Chapter 1

Data Collection

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called **data**. Statistics is the study of how best to collect, analyze, and draw conclusions from data. It is helpful to put statistics in the context of a general process of investigation:

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics as a subject focuses on making stages 2-4 objective, rigorous, and efficient. That is, statistics has three primary components: How best can we collect data? How should it be analyzed? And what can we infer from the analysis?

Researchers from a wide array of fields have questions or problems that require the collection and analysis of data. Let's consider three examples.

- Climate scientists: how will the global temperature change over the next 100 years?
- Psychology: can a simple reminder about saving money cause students to spend less?
- Political science: what fraction of Americans approve of the job Congress is doing?

While the questions that can be posed are incredibly diverse, many of these investigations can be addressed with a small number of data collection techniques, analytic tools, and fundamental concepts in statistical inference.

This chapter focuses on collecting data. We'll discuss basic properties of data, common sources of bias that arise during data collection, and several techniques for collecting data through both sampling techniques and experiments. After finishing this chapter, you will have the tools for identifying weaknesses and strengths in data-based conclusions, tools that are essential to be an informed citizen and a savvy consumer of information.

1.1 Case study: using stents to prevent strokes

Section 1.1 introduces a classic challenge in statistics: evaluating the efficacy of a medical treatment. Terms in this section, and indeed much of this chapter, will all be revisited later in the text. The plan for now is simply to get a sense of the role statistics can play in practice.

In this section we will consider an experiment that studies effectiveness of stents in treating patients at risk of stroke.[1] Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death. Many doctors have hoped that there would be similar benefits for patients at risk of stroke. We start by writing the principal question the researchers hope to answer:

Does the use of stents reduce the risk of stroke?

The researchers who asked this question collected data on 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

Treatment group. Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.

Control group. Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.

Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment. The results of 5 patients are summarized in Table 1.1. Patient outcomes are recorded as "stroke" or "no event", representing whether or not the patient had a stroke at the end of a time period.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Table 1.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Table 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study. For instance, to identify the number of patients in the treatment group who had a stroke

[1] Chimowitz MI, Lynn MJ, Derdeyn CP, et al. 2011. Stenting versus Aggressive Medical Therapy for Intracranial Arterial Stenosis. New England Journal of Medicine 365:993-1003. http://www.nejm.org/doi/full/10.1056/NEJMoa1105335. NY Times article reporting on the study: http://www.nytimes.com/2011/09/08/health/research/08stent.html.

within 30 days, we look on the left-side of the table at the intersection of the treatment and stroke: 33.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Table 1.2: Descriptive statistics for the stent study.

⊙ **Guided Practice 1.1** Of the 224 patients in the treatment group, 45 had a stroke by the end of the first year. Using these two numbers, compute the proportion of patients in the treatment group who had a stroke by the end of their first year. (Please note: answers to all in-text exercises are provided using footnotes.)[2]

We can compute summary statistics from the table. A **summary statistic** is a single number summarizing a large amount of data.[3] For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: $45/224 = 0.20 = 20\%$.

Proportion who had a stroke in the control group: $28/227 = 0.12 = 12\%$.

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would *reduce* the rate of strokes. Second, it leads to a statistical question: do the data show a "real" difference between the groups?

This second question is subtle. Suppose you flip a coin 100 times. While the chance a coin lands heads in any given coin flip is 50%, we probably won't observe exactly 50 heads. This type of fluctuation is part of almost any type of data generating process. It is possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance. So what we are really asking is the following: is the difference so large that we should reject the notion that it was due to chance?

While we don't yet have our statistical tools to fully address this question on our own, we can comprehend the conclusions of the published analysis: there was compelling evidence of harm by stents in this study of stroke patients.

Be careful: do not generalize the results of this study to all patients and all stents. This study looked at patients with very specific characteristics who volunteered to be a part of this study and who may not be representative of all stroke patients. In addition, there are many types of stents and this study only considered the self-expanding Wingspan stent (Boston Scientific). However, this study does leave us with an important lesson: we should keep our eyes open for surprises.

[2] The proportion of the 224 patients who had a stroke within 365 days: $45/224 = 0.20$.

[3] Formally, a summary statistic is a value computed from the data. Some summary statistics are more useful than others.

1.2 Data basics

Effective presentation and description of data is a first step in most analyses. This section introduces one structure for organizing data as well as some terminology that will be used throughout this book.

1.2.1 Observations, variables, and data matrices

Table 1.3 displays rows 1, 2, 3, and 50 of a data set concerning 50 emails received during early 2012. These observations will be referred to as the `email50` data set, and they are a random sample from a larger data set that we will see in Section 2.3.

Each row in the table represents a single email or **case**.[4] The columns represent characteristics, called **variables**, for each of the emails. For example, the first row represents email 1, which is a not spam, contains 21,705 characters, 551 line breaks, is written in HTML format, and contains only small numbers.

In practice, it is especially important to ask clarifying questions to ensure important aspects of the data are understood. For instance, it is always important to be sure we know what each variable means and the units of measurement. Descriptions of all five email variables are given in Table 1.4.

	spam	num_char	line_breaks	format	number
1	no	21,705	551	html	small
2	no	7,011	183	html	big
3	yes	631	28	text	none
⋮	⋮	⋮	⋮	⋮	⋮
50	no	15,829	242	html	small

Table 1.3: Four rows from the `email50` data matrix.

variable	description
`spam`	Specifies whether the message was spam
`num_char`	The number of characters in the email
`line_breaks`	The number of line breaks in the email (not including text wrapping)
`format`	Indicates if the email contained special formatting, such as bolding, tables, or links, which would indicate the message is in HTML format
`number`	Indicates whether the email contained no number, a small number (under 1 million), or a large number

Table 1.4: Variables and their descriptions for the `email50` data set.

The data in Table 1.3 represent a **data matrix**, which is a common way to organize data. Each row of a data matrix corresponds to a unique case, and each column corresponds to a variable. A data matrix for the stroke study introduced in Section 1.1 is shown in Table 1.1 on page 2, where the cases were patients and there were three variables recorded for each patient.

Data matrices are a convenient way to record and store data. If another individual or case is added to the data set, an additional row can be easily added. Similarly, another column can be added for a new variable.

[4] A case is also sometimes called a **unit of observation** or an **observational unit**.

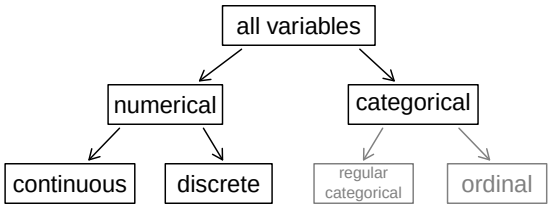

Figure 1.5: Breakdown of variables into their respective types.

⊙ **Guided Practice 1.2** We consider a publicly available data set that summarizes information about the 3,143 counties in the United States, and we call this the `county` data set. This data set includes information about each county: its name, the state where it resides, its population in 2000 and 2010, per capita federal spending, poverty rate, and five additional characteristics. How might these data be organized in a data matrix? Reminder: look in the footnotes for answers to in-text exercises.[5]

Seven rows of the `county` data set are shown in Table 1.6, and the variables are summarized in Table 1.7. These data were collected from the US Census website.[6]

1.2.2 Types of variables

Examine the `fed_spend`, `pop2010`, `state`, and `smoking_ban` variables in the `county` data set. Each of these variables is inherently different from the other three yet many of them share certain characteristics.

First consider `fed_spend`, which is said to be a **numerical** variable since it can take a wide range of numerical values, and it is sensible to add, subtract, or take averages with those values. On the other hand, we would not classify a variable reporting telephone area codes as numerical since their average, sum, and difference have no clear meaning.

The `pop2010` variable is also numerical, although it seems to be a little different than `fed_spend`. This variable of the population count can only take whole non-negative numbers (`0`, `1`, `2`, ...). For this reason, the population variable is said to be **discrete** since it can only take numerical values with jumps. On the other hand, the federal spending variable is said to be **continuous**.

The variable `state` can take up to 51 values after accounting for Washington, DC: `AL`, ..., and `WY`. Because the responses themselves are categories, `state` is called a **categorical** variable,[7] and the possible values are called the variable's **levels**.

Finally, consider the `smoking_ban` variable, which describes the type of county-wide smoking ban and takes values `none`, `partial`, or `comprehensive` in each county. This variable seems to be a hybrid: it is a categorical variable but the levels have a natural ordering. A variable with these properties is called an **ordinal** variable. To simplify analyses, any ordinal variables in this book will be treated as categorical variables.

[5] Each county may be viewed as a case, and there are eleven pieces of information recorded for each case. A table with 3,143 rows and 11 columns could hold these data, where each row represents a county and each column represents a particular piece of information.

[6] http://quickfacts.census.gov/qfd/index.html

[7] Sometimes also called a **nominal** variable.

	name	state	pop2000	pop2010	fed_spend	poverty	homeownership	multiunit	income	med_income	smoking_ban
1	Autauga	AL	43671	54571	6.068	10.6	77.5	7.2	24568	53255	none
2	Baldwin	AL	140415	182265	6.140	12.2	76.7	22.6	26469	50147	none
3	Barbour	AL	29038	27457	8.752	25.0	68.0	11.1	15875	33219	none
4	Bibb	AL	20826	22915	7.122	12.6	82.9	6.6	19918	41770	none
5	Blount	AL	51024	57322	5.131	13.4	82.0	3.7	21070	45549	none
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Washakie	WY	8289	8533	8.714	5.6	70.9	10.0	28557	48379	none
3143	Weston	WY	6644	7208	6.695	7.9	77.9	6.5	28463	53853	none

Table 1.6: Seven rows from the `county` data set.

variable	description
`name`	County name
`state`	State where the county resides (also including the District of Columbia)
`pop2000`	Population in 2000
`pop2010`	Population in 2010
`fed_spend`	Federal spending per capita
`poverty`	Percent of the population in poverty
`homeownership`	Percent of the population that lives in their own home or lives with the owner (e.g. children living with parents who own the home)
`multiunit`	Percent of living units that are in multi-unit structures (e.g. apartments)
`income`	Income per capita
`med_income`	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older
`smoking_ban`	Type of county-wide smoking ban in place at the end of 2011, which takes one of three values: `none`, `partial`, or `comprehensive`, where a `comprehensive` ban means smoking was not permitted in restaurants, bars, or workplaces, and `partial` means smoking was banned in at least one of those three locations

Table 1.7: Variables and their descriptions for the `county` data set.

Example 1.3 Data were collected about students in a statistics course. Three variables were recorded for each student: number of siblings, student height, and whether the student had previously taken a statistics course. Classify each of the variables as continuous numerical, discrete numerical, or categorical.

The number of siblings and student height represent numerical variables. Because the number of siblings is a count, it is discrete. Height varies continuously, so it is a continuous numerical variable. The last variable classifies students into two categories – those who have and those who have not taken a statistics course – which makes this variable categorical.

Guided Practice 1.4 Consider the variables `group` and `outcome` (at 30 days) from the stent study in Section 1.1. Are these numerical or categorical variables?[8]

1.2.3 Relationships between variables

Many analyses are motivated by a researcher looking for a relationship between two or more variables. A social scientist may like to answer some of the following questions:

(1) Is federal spending, on average, higher or lower in counties with high rates of poverty?

(2) If homeownership is lower than the national average in one county, will the percent of multi-unit structures in that county likely be above or below the national average?

(3) Which counties have a higher average income: those that enact one or more smoking bans or those that do not?

To answer these questions, data must be collected, such as the `county` data set shown in Table 1.6. Examining summary statistics could provide insights for each of the three questions about counties. Additionally, graphs can be used to visually summarize data and are useful for answering such questions as well.

Scatterplots are one type of graph used to study the relationship between two numerical variables. Figure 1.8 compares the variables `fed_spend` and `poverty`. Each point on the plot represents a single county. For instance, the highlighted dot corresponds to County 1088 in the `county` data set: Owsley County, Kentucky, which had a poverty rate of 41.5% and federal spending of $21.50 per capita. The scatterplot suggests a relationship between the two variables: counties with a high poverty rate also tend to have slightly more federal spending. We might brainstorm as to why this relationship exists and investigate each idea to determine which is the most reasonable explanation.

Guided Practice 1.5 Examine the variables in the `email50` data set, which are described in Table 1.4 on page 4. Create two questions about the relationships between these variables that are of interest to you.[9]

The `fed_spend` and `poverty` variables are said to be associated because the plot shows a discernible pattern. When two variables show some connection with one another, they are called **associated** variables. Associated variables can also be called **dependent** variables and vice-versa.

[8] There are only two possible values for each variable, and in both cases they describe categories. Thus, each are categorical variables.

[9] Two sample questions: (1) Intuition suggests that if there are many line breaks in an email then there would tend to also be many characters: does this hold true? (2) Is there a connection between whether an email format is plain text (versus HTML) and whether it is a spam message?

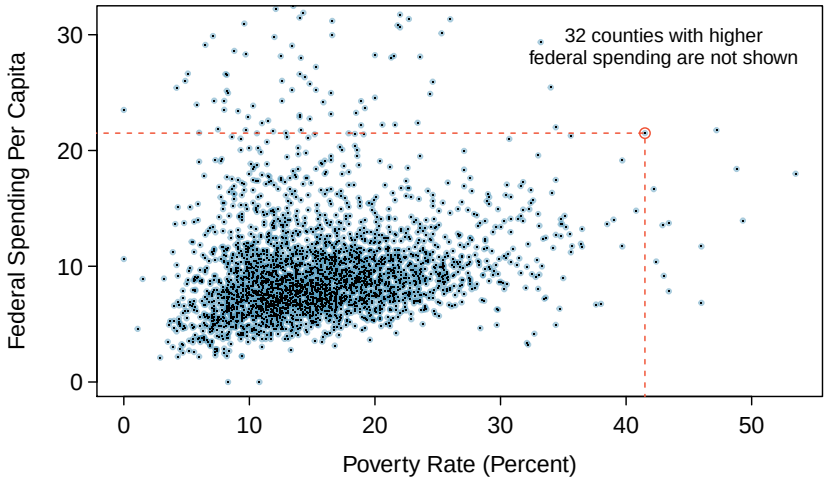

Figure 1.8: A scatterplot showing `fed_spend` against `poverty`. Owsley County of Kentucky, with a poverty rate of 41.5% and federal spending of $21.50 per capita, is highlighted.

Example 1.6 This example examines the relationship between homeownership and the percent of units in multi-unit structures (e.g. apartments, condos), which is visualized using a scatterplot in Figure 1.9. Are these variables associated?

It appears that the larger the fraction of units in multi-unit structures, the lower the homeownership rate. Since there is some relationship between the variables, they are associated.

Because there is a downward trend in Figure 1.9 – counties with more units in multi-unit structures are associated with lower homeownership – these variables are said to be **negatively associated**. A **positive association** is shown in the relationship between the `poverty` and `fed_spend` variables represented in Figure 1.8, where counties with higher poverty rates tend to receive more federal spending per capita.

If two variables are not associated, then they are said to be **independent**. That is, two variables are independent if there is no evident relationship between the two.

Associated or independent, not both
A pair of variables are either related in some way (associated) or not (independent). No pair of variables is both associated and independent.

1.3 Overview of data collection principles

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider *how* data are collected so that they are reliable and help achieve the research goals.

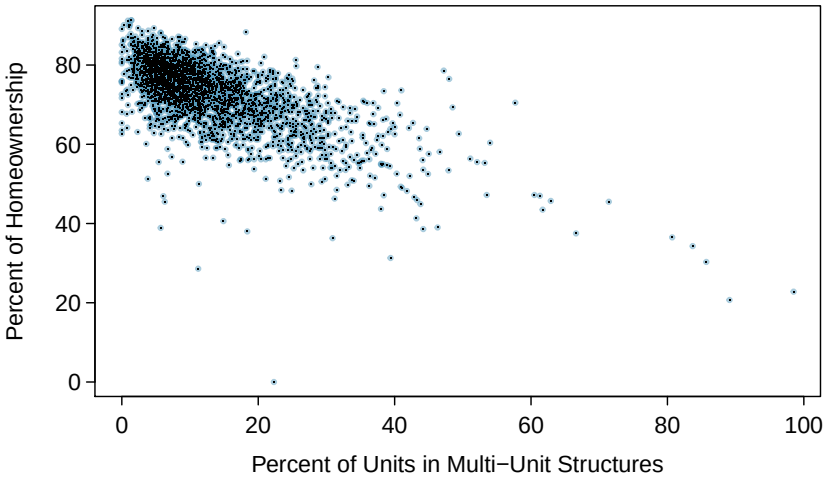

Figure 1.9: A scatterplot of homeownership versus the percent of units that are in multi-unit structures for all 3,143 counties. Interested readers may find an image of this plot with an additional third variable, county population, presented at www.openintro.org/stat/down/MHP.png.

1.3.1 Populations and samples

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to complete a degree for Duke undergraduate students?
3. Does a new drug reduce the number of deaths in patients with severe heart disease?

Each research question refers to a target **population**. In the first question, the target population is all swordfish in the Atlantic ocean, and each fish represents a case. Often times, it is too expensive to collect data for every case in a population. Instead, a sample is taken. A **sample** represents a subset of the cases and is often a small fraction of the population. For instance, 60 swordfish (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

⊙ **Guided Practice 1.7** For the second and third questions above, identify the target population and what represents an individual case.[10]

We collect a sample of data to better understand the characteristics of a population. A **variable** is a characteristic we measure for each individual or case. The overall quantity of interest may be the mean, median, proportion, or some other summary of a population. These population values are called **parameters**. We estimate the value of a parameter

[10](2) Notice that the first question is only relevant to students who complete their degree; the average cannot be computed using a student who never finished her degree. Thus, only Duke undergraduate students who have graduated in the last five years represent cases in the population under consideration. Each such student would represent an individual case. (3) A person with severe heart disease represents a case. The population includes all people with severe heart disease.

by taking a sample and computing a numerical summary called a **statistic** based on that sample. Note that the two p's (population, parameter) go together and the two s's (sample, statistic) go together.

● **Example 1.8** Earlier we asked the question: what is the average mercury content in swordfish in the Atlantic Ocean? Identify the variable to be measured and the parameter and statistic of interest.

The variable is the level of mercury content in swordfish in the Atlantic Ocean. It will be measured for each individual swordfish. The parameter of interest is the average mercury content in *all* swordfish in the Atlantic Ocean. If we take a sample of 50 swordfish from the Atlantic Ocean, the average mercury content among just those 50 swordfish will be the statistic.

Two statistics we will study are the **mean** (also called the **average**) and **proportion**. When we are discussing a population, we label the mean as μ (the Greek letter, *mu*), while we label the sample mean as $\bar{x}$. When we are discussing a proportion in the context of a population, we use the label p, while the sample proportion has a label of $\hat{p}$ (read as *p-hat*). Generally, we use $\bar{x}$ to estimate the population mean, μ. Likewise, we use the sample proportion $\hat{p}$ to estimate the population proportion, p.

● **Example 1.9** Is μ a parameter or statistic? What about $\hat{p}$?

μ is a parameter because it refers to the average of the *entire* population. $\hat{p}$ is a statistic because it is calculated from a sample.

● **Example 1.10** For the second question regarding time to degree for a Duke undergraduate, is the variable numerical or categorical? What is the parameter of interest?

The characteristic that we record on each individual is the number of years until graduation, which is a numerical variable. The parameter of interest is the average time to degree for all Duke undergraduates, and we use μ to describe this quantity.

⊙ **Guided Practice 1.11** The third question asked whether a new drug reduces deaths in patients with severe heart disease. Is the variable numerical or categorical? Describe the statistic that should be calculated in this study.[11]

If these topics are still a bit unclear, don't worry. We'll cover them in greater detail in the next chapter.

1.3.2 Anecdotal evidence

Consider the following possible responses to the three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.

2. I met two students who took more than 7 years to graduate from Duke, so it must take longer to graduate at Duke than at many other colleges.

[11] The variable is whether or not a patient with severe heart disease dies within the time frame of the study. This is categorical because it will be a yes or a no. The statistic that should be recorded is the proportion of patients that die within the time frame of the study, and we would use $\hat{p}$ to denote this quantity.

Figure 1.10: In February 2010, some media pundits cited one large snow storm as valid evidence against global warming. As comedian Jon Stewart pointed out, "It's one storm, in one region, of one country."

February 10th, 2010.

3. My friend's dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work.

Each of the conclusions are based on some data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

Anecdotal evidence
Be careful of making inferences based on anecdotal evidence. Such evidence may be true and verifiable, but it may only represent extraordinary cases. The majority of cases and the average case may in fact be very different.

Anecdotal evidence typically is composed of unusual cases that we recall based on their striking characteristics. For instance, we may vividly remember the time when our friend bought a lottery ticket and won $250 but forget most the times she bought one and lost. Instead of focusing on the most unusual cases, we should examine a representative sample of many cases.

Appendix A

End of chapter exercise solutions

1 Data collection

1.1 (a) Treatment: $10/43 = 0.23 \rightarrow 23\%$. Con- trol: $2/46 = 0.04 \rightarrow 4\%$. (b) There is a 19% difference between the pain reduction rates in the two groups. At first glance, it appears pa- tients in the treatment group are more likely to experience pain reduction from the acupuncture treatment. (c) Answers may vary but should be sensible. Two possible answers: [1] Though the groups' difference is big, I'm skeptical the results show a real difference and think this might be due to chance. [2]The difference in these rates looks pretty big, so I suspect acupuncture is having a positive impact on pain.

1.3 (a-i) 143,196 eligible study subjects born in Southern California between 1989 and 1993. (a-ii) Measurements of carbon monoxide, nitro- gen dioxide, ozone, and particulate matter less than $10\mu g/m^3$ (PM_{10}) collected at air-quality- monitoring stations as well as length of gesta- tion. These are continuous numerical variables. (a-iii) The research question: “Is there an as- sociation between air pollution exposure and preterm births?” (b-i) 600 adult patients aged 18-69 years diagnosed and currently treated for asthma. (b-ii) The variables were whether or not the patient practiced the Buteyko method (categorical) and measures of quality of life, ac- tivity, asthma symptoms and medication reduc- tion of the patients (categorical, ordinal). It may also be reasonable to treat the ratings on a scale of 1 to 10 as discrete numerical variables. (b-iii) The research question: “Do asthmatic patients who practice the Buteyko method experience improvement in their condition?"