99	60.87	58.70	.2197
normal	woman	2.19	1.48	-.71	73.91	71.74	.5751
normal	woman	5.87	1.95	-3.92	47.83	45.65	-.1092
normal	woman	9.87	4.79	-5.08	21.74	19.57	-.8573
normal	woman	6.73	3.22	-3.51	56.52	54.35	.1092