Introduction to Medical Statistics with Jamovi

1st Edition

Konstantinos I. Bougioukas, PhD

August 2, 2022

Table of contents

Preface 1
License 2

1 Introduction 3
1.1 Statistics and Medicine 3
1.2 Why Jamovi? 6
1.3 Types of Data 7
1.4 Summarizing Categorical Data 15
1.5 Displaying Categorical Data 19
1.6 Summarizing Numerical Data 21
1.7 Displaying Numerical Data 23

2 LAB I: Introduction to Jamovi (Part I) 25

3 Sampling methods and study designs 27

4 LAB II: Introduction to Jamovi (Part II) 29

5 Probability and distributions 31

6 LAB III: Probability and distributions 33

7 Normal distribution 35

8 LAB IV: Normal distribution 37

9 Foundations for statistical inference 39

10 LAB V: Foundations for inference 41

Table of contents

11 Inference for numerical data: 2 samples 43

12 LAB VI: Inference for numerical data (2 samples) 45

13 Inference for numerical data: >2 samples 47

14 LAB VII: Inference for numerical data (>2 samples) 49

15 Inference for categorical data 51

16 LAB VIII: Inference for categorical data 53

17 Correlation 55

18 LAB IX: Correlation 57

19 Simple linear regression 59

20 LAB X: Simple linear regression 61

21 Reporting the results of statistical analysis 63

References 65

List of Figures

1.1 Jamovi is free and open statistical software to bridge the gap between researcher and statistician 6
1.2 Broad classification of the different types of data with examples 8
1.3 Bar plot showing where 202 patients with corns were treated. 20
1.4 Clustered bar plot showing where 202 patients with corns were treated by randomized group. 21
1.5 Clustered bar plot showing where 202 patients with corns were treated by randomized group. 22

List of Figures

List of Tables

1.1 Baseline characteristics of participants in a RCT of the effectiveness of salicylic acid plasters compared with “usual” scalpel debridement of foot corns by treatment group 8
1.2 Treatment center for 202 patients with corns who were recruited to a RCT 17
1.3 Cross-tabulation of treatment center by randomized group for 202 patients with corns 17
1.4 Reporting numbers and percentages 18

Preface

This textbook is for medical students, doctors, medical researchers, nurses, members of professions allied to medicine, and all others concerned with medical data.

While statistics books focus on mathematics, this textbook focuses on using a computer to conduct data analysis. That means using a statistical software program, in this case the Jamovi software for statistics and graphics. Our aim is to keep a balance between mathematical rigor and readability as well as learning Jamovi and statistics simultaneously.

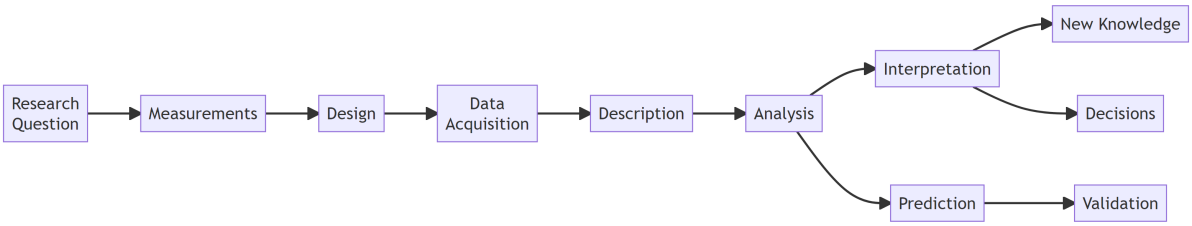
Most of the examples discussed in this textbook are based on scientific studies whose data are publicly available. For each example, we provide the step-by-step application in Jamovi. Readers are encouraged to follow these steps while reading the textbook so that they can learn statistics through hands-on experience.

All sections of this textbook are reproducible as they were made using Quarto® which is an open-source scientific and technical publishing system built on Pandoc.

To learn more about Quarto books visit https://quarto.org/docs/books.

License

1 Introduction



1.1 Statistics and Medicine

Although some healthcare professionals may not carry out medical research, they will definitely be consumers of medical research. Thus, it is incumbent on them to be able to discern high quality research studies from low quality, to be able to verify whether the conclusions of a study are valid and to understand the limitations in methods of a study. The current emphasis on evidence-based medicine (EBM) requires that healthcare professionals consider critically all evidence about whether a specific treatment works and this requires basic statistical knowledge.

Statistics is not only a discipline in its own right but it is also a fundamental tool for investigation in all biological and medical sciences. As such, any serious investigator in these fields must have a grasp of the basic principles. With modern computer facilities there is little need for familiarity with the technical details of statistical calculations. However, a healthcare professional should understand when such calculations are valid, when they are not and how they should be interpreted.

The use of statistical methods pervades the medical literature. In a survey of 350 original articles published in three UK journals of general practice: *British Medical Journal (General Practice Section)*, *British Journal of General Practice* and *Family Practice*, over a one-year period, Rigby et al. (2004) found that 66% used some form of statistical analysis. Another review by Strasak et al. (2007) of 91 original research articles published in *The New England Journal of Medicine* (one of the prestigious peer-reviewed medical journals) found an even higher percentage (95%) of using inferential statistics, for example, hypothesis testing and deriving estimates. It appears, therefore, that the majority of papers published in these journals require some statistical knowledge for a complete understanding.

To students schooled in the 'hard' sciences of physics and chemistry it may be difficult to appreciate the variability of biological data. If one repeatedly puts blue litmus paper into acid solutions it turns red 100% of the time, not most (say 95%) of the time. In contrast, if one gives aspirin to a group of people with headaches, not all of them will experience relief. Penicillin was perhaps one of the few 'miracle' cures where the results were so dramatic that little evaluation was required. Absolute certainty in medicine is rare.

Measurements on human subjects seldom give exactly the same results from one occasion to the next. For example, O' Sullivan et al (1999), found that systolic blood pressure (SBP) in normal healthy children has a wide range, with 95% of children having SBPs below 130 mmHg when they were resting, rising to 160 mmHg during the school day, and falling again to below 130 mmHg at night. Furthermore, Hansen et al. (2010) in a study of over 8000 subjects found that increasing variability in blood pressure over 24 hours was a significant and independent predictor of mortality and a cardiovascular and stroke events.

This variability is also inherent in responses to biological hazards. Most people now accept that cigarette smoking causes lung cancer and heart disease, and yet nearly everyone can point to an apparently healthy 80-year-old who has smoked for many years without apparent ill effect. Although it is now known from the report of Doll et al (2004) that about half of all persistent cigarette smokers are killed by their habit, it is usually forgotten that until the 1950s, the cause of the rise in lung cancer deaths was a mystery and commonly associated with general atmospheric pollution such as the exhaust fumes of cars. It was not until the carefully designed and statistically analysed case–control and cohort studies of Richard Doll and Austin Bradford Hill and others, that smoking was identified as the true cause. Enstrom et al. (2003) moved the debate on to ask whether or not passive smoking causes lung cancer. This is a more difficult question to answer since the association is weaker. However, studies by Cao et al. (2015) have now shown that it is a major health problem and scientists at the International Agency for Rsearch on Cancer (IARC) have concluded that there is sufficient evidence

that second-hand smoke causes lung cancer (IARC 2012). Restrictions on smoking in public places have been imposed to smokers.

With such variability, it follows that in any comparison made in a medical context, such as people on different treatments, differences are almost bound to occur. These differences may be due to real effects, random variation or variation in some other factor that may affect an outcome. It is the job of the analyst to decide how much variation should be ascribed to chance or other factors, so that any remaining variation can be assumed to be due to a real effect. This is the art of statistics.

1.2 Why Jamovi?

Jamovi is a new fee open "3rd generation" statistical software that is built on top of the programming language R (Figure 1.1). Designed from the ground up to be easy to use, Jamovi is a compelling alternative to costly statistical products such as SPSS and SAS.

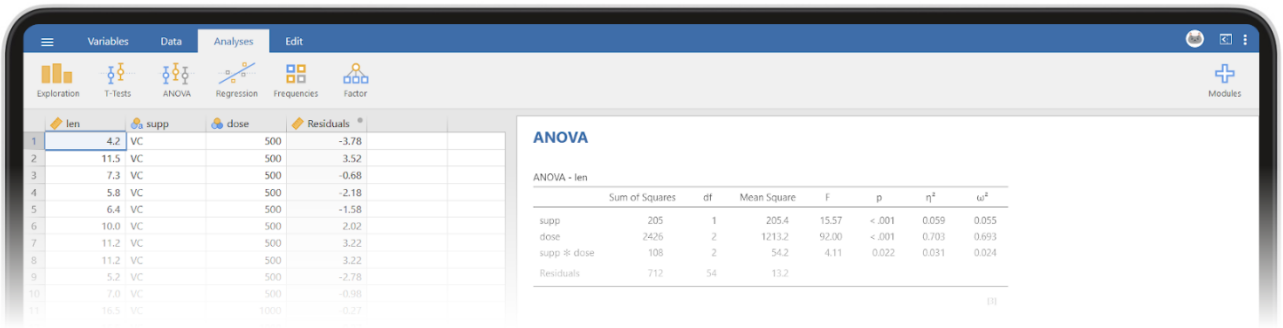


Figure 1.1: Jamovi is free and open statistical software to bridge the gap between researcher and statistician

Some other advantages are:

1. Enables integration with R
2. It provides informative tables and neat visuals
3. Gives access to a user guide and community resources from the Jamovi website

1.3 Types of Data

Data can be either categorical or numerical (otherwise known as qualitative and quantitative) in nature (Figure 1.2).

Example from the literature - Salicylic acid plasters for treatment of foot corns

Table 1.1 presents a typical table with basic characteristics of a set of patients entered into a randomized controlled trial (RCT) that investigated the effectiveness of salicylic acid plasters compared with usual scalpel debridement for treatment of foot corns (Farndon et al. 2013). Corns and calluses are areas of hard, thickened skin that develop when the skin is exposed to excessive pressure or friction. They commonly occur on the feet and can cause pain and discomfort when you walk.

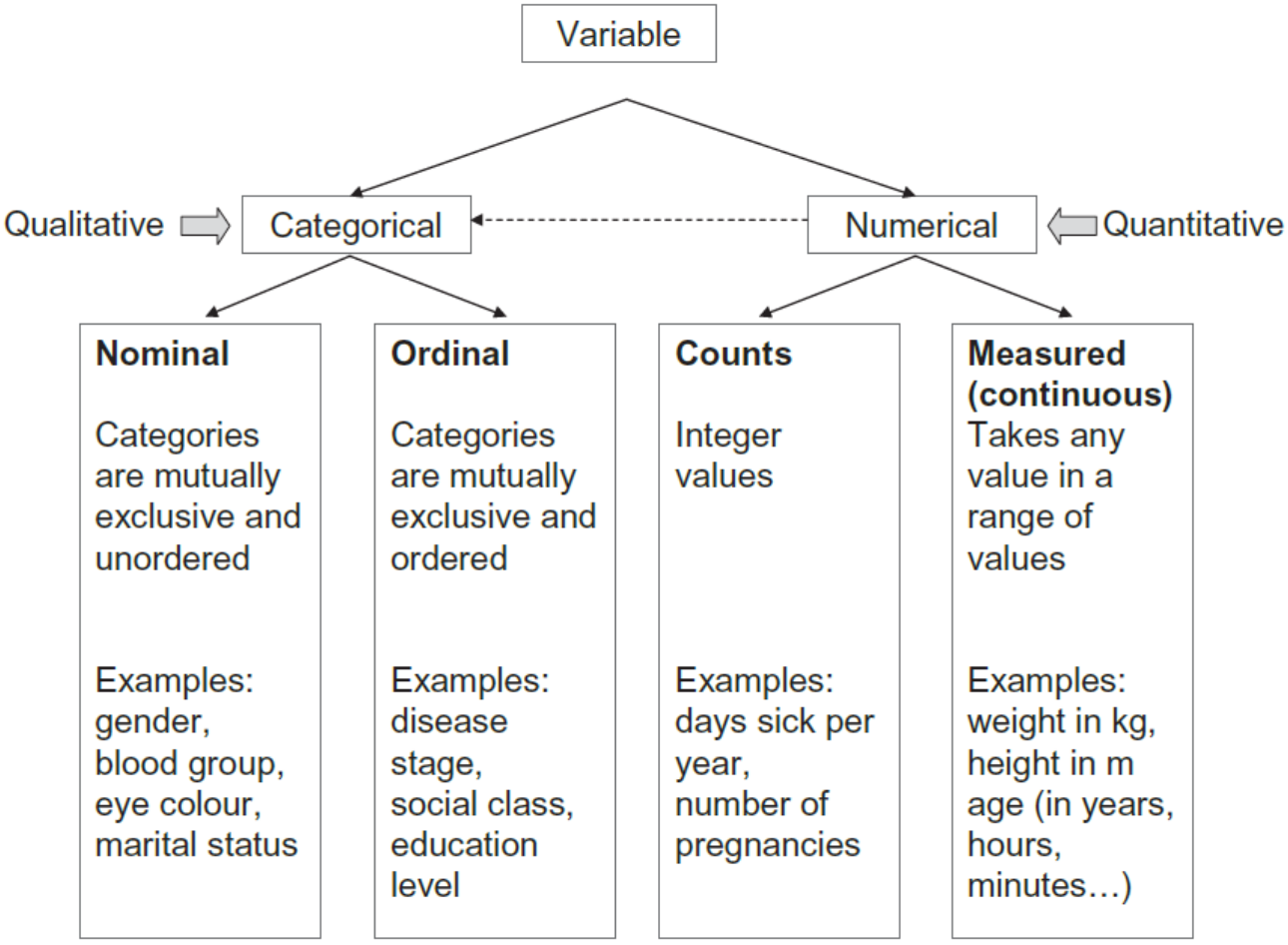


Figure 1.2: Broad classification of the different types of data with examples

Table 1.1: Baseline characteristics of participants in a RCT of the effectiveness of salicylic acid plasters compared with "usual" scalpel debridement of foot corns by treatment group

		Corn Plaster, n (%) (n=101)	Scalpel, n (%) (n=101)
Gender	Male	42 (42)	42 (42)
	Female	59 (59)	59 (58)
Center	Central	58 (57)	52 (52)

		Corn Plaster, n (%) (n=101)	Scalpel, n (%) (n=101)
	Manor	13 (13)	20 (20)
	Jordanthorpe	10 (10)	14 (14)
	Limbrick	3 (3)	6 (6)
	Firth Park	7 (7)	4 (4)
	Huddersfield	5 (5)	4 (4)
	Darnall	5 (5)	1 (1)
Smoking History	Non-smoker	34 (35)	40 (40)
	Previous smoker	22 (22)	16 (16)
	Current smoker	42 (43)	43 (43)
	Missing	3 (3)	2 (2)
Numbers of corns	1	48 (48)	66 (65)
	2	28 (28)	23 (23)
	3	24 (24)	12 (12)
	Missing	1 (1)	0 (0)
Age (yrs), mean (sd)		58.5 (15.6)	59.7 (17.5)
Corn size (mm), median (IQR)		4 (3, 5)	3 (3, 5)
EQ-5D, median (IQR)		0.73 (0.59, 0.80)	0.73 (0.66, 0.80)

In this example since we have 101 patients in each randomized group the percentages are almost the same as the raw counts. However, for most studies we are unlikely to have exactly 100 participants in each group!

1.3.1 Categorical Data

A. Nominal Data

Nominal categorical data are data that one can name and put into categories. They are not measured but simply **counted**. They often consist of **unordered** 'either–or' type observations which have two categories and are often know as **binary**. For example: dead or alive; male or female; cured or not cured; pregnant or not pregnant. In Table 1.1 gender is a binary variable. However, nominal categorical data often can have **more that two categories**, for example, blood group A, B, AB, 0; country of origin; ethnic group; eye color. The Table 1.1 gives the number and percentages of people treated at each of the seven centers in each of the two randomized groups.

> ⚠ Warning
>
> **Numerical representation of categories are just codes**
> We can denote a male and female as 1 and 2 for gender and denote A, B, AB and 0, as 1, 2, 3, and 4 for blood type. Unlike numerical data, the numbers representing different categories do not have mathematical meanings (they are just codes).

B. Ordinal Data

If there are more than two categories of classification it may be possible to **order** them in some way. For example, after treatment a patient may be either improved, the same or worse. Another example of an ordinal variable is the variable **pain** where a subject is asked to describe their pain verbally as minimal, moderate, severe, or unbearable. In Table 1.1 **smoking history** is given in three categories: non-smoker, previous smoker, and current smoker. Thus, someone who is a current smoker has more recent exposure to tobacco than someone who is an ex-smoker and someone who has never smoked. However, **without further knowledge** (of the current and past levels of tobacco consumption) it would be wrong to ascribe a numerical quantity to the category, for example, non-smoker=0, previous smoker=1, and current smoker=2, as one cannot say that someone who is current smoker has twice the levels of tobacco consumption as someone who is a previous smoker.

> ⚠ Warning
>
> **Collapsion of categories leads to a loss of information**
> Ordinal data are often reduced to two categories to simplify analysis and presentation, which may result in a considerable loss of information.

1.3.2 Numerical Data

A. Count (or discrete) Data

Table 1.1 gives details of the number of corns each participant had at the start of the trial, since this can only be a whole number or integer value, for example, 0, 1, 2, or 3 in this trial, this is termed count data. Other examples

are often counts per unit of time such as the number of deaths in a hospital per year, the number of visits to the GP in a year, or the number of attacks of asthma a person has per month. In dentistry, a common measure is the number of decayed, filled or missing teeth (DFM).

The difference between such data as these and the ordered categorical data described earlier can be seen by considering an example of each:

Illustrative Example: Ordinal Vs Discrete data

Ordinal categorical: Stage of breast cancer: I II III IV

Discrete numerical: Number of children: 0 1 2 3 4 5+

We cannot say that stage IV is twice as bad as stage II nor that the difference between stages I and II is equivalent to that between stages III and IV. In contrast, three children are three times as many as one, and a difference of one means the same throughout the range of values.

⚠ Warning

Discrete Vs Ordinal Data

In practice discrete data are often treated in statistical analyses as if they were ordered categories. This is not wrong, but it may not be getting the most out of the data. Conversely, when ordered categories are numbered, as with stage of disease, the temptation to treat these numbers as statistically meaningful must be resisted. For example, it is not sensible to calculate the average stage of cancer. The only information the numbers contain is in the ordering, which would be conveyed equally by calling

them A, B, C, D and so on.

B. Continuous (or measured) Data

Such data are measurements that can, in theory at least, take any value within a given range (they are restricted by the accuracy of the measuring instrument). These data contain the most information, and are the ones most commonly used in statistics. Examples of continuous data in Table 1.1 are: age, corn size, and EQ-5D.

Sometimes it is reasonable to treat discrete data as if they were continuous, at least as far as statistical analysis goes. While age is a continuous measurement, **age at last birthday is discrete**. In studies of adults with ages ranging from, say, 16 to 80, no harm is done in considering age in years as a continuous measurement (and this is standard practice), but for studies of pre-school children it would be better to use age in months. **Heart rate** (in beats per minute) is another discrete measurement that is usually regarded as continuous. Although the essential requirement for this change of status is that there should be a large number of different possible values, in practice we do not worry too much about analysing discrete measurements as if they were continuous.

⚠ Warning

Categorization of numerical data leads to a loss of information

For simplicity, it is often the case in medicine that continuous data are dichotomized to make nominal data. Thus diastolic blood pressure (DBP),

which is continuous, is converted into hypertension (>90 mmHg) and normotension (≤90 mmHg). This clearly leads to a loss of information. There are two main reasons for doing this. It is easier to describe a population by the proportion of people affected, for example, the proportion of people in the population with hypertension is 10%. Further, one often has to make a decision: if a person has hypertension, then they will get treatment, and this too is easier if high blood pressure has been categorized.

One can also divide a continuous variable into more than two groups. For example, we could divide age into age bands of equal lengths of, say 10 years such as: 0-9, 10-19, 20-29, etc. When categorizing continuous data authors should give an indication as to why they chose these cutt-off points, and a reader has to be very wary to guard against the fact that the cuts may be chosen to make a particular point. Some statisticians have termed the habit of categorizing continuous variables as "dichotomania", which they regard as poor practice since it loses information and assumes a discontinuous relationship that is unlikely in nature.

Tip

Record the actual values

It is best to record the actual value of blood pressure, haemoglobin, etc. It is easy to convert to categories in the analysis, but the raw data cannot be retrieved later if only categories are recorded. Information is lost with no compensatory gain. Indeed, the statistical analysis of continuous data

> is more powerful, and often simpler.
>
> When some calculation is necessary to derive the observation of interest this should be done by the computer. Thus it is much better to record date of birth and date of examination for subsequent calculation of age rather than to rely on mental arithmetic.

The degree of measurement accuracy and the type of data are both important in relation to carrying out a proper statistical analysis.

1.4 Summarizing Categorical Data

Binary data are the simplest type of data. Each individual has a label which takes one of two values such as male or female, corn healed or not healed. A simple summary would be to count the different types of labels and find the **frequencies**. The set of frequencies of all the possible categories is called the **frequency distribution** of the variable.

However, a raw count is rarely useful. For example, in Table 1.1 there are more non-smokers in the scalpel group (40 out of 99 or 40%) compared to corn plaster group (34 out of 98 or 35%). It is only when this count is expressed as a proportion (relative frequency) that it becomes useful. Hence the first step to analyzing categorical data is to count the number of observations in each category (frequencies) and express them as **proportions** of the total sample size (relative frequencies).

Illustrative Example - Salicylic acid plasters for treatment of foot corns

Farndon et al. (2013) reports a RCT that investigated the effectiveness of salicylic acid plasters compared with usual scalpel debridement for treatment of foot corns. As we have already mentioned one categorical variable recorded was the centre where each trial participant was treated. Trial participants were treated at one of seven centers and the corresponding categories as displayed in Table 1.2. The first column shows category (treatment center) names, whilst the second shows the number of individuals in each category together with its percentage contribution to the total. Table 1.2 clearly shows that the majority (54.5%) of patients were treated at the "Central" treatment center.

In addition to tabulating each variable separately, we might be interested in whether the distribution of patients across each center is the same for each randomized group. Table 1.3 shows the distribution of the number of patients treated at center by randomized group; in this case it can be said that the treatment center has been ***cross-tabulated*** with randomized group. Table 1.3 is an example of a ***contingency*** table with seven rows (representing treatment center) and two columns (randomized group). Note that we are interested in the distribution of patients across the seven centers in each randomized group (to see whether or not we have similar numbers of patients randomized to each treatment within each center), and so the percentages add to 100 down each column, rather than across the rows.

Table 1.2: Treatment center for 202 patients with corns who were recruited to a RCT

Treatment center	Frequency	Percentage
Central	110	54.5%
Manor	33	16.3%
Jordanthorpe	24	11.9%
Limbrick	9	4.5%
Firth Park	11	5.4%
Huddersfield	9	4.5%
Darnall	6	3.0%
Total	202	100.0%

Table 1.3: Cross-tabulation of treatment center by randomized group for 202 patients with corns

	Corn plaster, n(%)	Scalpel, n(%)	All, n(%)
Central	58 (57)	52 (52)	110 (54.5)
Manor	13 (13)	20 (20)	33 (16.3)
Jordanthorpe	10 (10)	14 (14)	24 (11.9)
Limbrick	3 (3)	6 (6)	9 (4.5)
Firth Park	7 (7)	4 (4)	11 (5.4)
Huddersfield	5 (5)	4 (4)	9 (4.5)
Darnall	5 (5)	1 (1)	6 (3.0)
Total	101 (100)	101 (100)	202 (100)

	Corn plaster, n(%)	Scalpel, n(%)	All, n(%)

Tip

Recommendations for reporting numbers

Table 1.4: Reporting numbers and percentages

Recommendation	Correct expression
Numbers	
In a sentence, numbers less than 10 are words.	Smoking history was missing from three patients in the corn plaster study group.
In a sentence, numbers 10 or more are numbers.	There are 34 non-smokers patients in the corn plaster group.
Avoid starting a sentence with numbers.	Thirty-four non-smokers patients recorded in the cord plaster group.
Percentages	
Report percentages to only one decimal place if the sample size is larger than 100.	In the sample of 202 patients, 4.5% were treated at the “Limbrick” treatment center.

Report percentages with no decimal places if the sample size is less than 100.	In the sample of 98 patients in the corn plaster group, 35% were non-smokers.
Do not use percentages if the sample size is less than 20.	From 16 previous smokers in the scalpel group, 7 were females.

1.5 Displaying Categorical Data

The best way to investigate a dataset is of course to plot it. For categorical variables, such as gender and treatment center, it is straightforward to present the number in each category, usually indicating the frequency and percentage of the total number of patients. When shown graphically this is called a bar plot. Figure 1.3 shows in a **bar plot** the reqruiting centers of the 202 patients with foot corns treated in the trial of Farndon et al. (2013). Along the horizontal axis (x-axis) are the different treatment center categories whilst on the vertical axis (y-axis) is the percentage. Each bar represents the percentage of the total patients in that category. For example, it can be seen that the percentage of participants who were treated in the “Central” center was about 55%.

If the sample is further classified into whether the patient was treated with corn plasters or scalpel then it becomes impossible to present the data as

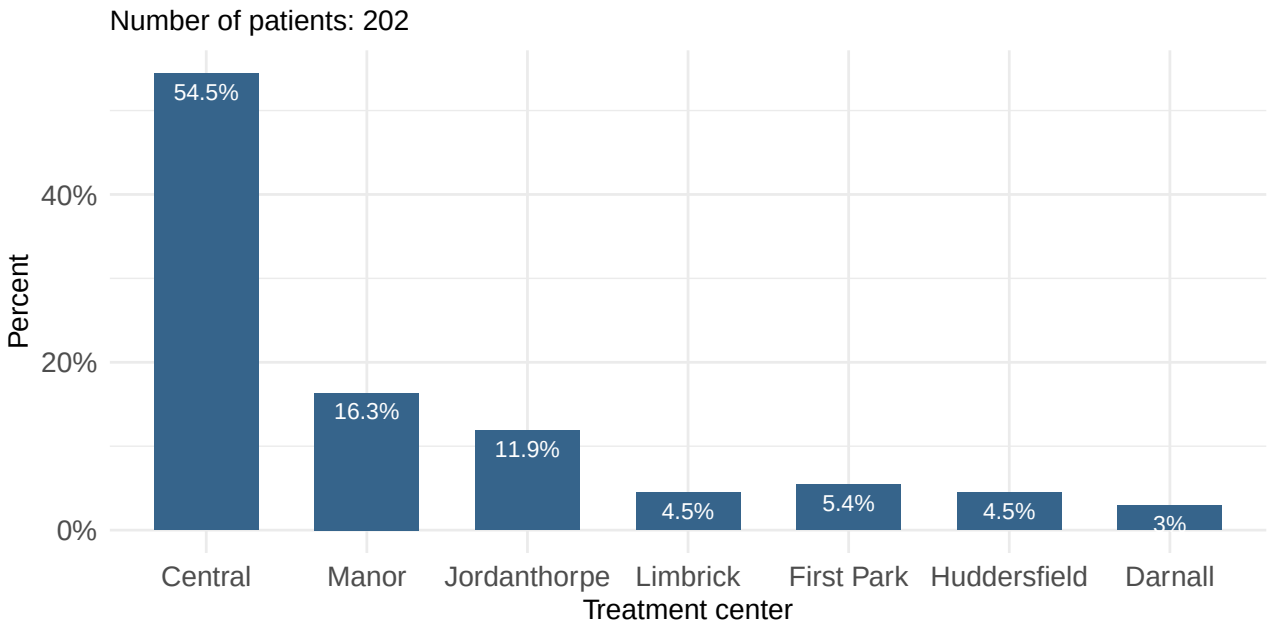


Figure 1.3: Bar plot showing where 202 patients with corns were treated.

a single bar plot. We could present the data as a side by side bar plot (see Figure 1.4) but is preferable to present the data in one graph with the same scales and axes to make the visual comparisons easier (clustered or grouped bar plot) (see Figure 1.5).

If you do use the relative frequency scale as we have, then it is recommended to report the actual total sample sizes for each group (e.g., in the legend or caption). In this way, given the total sample size and relative frequency (from the height of the bars) we can work out the actual numbers treated in each center.

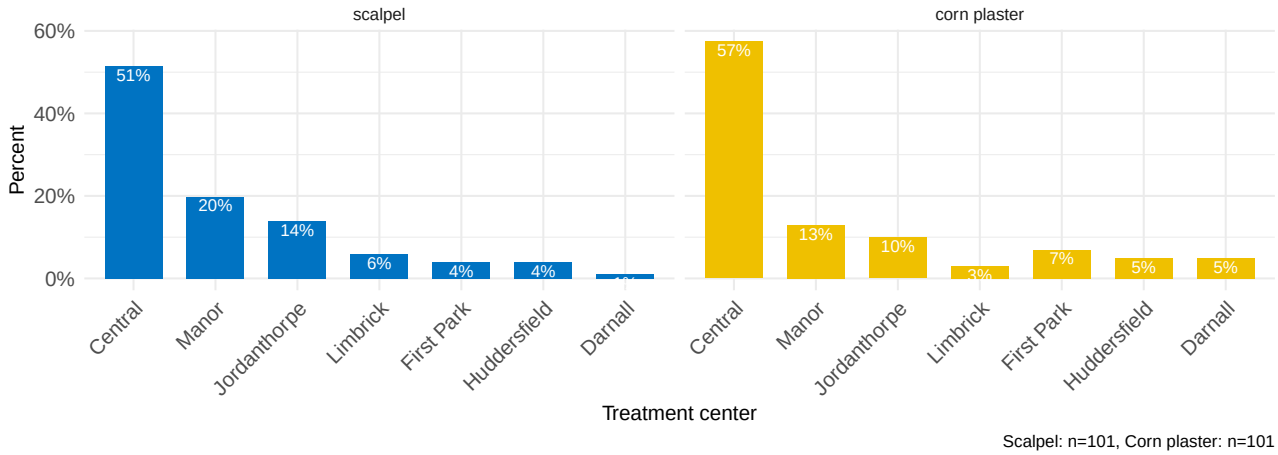


Figure 1.4: Clustered bar plot showing where 202 patients with corns were treated by randomized group.

1.6 Summarizing Numerical Data

A quantitative measurement contains more information than a categorical one, and so summarizing these data is more complex. One chooses summary statistics to condense a large amount of information into a few intelligible numbers, the sort that could be communicated verbally. The two most important pieces of information about a quantitative measurement are **'where is it?'** and **'how variable is it?'** These are categorized as **measures of location** (or sometimes 'central tendency') and **measures of spread** or variability.

Two summary measures should be reported for a numerical variable

A measure of **location** (where the center of the distribution of the values is located) and **variability** (how widely the values are spread above and below the central value) provides an informative but brief summary of a set of observations.

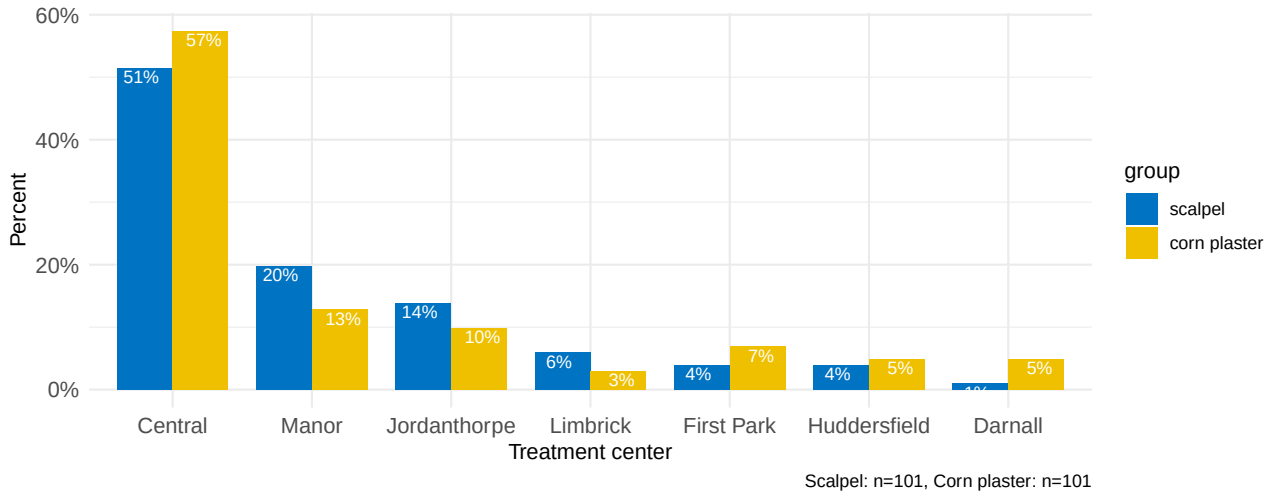


Figure 1.5: Clustered bar plot showing where 202 patients with corns were treated by randomized group.

1.6.1 Measures of Location

"Center" of a sample

- Mean: arithmetic average

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Population mean μ is the long-run average (let $n \to \infty$ in computing $\bar{x}$) + center of mass of the data (balancing point) + highly influenced by extreme values even if they are highly atypical

- Median: middle sorted value, i.e., value such that $\frac{1}{2}$ of the values are below it and above it
  - always descriptive
  - unaffected by extreme values

- not a good measure of central tendency when there are heavy ties in the data

1.7 Displaying Numerical Data

2 LAB I: Introduction to Jamovi (Part I)

2 LAB I: Introduction to Jamovi (Part I)

3 Sampling methods and study designs

4 LAB II: Introduction to Jamovi (Part II)

4 LAB II: Introduction to Jamovi (Part II)

5 Probability and distributions

6 LAB III: Probability and distributions

7 Normal distribution

8 LAB IV: Normal distribution

9 Foundations for statistical inference

10 LAB V: Foundations for inference

11 Inference for numerical data: 2 samples

Two sample t-test (Student's t-test) can be used if we have two independent (unrelated) groups (e.g., males-females, unmatched case-controls, treatment-non treatment) and one quantitative variable of interest (e.g., age, weight, systolic blood pressure). For example, we may want to compare the age in males and females or the weights in two groups of children, each child being randomly allocated to receive either a dietary supplement or placebo.

Assumptions for conducting a Student's t-test

1. The groups are independent
2. The outcome of interest is continuous
3. The data is normally distributed in both groups
4. The data in both groups have similar standard deviations

$$t = \frac{\bar{x}_1 - \bar{x}_2}{se_{dif}}$$

12 LAB VI: Inference for numerical data (2 samples)

12 LAB VI: Inference for numerical data (2 samples)

13 Inference for numerical data: $>$2 samples

14 LAB VII: Inference for numerical data ($>$2 samples)

14 LAB VII: Inference for numerical data (>2 samples)

15 Inference for categorical data

16 LAB VIII: Inference for categorical data

17 Correlation

18 LAB IX: Correlation

19 Simple linear regression

20 LAB X: Simple linear regression

21 Reporting the results of statistical analysis

References