# 1 Introduction: Experiments and Variables

Each chapter of this book will involve the analysis or discussion of data resulting from two perceptual experiment carried out by a group of fifteen listeners. The first 8 chapters will focus on an experiment investigating the perception of height from speech, and a second experiment investigating vowel perception will be introduced in Chapter 9. As noted in the preface, a basic working knowledge of R is assumed and a familiarity with basic statistics is helpful, though not strictly necessary. The preface also provides suggested readings for those wanting to do some background reading on R or statistical inference. In this chapter we will discuss the data for the first experiment, introduce some basic concepts related to variables and probabilities, and provide a very basic introduction to R along the way. For information about the software you need installed to follow along with the examples in the chapter, please see section X.

## 1.1 Experiments and effects

An **experiment** is a procedure or process than can help answer some research question. Obviously, when defined so broadly almost anything can be an experiment. In fact, when a child touches a hot stove to see what it feels like, they are conducting an experiment which provides essential information about their world. In an academic context, experiments are expected to be **scientific**. However, there is no definition of *scientific* that transcends space and time. What is considered ‘scientific’ is determined by what scientists in a specific time and place consider to be scientific, and this can change, and has changed, substantially over time. At the moment, in most contexts, a research project is ‘scientific’ when it generally conforms the the **scientific method**. Of course, just as with **science** and scientific, there is no *single* scientific-method, no single true definition that can be referred to. Instead, the scientific method consists generally of a process in which researchers: 1) Ask questions based on gaps in their knowledge about the world, 2) Collect data that can help answer their questions, 3) Evaluate their questions in light of their data, and 4) Reach conclusions where possible, synthesize their conclusions with their previous knowledge about the world.

Modern ‘scientific’ work usually involves the collection of empirical measurements, the quantification of patterns in these measurements, and the qualitative description of the quantitative patterns in the measurements. As a result, much modern scientific work yields large quantities of numeric values, observed under different conditions, which the researcher must then analyze in order to understand. For example, imagine an experiment about whether caffeine makes people talk faster. Subjects are asked to drink either a cup of regular or decaffeinated coffee. After a 30-minute wait they are asked to read a passage aloud and the duration of the reading is measured. Basically we are measuring two different values, “the amount of time it takes people to read a passage of text”, and “the amount of time it takes people to read this passage of text after consuming caffeine”. Our experiment allows us to ask: is “the amount of time it takes people to read this passage of text” *the same thing* as “the amount of time it takes people to read this passage of text after consuming caffeine”? Another way to look at this is that we are interested in the **effect** of caffeine on reading times. By ‘effect’ we mean the ability of caffeine to change the characteristics of our observation (reading times) in some way. For example, if the average reading times were the same in both groups we would conclude that “caffeine has no *effect* on reading times”. In contrast, if reading times were 800 milliseconds seconds shorter in the caffeine group, we might conclude “caffeine has the effect of reducing reading times by 800 milliseconds”.

Our experiment on reading times is specifically constructed to investigate the effect of caffeine on readings times. If the speakers in our experiment were randomly assigned to conditions, there is no particular reason to expect that their reading times would be different *in the absence* of the caffeine. So, if we find that people read faster in the caffeine group, we may infer that it is the caffeine that has had the *effect* of causing the increase in speaking rate. This same logic applies in situations where we do not randomly assign subjects to groups, as long as we are careful in creating equivalent groups. Consider the same experiment about speaking rate carried out with groups based on speaker gender rather than drinking coffee. In this case the question would be “is the amount of time it takes men to read this passage of text the same amount of time that it takes women to read this passage of text”. If the speakers are generally similar in important characteristics (e.g., dialect, age, cultural background) *apart* from gender, then any group differences may be attributable to the effect for gender on speaking rate.

What we are describing in the above paragraph are **controlled experiments**, experiments where the researcher takes an active role in ensuring the ‘fairness’ of the experiment. The notions of control and fairness and somewhat hazy, and are perhaps more gradient than discrete (i.e. ‘controllved’ vs. ‘uncontrolled’). However, some situations clearly do not lead to ‘fair’ outcomes. For example, what if the caffeine group of readers were all first language English speakers, and the decaf group had substantial number of second language speakers. The ceffeine group may very well read faster simply because they are more polished readers, independently of the effects of caffeine. Whenever possible, researchers avoid situations like this by exerting *control* over their experiments, both in the structure of their experiments and in the recruiting and assignment of their participants to experimental conditions.

All of the experiments discussed to this point would result in one (or more) reading time per subject per group. Due to random between-speaker variation (among other things), there is no chance whatsoever that the average readings times across both groups will be exactly identical, even if caffeine has no effect on reading times at all. Actually, if you re-ran the experiment, there is basically no chance that the group means would match themselves exactly in the replication. And yet, there is the possibility that caffeinated reading times are *actually* different in a way that the random variation of groups across replications is not. So, how can we ever establish that our measures are *actually* different and don’t just *appear* to be different because of randomness? It is precisely this problem that has motivated scientists use statistical analyses to help answer their research questions.

## 1.2 Experiments and inference

This book is about statistical inference. We will talk about the ‘statistics’ part in more detail in the next chapter, but we can talk about the ‘inference’ part now. **Inference** is a form of reasoning that allows you to go from a limited number of observations to a general conclusion. For example, you may arrive at a newly discovered island and see white cats wandering around. If you are there for a while and continue to observe only white cats, you may conclude “all the cats on this island are white”. If you do this you have made what is called an **inductive** inference: You have gone from a set of observations (the cats you saw) to a general conclusion about all the cats on the island. Often, experiments are not just about observing and measuring certain effects, but also about making inferences regarding those effects. For example, in the reading time experiment described above the researchers are not specifically interested in the reading times of the people in the experiments (i.e., the cats they saw) but rather about the reading times of people more generally (i.e. all the cats on the island).

Since inductive inference seeks to go from limited observation to general rules or principles, it has a central weakness. For example, your inference that only white cats exist on the island is on solid ground until you see a cat that is not white. Can you be sure this won’t happen? You can’t, because fundamentally you don’t know what you don’t know and you can’t be sure that what hasn’t happened yet will never happen. This is called **the problem of induction** and it is a fundamental weakness of inductive reasoning.

It’s useful to be aware of the fundamental limitations of trying to understand general patterns given limited sets of observations, and to be generally skeptical. It is also useful to think about how we can reason in a way that might minimize the odds of inferential mistakes, especially by including our general knowledge of the world (and the specific topic) in our reasoning. For example, rather than observing white cats and leaving it at that we can ask: Why are the cats white? Do evolutionary pressures cause them to be white? How do their genetics ensure that all members of the species will be white? Is there any chance non-white cats could enter into the population? Considering the answers to questions like this, in combination with our observations, can make inferences like “all cats on this island are white” more reliable.

For example, the examples above involved the effect of caffeine on reading times. We are interested in generalizing to the human population based on what is a tiny sample of humans (relatively speaking). If we make the claim “caffeine speeds up reading times”, are we extending that to all humans, or at least to all English speakers? Past, present and future? That is a bold claim based on a small number of data points, or it would be in the total absence of any world knowledge and prior expectations. Of course, we know that caffeine is a stimulant and seems reasonably likely to make people read faster. As a result, the finding fits within our larger world view and, as a result, we may accept as likely to be ‘true’. In contrast, suppose that the two groups had instead drank plain water, one ‘regular’ and one dyed with blue food coloring. In this situation we may be skeptical of any finding for an effect for the food coloring. This is because there is no reason to suppose that there is an effect. Since this finding does not conform to any prior knowledge about the world, it is the sort of inference that may turn out to be less reliable, in the long run.

## 1.3 Our experiment

As noted above, each chapter in this book will feature the analysis of data from a perceptual experiment. In this section we provide information about the experiment in general, the design of the experiment, the general research questions this experiment can address, and an overview of the data resulting from our experiment.

### 1.3.1 Our experiment: Introduction

Any group of speakers will ‘sound’ different from each other even when they are all saying the ‘same’ word. These between-speaker differences can, in some cases, be systematically be associated with speaker characteristics such as age, height, and gender. So, tall speakers may tend to sound one way, while shorter speakers may tend to sound other ways. As a result, although it may sound odd to talk about how tall someone *sounds*, listeners are able to use the information in a speaker’s voice to *guess* information about the speaker. We call this information the speaker **indexical characteristics**, social and physical information regarding the speaker that is understood from the way someone speaks. We can ask two different question with respect to assessments of indexical characteristics from speech: 1) Are they accurate, and 2) How do they arrive at their guesses? Generally speaking, listeners are often not very accurate in their judgments of indexical characteristics, however, they are very consistent in the errors that they tend to make. For example, if one voice is incorrectly assumed to belong to some sort of speaker, it will often be the case that this mistake is a regular occurrence.

Generally, the ‘guessing’ of speaker characteristics is dominated by two acoustic cues: Voice pitch and voice resonance. Voice **pitch** can be thought of as the ‘note’ someone produces with their speech. When you sing you produce different notes by producing different pitches. The pitch of a sound is related to the vibration rate of the thing that produced the sound, because repetitive vibration produces a repetitive sound wave that humans perceive as musicality. Human voice pitch is regulated by changing the vibration rate of the vocal folds in your larynx. You can feel this vibration if you hum a song and press your fingers against the middle of the front of your neck. Pitch is an **auditory sensation**, a *feeling* you have in relation to an acoustic event, a sound. When you hear two sounds, you can order them based on which *sounds* lower/higher than the other. That’s pitch. Since this quality cannot be directly measured, scientists measure the **fundamental frequency** (f0) of the sound to quantify its pitch. The f0 of a sound is measured in **Hertz** (Hz), which measures how many times a sound repeats itself in a given second.

Generally speaking, smaller things tend vibrate at higher rates than larger things. This holds for vocal folds as well; shorter vocal folds tends to want to vibrate at higher rates thereby producing speech with a higher pitch. As a result, generally speaking, larger speakers tend to produce speech with a lower pitch. Since the vocal folds generally grows as one ages into adulthood, voice pitch may be an indicator of age between young childhood and adulthood. What we mean is that pitch may be able to help you distinguish a 5 year old from an 18 year old but maybe not an 18 year old from a 30 year old. In addition to general age-related changes, the vocal folds tend to increase in size quite a bit during male puberty so that post-pubescent males tend to produce speech with a lower pitch than the rest of the human population. As a result of these relations, a voice with a lower voice pitch is more likely to *older*, *taller*, and *more male* than a voice with a higher pitch. The relationships between age, height, gender and f0 are presented in Figure 1.1

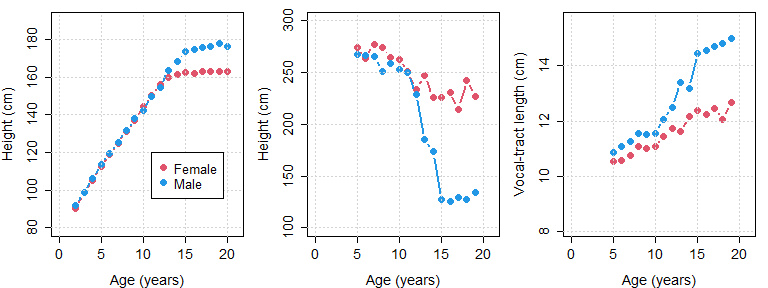


Figure 1.1: (left) Average height of males and females in the United states of America, organized by age (cite). (middle) Estimated vocal-tract length for male and female speakers between from 5 years of age until adulthood, based on the acoustic data provided in Lee at al. (?). (right) Average f0 produced by male and female speakers between from 5 years of age until adulthood.

**Resonance** can be thought of as the ‘size’ of a sound. For example, a violin and a cello can be playing the same note (with the same pitch), but a cello ‘sounds’ bigger. This is because it resonates lower frequencies by virtue of being a larger structure. In the same way, speakers with longer vocal tracts (the space from the vocal folds to the lips) tend to ‘sound’ bigger by producing speech with lower frequencies overall. We don’t really have good words to describe what resonance ‘sounds’ like, but a small resonance (short vocal tract) sounds ‘heliumy’. When a person breathes helium and speaks, their speech does not go up, but their **resonance** frequencies increase (for more information on this, see the Appendix). Long vocal tracts sounds like slow motion speech (think of someone saying “noooooooooooooooooo….” when something bad is happening in slow motion in a movie), and this is because slowing down the playback of a recording simulates a lowering of resonance frequencies in speech. In fact, size simulation by resonance manipulation is how the recordings for ‘Alvin and the chipmunks’ were originally created. A low-pitched male singer was recorded singing abnormally slow, and the recording was sped up in order to simulate a speech with a very high resonance (and an associated very short vocal tract).

There are many ways to measure the resonance of a voice. In our data we will use speech acoustics to directly estimate the length of the vocal tract that produced it, in centimeters (in the manner described in the Appendix). So, our measure of voice resonance will not be acoustic at all but will instead measure the physical correlate of the vocal tract expected to have produced the speech sound. In general, lower frequencies overall suggests a lower voice resonance which in turn suggests a longer vocal tract length in centimeters. There is a strong positive relationship between vocal-tract length and body length (i.e. height) across the entire human population. This means that as a person is taller, their vocal-tract is expected to be longer and their voice resonance is expected to be lower. Since height increases from birth into adulthood, this means that voice resonance can be used to predict both height and age. In addition, adult males tend to be somewhat taller than adult females in most populations, with the difference of about 15 cm in the United States. As a result, voice resonance can be used to infer the gender of adult speakers, and possibly that of children as well. These relationships are shown in Figure 1.1.

So, voice pitch and voice size are independent ways that someone can acoustically ‘sound’ bigger/smaller, older/younger, and male/female. The experiment to be described below involves a **perceptual experiment** involving **behavioral measures**. This means that in this experiment human listeners were played auditory stimuli (words) and were asked to listen to then and answer questions regarding what they heard. The experiment was designed to investigate the way that speech acoustics are used by listeners to determine the age, gender, and height of speakers, and the way that these decisions affect each other.

### 1.3.2 Our experimental methods

Our listeners were 15 native speakers of American English. Listeners were presented with the word “heed” produced by 139 different speakers of Michigan English. These speech samples were recorded by Hillenbrand et al (1995) and are available on the GitHub page associated with this book. So, this experiment featured 139 unique **stimulus** sounds that the listeners in the experiment were asked to respond to. The stimuli used were productions by 48 adult females, 45 adult males, 19 girls (10-12 years of age), and 27 boys (10-12 years of age). These speakers showed substantial variation in their voice pitch and resonance as measured by their f0 and estimated vocal-tract length (as will be discussed in section 1.5). In addition to the natural acoustic variation that exists between speakers, voice resonance was also manipulated experimentally. All stimuli were manipulated by shifting the spectral envelope down by 10%, simulating an increase in speaker size of approximately 10%. This acoustic manipulation is similar to the one carried out to make voices such as those of ‘Alvin and the Chipmunks’ sound small, but in reverse. (and pitch was not affected) By manipulating the spectral envelope of each word, we created two versions, the original and a manipulated version intended to ‘sound bigger’.

Each listener responded to all 278 stimuli (139 speakers x 2 resonance levels), for a total of 4170 observations across all listeners (15 listeners x 278 stimuli). Stimuli were presented one at a time, randomized along all stimulus dimensions. This means that tokens were thrown in one big pile and selected at random in a way that a stimulus was never predictable based on the previous one. For each trial, listeners were presented with a single word at random and were asked to:

1. Indicate whether they thought the speaker was a “boy 10-12 years old”, a “girl 10-12 years old”, a “man 18+ years old”, or a “woman 18+ years old”. This is the **apparent speaker category**.
2. Estimate the height of the speaker in feet and inches (converted to centimeters for this the discussion in this book). This is the **apparent speaker height**.

Our intention is to analyze the apparent height judgments provided by listeners in order to better understand them. To do this we will use acoustic descriptions of the different speakers’ voices, focusing on their fundamental frequency of their speech, and the vocal-tract length implied by their speech (estimated using the method described in the appendix). In addition, we will use the judgments made by listeners regarding the age and gender of the listener to better understand their use of acoustic in speaker height estimation.

### 1.3.3 Our research questions

This experiment is meant to investigate how listeners use speech acoustics to estimate the height of unknown talkers. Also, the results will let us investigate the possible relationship between the perception of talker size and the perception of talker category. Specific research questions will be discussed in each chapter, however, a general overview will be provided here. The expectations to be outlined below are based on the empirical relationships between these measurements and characteristics outlined above, and shown in Figure 1.1. The assumption is that listeners are familiar with the relationships between height and speech acoustics, and ‘somehow’ use the information in speech to guess the height of speakers. So, for example, if we know that a speaker with an f0 of 100 Hz is usually an adult male and is usually about 176 cm tall, we expect listeners will identify speech stimuli with an f0 near 100 Hz as produced adult male speakers who are about 176 cm tell.

Listeners were asked to provide two responses, speaker height and speaker group. The four speaker groups can be split according to two characteristics: The age of the group and the gender of the group (boy = male child, girl = female child, man = male adult, woman = female adult). So, we can consider that listeners reported the height, the age and the gender of the speaker, for each sound they listened to. In general, we expect that the perception of maleness will be associated with the perception of taller speakers, in particular for older speakers. The perception of adultness should be associated with taller speakers for either gender.

In terms of the acoustic variation in speaker voices, lower frequencies, whether f0 or resonances, are expected to be associated with taller and older speakers. For postpubescent speakers, low frequencies, particularly in f0, can also be an indicator of maleness. It is possible that the acoustic information in voices might be used differently based on the apparent class of the speaker. For example, maybe listeners used f0 one way when they think the speaker was an adult and another way when they think the speaker was a child. In addition, it is possible that different listeners used the acoustic information in ways that were systematic within-listener, but which differ arbitrarily from each other between listener.

### 1.3.4 Our experimental data

The data associated with this experiment is available in the bmmb package (discussed in section X), and can be accessed using the code below:

library ("bmmb")  
##   
## Attaching package: 'bmmb'  
## The following objects are masked \_by\_ '.GlobalEnv':  
##   
## findauc, h95  
data (height\_exp)

The code above loads our data and places it into our workspace in a, object called height\_exp. Below we use the head function see the first six lines of the data for the experiment. Our data is in *long* format so each row is a different individual observation and each column is a different piece of information regarding that observation. Each individual trial (a single row) represents an individual listener’s response to a single stimulus word played to them. So, we know that this data frame has 4170 rows to represent the 4170 observations in our data.

# see first 6 rows  
head (height\_exp)  
## L C height R S C\_v vtl f0 dur G A  
## 279 01 b 121.9 a 1 b 12.2 277 237 m c  
## 280 01 b 132.1 b 1 b 12.2 277 237 m c  
## 281 01 g 129.3 a 2 b 12.4 287 317 f c  
## 282 01 w 156.2 b 2 b 12.4 287 317 f a  
## 283 01 b 141.0 a 3 b 11.6 219 277 m c  
## 284 01 b 130.6 b 3 b 11.6 219 277 m c

If this were data that you collected and wanted to analyze, you would likely have it somewhere on your hard drive in a csv file, or some equivalent data file. If you were to open this data in Excel (or a similar software) you would see your data arranged in rows and columns. Below we write our data out as a csv file so that we can have a look at it outside of R.

write.csv (height\_exp, "height\_exp.csv", row.names = FALSE)

We can get more information about our data using the str function, which tells us that our data is stored in a data frame. A **data frame** is a collection of vectors that can be of different types, but which must be of the same length. A **vector** is a collection of elements of the same kind. Below, we see that the str function tells us about the vectors comprised by our data frame.

str (height\_exp)  
## 'data.frame': 4170 obs. of 11 variables:  
## $ L : chr "01" "01" "01" "01" ...  
## $ C : chr "b" "b" "g" "w" ...  
## $ height: num 122 132 129 156 141 ...  
## $ R : chr "a" "b" "a" "b" ...  
## $ S : num 1 1 2 2 3 3 4 4 5 5 ...  
## $ C\_v : chr "b" "b" "b" "b" ...  
## $ vtl : num 12.2 12.2 12.4 12.4 11.6 11.6 11.9 11.9 12.1 12.1 ...  
## $ f0 : int 277 277 287 287 219 219 260 260 244 244 ...  
## $ dur : int 237 237 317 317 277 277 318 318 242 242 ...  
## $ G : chr "m" "m" "f" "f" ...  
## $ A : chr "c" "c" "c" "a" ...

We see three kinds of vectors in our data: int indicating that the vector contains integers, num indicating that the vector contains real numbers, and chr indicating that the vector contains elements made up of characters (i.e. letters or words), or numbers being treated as if they were letters (i.e. as symbols with no numeric value). For example our data contains a column called height that contains the numeric values 122, 132, 129, 156, 141, and so on. The information represented in each column is:

* L: A number from 1-15 indicating which *listener* responded to the trial, being treated as a character.
* C: A letter representing the speaker *category* (b=boy, g=girl, m=man, w=woman) reported by the listener for each trial.
* height: A number representing the *height* (in centimeters) reported for the speaker on each trial.
* R: A letter representing the *resonance* scaling for the stimulus on each trial. The coding is a (actual) for the unmodified resonance and b (big) for the modified resonance (intended to sound bigger).
* S: A number from 1-139 indicating which *speaker* produced the trial stimulus.
* C\_v: A letter representing the *veridical* (actual) speaker category (‘b’=boy, ‘g’=girl, ‘m’=man, ‘w’=woman) for each speaker for each trial.
* vtl: An estimate of the speaker’s *vocal-tract length* in centimeters.
* f0: The speaker’s average *fundamental frequency* (f0) measured in Hertz.
* dur: The *duration* of the vowel sound, in milliseconds.
* G: The *apparent gender* of the speaker indicated by the listener, f (female) or m (male).
* A: The *apparent age* of the speaker indicated by the listener, a (adult) or c (child).

We can access the individual vectors that make up our data frame in many ways. One way is to add a $ after the name of our data frame, and then write the name of the vector after. This is shown below for our vector of heights.

height\_exp$height

Calling the command above will write out the entire vector to your screen, all 2780 observations of height responses that make up our data. Using the head function will show you the first six elements of an object, and you can get specific elements of the vector using brackets as shown below.

# show the first six  
head (height\_exp$height)  
## [1] 121.9 132.1 129.3 156.2 141.0 130.6  
  
# show the first element  
height\_exp$height[1]  
## [1] 121.9  
  
## show elements 2 to 6  
height\_exp$height[2:6]  
## [1] 132.1 129.3 156.2 141.0 130.6

Below, we use two sets of brackets to retrieve the height vector using its position in the data frame (first example), or its name (second example).

head( height\_exp[[3]] )  
## [1] 121.9 132.1 129.3 156.2 141.0 130.6  
  
head( height\_exp[["height"]] )  
## [1] 121.9 132.1 129.3 156.2 141.0 130.6

We can also retrieve the height vector by using a single set of parentheses as shown below. This method relies on treating the data frame as a matrix whose elements are arranged on a grid. Each element of the grid can then be accessed by providing x and y grid coordinates in single brackets as in [x,y]. Below we retrieve the entire third column by specifying a column number (or name) but leaving the row number unspecified.

head( height\_exp[,3] )  
## [1] 121.9 132.1 129.3 156.2 141.0 130.6  
  
head( height\_exp[,"height"] )  
## [1] 121.9 132.1 129.3 156.2 141.0 130.6

Below we use the same method to recover the entire first row of the data frame, and then the second element of the first row (or, from another perspective, the first element of the second column).

height\_exp[1,]  
## L C height R S C\_v vtl f0 dur G A  
## 279 01 b 121.9 a 1 b 12.2 277 237 m c  
height\_exp[1,2]  
## [1] "b"

## 1.4 Variables

Each of the columns in the height\_exp data frame can be thought of as a different variable. **Variables** are placeholders for some value, whether we know it or not. For example I can say “my weight is pounds”, or “this data represents a response provided by experimental subject ”. Ir our data, our variables take on different values from trial to trial, and the values of these variables tell us about the different outcomes and conditions associated with the trial. In this section we are going to discuss different aspects of variables, especially as they pertain to the analysis of experimental data.

### 1.4.1 Populations and samples

Anything that varies from observation to observation in an unpredictable manner can be though of a **random variable**. For example, your exact weight varies from day to day around your ‘average’ weight. In principle, you could probably explain exactly why your weight varies from day if you were so inclined. However, in practice you are probably not exactly sure *why* your weight is a bit higher one day and a bit lower the next. So, your weight is a random variable not necessarily because it is *impossible* to know why it varies, but simply because you don’t currently have the means to predict its value on any given observation.

In order to answer questions about reasonable values for variables of interest, scientists often collect measurements of that variable. These measurements can help us understand the most probable values of this variable, and the expected range of the variable, even if its value for any given observation in unpredictable. For example, although you may not know your exact weight in any given day, if you weigh yourself with some regularity you may have enough observations to have a pretty good idea of what your weight might be tomorrow. In addition, your expectation may be so strong that a large deviation from it would be more likely to result in your buying a new scale than believing the measurement.

The measurements you make of a random variable are called **samples**. A sample is a finite set of observations that you actually have. The **population** is the (hypothetical) larger group of all possible observations that you are *actually* interested in. The population is the entire set of possible values of the random variable. For example, the population of “f0 produced by adult women in the United States” contains all possible values of f0 produced by the entire set of women from the United States. Our sample is the specific set of observations we have from our set of speakers.

Usually, a scientist will collect a sample to make inferences about the population. In other words, we are interested in the general behavior of the variable itself, not just of the small number of instances that we observed. For example, Hillenbrand et al. collected their data to make inferences about speakers of American English in general, and not because they were particularly interested in the specific speakers in their sample. Hillenbrand et al. collected speech samples from a relatively small sample of speakers to make inferences about the whole population of speakers in the United States. Similarly, we are not specifically interested in the opinions of the 15 listeners in our data, but about what their behavior might tell us about the population of human listeners in general.

### 1.4.2 Dependent and Independent Variables

We can make a very basic distinction between variables that we want to explain or understand, and variables that we *use* to explain and understand. The variables we want to explain are our **dependent variables**, they are usually the variables we measure or observe in an experiment. The variables that we use to explain and understand our measurements are our **independent variables** (sometimes called explanatory variables).

Dependent variables can often be **random**, which means their values are not knowable **a priori** (before observation). For example, you may have some expectation about what your weight might be before you get on a scale, but in general you can’t know exactly what it will say with certainty before collecting the observation. Although the exact values of our dependent variables can vary somewhat unpredictably from trial to trial, in the context of an experiment there is the general expectation that these values will *depend* in some way on the other variables in the experiment. For example, in this experiment we modified the stimuli so that some are expected to ‘sound’ bigger than others. As a result, the reported height we expect for any given trial *depends* on the value of the R (Resonance) variable in our data, among other things.

Variables that help predict the response (dependent) variables and are sometimes referred to as independent variables because their values are not considered to depend on those of the other predictors. More specifically, we can say the values of our independent variables are not assumed to depend on the values of the other variables in our experiment within the context of our experiment, or in a manner that directly relates to the relevant research questions.

Our experiment has two response variables: the apparent height (height) reported for each trial, and the apparent speaker category (C) reported for each trial. Our experiment also involves several variables that could be used to understand our responses (i.e. every other variable in the data). Whether a variable is dependent or independent depends on the research question and on the structure of the model more than on some inherent property of variables and data. For example, the data in height\_exp could be used to understand variation in voice pitch (f0) across speakers groups. In this case f0 would be the dependent variable and the veridical speaker category (C\_v) would be the independent variable. Another researcher may chose to model how perceived height varies as a function of f0 and speaker group. In this case height would be the dependent variable and f0 and C\_v would be the independent variables.

### 1.4.3 Categorical variables and ‘factors’

**Categorical** variables, also sometimes called **nominal variables**, are variables that take on some set of non-numeric, usually character values. Often, categorical variables are the labels that we apply to objects or groups of objects. For example, gender is a nominal variables with possible values of ‘male’ and ‘female’ among others. In our experiment data, C, S, L, R, and C\_v are nominal variables. Categorical predictors are often called **factors**. Factors can take on a limited number of values, called **levels**. For example if your factor is “word category” you factor levels may be “verb” and “noun” (among others). If your factor is “first language” your levels may be “Mandarin” and “Hindi”. The factors and factor levels involved in your experiment are selected (and named) arbitrarily by you.

A factor is actually a data type in R. It’s very similar to a vector of words but it has some additional properties that are useful. For example, consider our C\_v predictor, which tells us which category each speaker falls into. Initially it is a character vector. We see that the first few tokens are produced by boys (b), and that there is no numerical value associated with these letter labels. The unique function returns all unique labels in the vector, in the order that they appear in the vector.

# see the first 6 observations  
head (height\_exp$C\_v)   
## [1] "b" "b" "b" "b" "b" "b"  
  
# class starts as a character vector  
class (height\_exp$C\_v)   
## [1] "character"  
  
# no numerical values, you will see NAs  
head (as.numeric (height\_exp$C\_v))   
## Warning in head(as.numeric(height\_exp$C\_v)): NAs introduced by coercion  
## [1] NA NA NA NA NA NA  
  
# we can see the number of unique groups  
unique (height\_exp$C\_v)   
## [1] "b" "g" "m" "w"

We can turn the character vector C\_v into a factor vector C\_v\_f. The benefit of this is that these nominal labels now have associated numerical values. Many R functions turn your nominal (non-numeric) predictors into factors, and doing this yourself gives you control over how this will be handled.

# we can turn it into a factor in R  
height\_exp$C\_v\_f = factor(height\_exp$C\_v)   
  
# now it has official levels  
levels(height\_exp$C\_v\_f)   
## [1] "b" "g" "m" "w"  
  
# now each level has numerical values  
table (height\_exp$C\_v\_f, as.numeric (height\_exp$C\_v\_f))   
##   
## 1 2 3 4  
## b 810 0 0 0  
## g 0 570 0 0  
## m 0 0 1350 0  
## w 0 0 0 1440

By default, factor levels are ordered alphabetically. You can control this behavior by re-ordering the factor levels as below:

height\_exp$C\_v\_f = factor (height\_exp$C\_v, levels = c('w','m','g','b'))  
  
levels (height\_exp$height\_exp)  
## NULL  
  
# note that 'm' is now the second category  
table (height\_exp$C\_v\_f, as.numeric (height\_exp$C\_v\_f))   
##   
## 1 2 3 4  
## w 1440 0 0 0  
## m 0 1350 0 0  
## g 0 0 570 0  
## b 0 0 0 810

Although our factors seem to have an ‘order’ this is only because items can only be discussed and presented one at a time, and so there must be some order in our nominal variables at some level of organization. For example, when presenting effects and plotting figures, you literally do have to decide to show one effect first and another second. However, the ordering of factors is **exchangeable** meaning it does not in any way affect our analysis. For example, the listeners and speakers in our experiment received unique numbers. However, listener 1 is not the listener who ‘most’ has the quality of listener, and speaker 8 is not twice the speaker that speaker 4 is. In other words, although we must commit to some order in our factors in order to organize our data, this ordering is arbitrary and not meaningful.

There is a special kind of nominal variable called an **ordinal** variable where the ordering of the categories *is* meaningful. These variables are halfway between numbers and labels: They faithfully represent the order (**rank**) of categories but not the difference between them. For example, consider the first, second, and third place runners in a race. These are ordinal labels. You know who finished before/after who, but don’t know anything about how much of a difference there was between the runners. As a result, these variables seem to have some of the properties of numbers, while not being totally like ‘real’ numbers. We will discuss the prediction of ordinal dependent variables in more detail in Chapter X.

### 1.4.4 Quantitative variables

Unlike nominal variables, quantitative variables let us represent the relative ordering of different observations *and* the relative differences between different observations. Some examples of quantitative variables are time, frequency, and weight. In our experiment data, height is a quantitative dependent variable, and f0, vtl, and duration are quantitative independent variables.

A distinction is made between **continuous** and **discrete** quantitative variables. Continuous variables have infinitely small spaces between adjacent elements (like the real numbers), at least in principle. On the other hand discrete variables have gaps between the possible values of the variable, like the integers. For example, things like time are naturally continuous while things like counts are naturally discrete.

When we are using a quantitative variable as our dependent variable, there is usually the expectation that is is continuous rather than discrete. In practice all measures stored on computers are discrete and many continuous values (e.g. reaction times) can be measured with a maximal precision, resulting in discrete values. For example, a chronometer that measures reaction times to the millisecond contains only 1000 possible values between zero and one second. Similarly, human height is difficult to measure to much less than a centimeter of precision, making height measurements effectively discrete. Below are some more questions that will help you decide if you should treat a variable as quantitative, even if it ‘discrete’:

* Is the underlying value continuous? Many variables are discrete in practice due to limitations in measurement. However, if the underlying value is continuous (e.g., height, time) then this can motivate treating the measurement as a quantitative dependent variable since fractional values ‘make sense’. For example, even if you measure time only down to the nearest millisecond, a value of 0.5 milliseconds is possible and interpretable. In contrast, a value of 0.5 people is not.
* Is the variable on a ratio or interval scale? An interval scale means that distances are meaningful, and an ratio scale means that 0 is meaningful. This characteristic is a prerequisite for a quantitative value to be used as a dependent variable.
* Are there a large number (>50) of possible values the measured variable can take? For example a die can only take on 6 quantitative values, which is not enough.
* Are most/all of the observed values far from their bounds? Human height does not really get much smaller than about 50 cm and longer than about 220 cm, so it is technically bounded. However, in most cases our observations are expected to not be clustered at the boundaries.

If you answered yes to all or most of these questions, it is probably ok to treat a quantitative variable as if it were continuous, though this determination really needs to be made on a case by case basis.

### 1.4.5 Logical variables

Before finishing with variables, we need to talk about one type that does not appear in our data, but that will come up often. **Logical** variables in R can only take one of two values: TRUE and FALSE. Below we use two equal signs to test for the equality of two values, and != to check for an inequality. Notice that we can check for the equality of numbers or characters.

2 == 1  
## [1] FALSE  
"hello" == "hello"  
## [1] TRUE  
"hello" != "hello"  
## [1] FALSE

We can also check for inequalities between numbers:

2 > 1  
## [1] TRUE  
2 >= 1  
## [1] TRUE  
2 < 1  
## [1] FALSE  
2 >= 1  
## [1] TRUE

One useful fact is that the logical values of TRUE and FALSE have numeric values of 1 and 0, as seen below. In each case, TRUE is equal to 1 so the expression evaluates to 2.

TRUE + 1  
## [1] 2  
(2 == 2) + 1  
## [1] 2

When logical operators are applied to vectors, the operation is evaluated for each element of the vector, as below, and a vector of logical values is returned. When combined with the numeric values of logical variables, this means that we can easily calculate the number of times a certain condition was met in the vector.

# is the values less than or equal to 3?  
c(1,2,3,4,5,6,7,8,9,10) <= 3  
## [1] TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

Below, we find whether each element of the vector is or is not greater then or equal to three. This results in a vector of logical values equivalent to a vector of ones and zeros. When we find the sum of the vector of logical values, we find the number of times in which the condition was met. Below, we see that three of the elements in this vector satisfy our condition.

logical\_vector = c(1,2,3,4,5,6,7,8,9,10) <= 3  
  
as.numeric (logical\_vector)  
## [1] 1 1 1 0 0 0 0 0 0 0  
  
sum (logical\_vector)  
## [1] 3  
sum (c(1,2,3,4,5,6,7,8,9,10) <= 3)  
## [1] 3

There is one other very important use for vectors of logical values, and this is to extract subsets of your data that meet certain conditions. Below we create a vector of logical values that indicate whether the f0 for a trial is below 175 Hz or not. We can see that this vector has 4170 elements, one for every row in our data, and that 1290 trials satisfied our condition. This is nothing more than a bigger version of the same process we just carried out above with our logical\_vector.

f0\_idx = height\_exp$f0 < 175  
str (f0\_idx)  
## logi [1:4170] FALSE FALSE FALSE FALSE FALSE FALSE ...  
sum (f0\_idx)  
## [1] 1290

Recall that we can access individual rows of our data frames by placing this information before a comma, inside brackets following the name of the data frame (as seen below). When we use a logical vector in this way, the effect is to include every row that equals TRUE and to omit every row that equals FALSE in the vector. Below we use our f0\_idx vector to create a new data frame called low\_f0 containing only productions with f0 below 175 Hz.

low\_f0 = height\_exp[f0\_idx,]  
str(low\_f0)  
## 'data.frame': 1290 obs. of 12 variables:  
## $ L : chr "01" "01" "01" "01" ...  
## $ C : chr "m" "m" "m" "b" ...  
## $ height: num 172 177 176 188 178 ...  
## $ R : chr "a" "b" "a" "b" ...  
## $ S : num 47 47 48 48 49 49 50 50 51 51 ...  
## $ C\_v : chr "m" "m" "m" "m" ...  
## $ vtl : num 14.8 14.8 15.6 15.6 15.5 15.5 14.5 14.5 15 15 ...  
## $ f0 : int 172 172 108 108 96 96 134 134 122 122 ...  
## $ dur : int 339 339 236 236 315 315 240 240 241 241 ...  
## $ G : chr "m" "m" "m" "m" ...  
## $ A : chr "a" "a" "a" "c" ...  
## $ C\_v\_f : Factor w/ 4 levels "w","m","g","b": 2 2 2 2 2 2 2 2 2 2 ...  
max(low\_f0$f0)  
## [1] 172

We can use the ! operator, which basically means ‘not’ to flip each TRUE to FALSE (and vice versa). When f0\_idx is flipped to select a subset of a data frame, the result is to select those rows where speaker f0 is *above* 175 Hz.

high\_f0 = height\_exp[!f0\_idx,]  
str(high\_f0)  
## 'data.frame': 2880 obs. of 12 variables:  
## $ L : chr "01" "01" "01" "01" ...  
## $ C : chr "b" "b" "g" "w" ...  
## $ height: num 122 132 129 156 141 ...  
## $ R : chr "a" "b" "a" "b" ...  
## $ S : num 1 1 2 2 3 3 4 4 5 5 ...  
## $ C\_v : chr "b" "b" "b" "b" ...  
## $ vtl : num 12.2 12.2 12.4 12.4 11.6 11.6 11.9 11.9 12.1 12.1 ...  
## $ f0 : int 277 277 287 287 219 219 260 260 244 244 ...  
## $ dur : int 237 237 317 317 277 277 318 318 242 242 ...  
## $ G : chr "m" "m" "f" "f" ...  
## $ A : chr "c" "c" "c" "a" ...  
## $ C\_v\_f : Factor w/ 4 levels "w","m","g","b": 4 4 4 4 4 4 4 4 4 4 ...  
min(high\_f0$f0)  
## [1] 175

## 1.5 Inspecting our data

After running an experiment but before running any kind of data analysis, you should inspect the patterns in your data. This gives you an opportunity to make sure your results make sense and that the data reflects the experimental structure, and results, you expect.

### 1.5.1 Inspecting caterogical variables

One of the most useful functions for understanding the distribution of categorical variables is the table function. This function make a **cross tabulation** (or **contingency table**) of the variables passed to the function. If a single factor is passed, the function returns the number of times each level of the factor is found in the data. Since each of our listeners listened to 278 stimuli, we expect that each level of the factor L (representing listeners) will appear 278 times in our data, confirmed below.

We can use this approach to confirm basic expectations about our data, and to rule out problems with the design of the experiment. This is always a good idea since mistakes happen, and sometimes only get noticed when attempting to process the data. For example, if any of the levels below appeared more than or fewer than 278 times, we would have a problem.

table (height\_exp$L)  
##   
## 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15   
## 278 278 278 278 278 278 278 278 278 278 278 278 278 278 278

We can also provide two (or more) factors at a time and the table function will return counts for every combination of factor levels. The table below reflects the fact that each listener heard 54 boys, 38 girls, 90 men, and 96 women, for a total of 276 total responses. When you provide multiple factors to table, it will vary the first factor along the rows of the table and the second factor along the columns of the table. If a third factor is provided, it makes a different table for factors one and two, for each level of factor three. More and more factors can be provided to the function, but these tables before harder and harder to work with.

table (height\_exp$C\_v, height\_exp$L)  
##   
## 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15  
## b 54 54 54 54 54 54 54 54 54 54 54 54 54 54 54  
## g 38 38 38 38 38 38 38 38 38 38 38 38 38 38 38  
## m 90 90 90 90 90 90 90 90 90 90 90 90 90 90 90  
## w 96 96 96 96 96 96 96 96 96 96 96 96 96 96 96

Below we see that unlike our veridical categories, the distribution of *apparent* speaker categories varies across listeners. This is because the equal distribution of speakers for each listener is an aspect of the experimental design. However, how listeners interpreted each voice, whether they though it sounded like a boy or girl for example, may vary across individual listeners.

table (height\_exp$C, height\_exp$L)  
##   
## 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15  
## b 59 89 59 32 51 12 95 98 63 40 68 58 59 90 38  
## g 24 45 44 67 44 60 23 16 43 34 24 29 34 50 61  
## m 90 89 99 88 91 95 88 88 89 95 86 90 92 84 97  
## w 105 55 76 91 92 111 72 76 83 109 100 101 93 54 82

We can visualize relationships between categorical variables using a mosaic plot. In figure 1.2 we mosaic plots representing the two tables shown immediately above. Mosaic plots use rectangles of different sizes to reflect the relative frequencies of different combinations of categorical variables. For example, in the left mosaic plot we see that the size of the rectangle for each category is identical across listeners. This tells us these variables do not affect each other: changing the listener does not affect the distribution of veridical speaker class in any way. In contrast, the distribution of apparent speaker class is affected by the listener and this is shown in the right plot where columns differ randomly from each other.

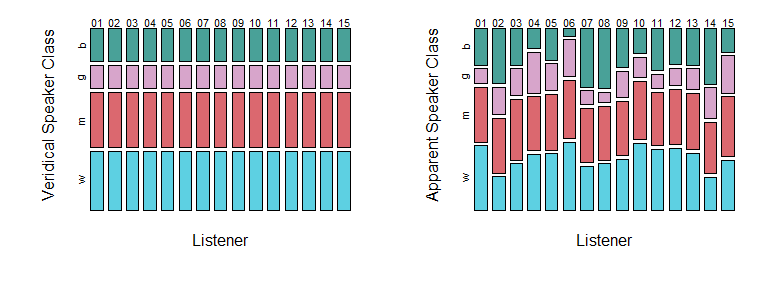


Figure 1.2: Comparisons of mosaic plots showing variables that do not (left), and do (right), affect each other.

Below we make a three-dimensional table, and inspect the table and each dimension. Notice that to index the table along the third dimension we need to add two commas inside the brackets.

tmp\_tab = table (height\_exp$C, height\_exp$L, height\_exp$R)  
tmp\_tab  
tmp\_tab[,,1]  
tmp\_tab[,,2]

When we plot the relationship between apparent speaker class, listener, and resonance, we see a three-way relationship between the variables. First, we see that the chances of observing different speaker categorizations depends on the listener. Second, we see that the chances of observing each category depends on resonance. And third, we see that the effect of resonance potentially affects each listener a somewhat different way. The first chapters of this book will focus on understanding patterns in continuous variables. However, we will discuss the prediction and modeling of categorical dependent variables beginning in Chapter X.

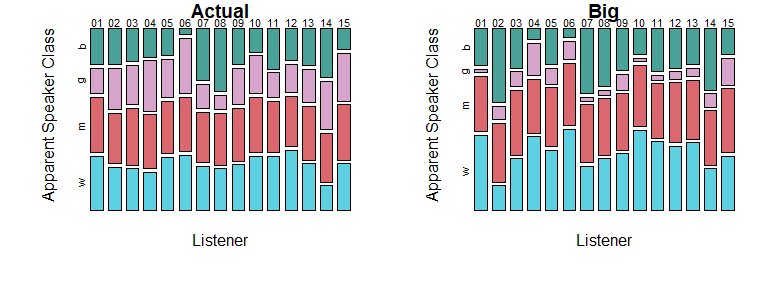


Figure 1.3: Mosaic plots highlighting a three-way relationship: The two-way relationship varies as a function of the third variable (indicated along the top of each plot).

### 1.5.2 Inspecting quantitative variables

Using R, we can easily find useful information about any quantitative variable. Below, we calculate the sample mean, the number of observations, the sample standard deviation, and some important quantiles for our speaker height judgments. The **quantiles** of a set of values such that a given percentage of observations fall above and below the value. Quantiles are found by ordering the observations and selecting the observation that is greater than of the sampled values and less than () of the sampled values. For example, the 50% quantile, also called the **median**, is the value such that 50% of the distribution is below it and 50% is above it, and the 25% quantile (the *first quartile*) is the value such that 25% of the distribution is below it and 75% is above it.

# calculate the mean  
mean (height\_exp$height)  
## [1] 162.7644  
  
# find the number of observations  
length (height\_exp$height)  
## [1] 4170  
  
# find quantiles  
quantile (height\_exp$height)  
## 0% 25% 50% 75% 100%   
## 106.7 154.9 164.8 173.5 198.1

We can use this information to make some basic, and potentially useful statements about our data. The mean and median are 162.8 and 164.8 cm respectively, and height values range from 106.7 to 198.1 cm. However, there are not many observations at the extremes, and 50% of values are between 154.9 and 173.5 cm. We know this because these are the values of the first and third **quartiles**, the 25% quantiles that divide our distribution into four equal parts. Since , we know that 50% of the distribution of observations must fall inside of these boundaries. We can look at the distribution of apparent height judgments in several ways, as seen in Figure 1.4. In the top row each point indicates an individual production. Points are jittered (randomly shifted) along the y axis to make them easier to distinguish so that dense and sparse locations can be compared. In the middle row we see a **box plot** of the same data. The edges of the box correspond to the 25% and 75% quantiles of the distribution, and the line in the middle of it corresponds to the median. So, the box spans the **interquartile range** of your observations and 50% of observations are contained in the box.The boxplot **whiskers** extend from the edge of the boxplots. By default, these extend out 1.5 times the interquartile range. These whiskers are simply intended to give you an estimate of the amount of ‘typical’ variation in your sample. Beyond the whiskers we see individual **outliers**, points considered to be substantially different from the rest of the sample. We can see that the boxplot does a good job of summarizing the information in the top plots, and provides information related to both average f0 values and to the expected variability in these values.

The bottom row presents what is knows as a **histogram** of the same data. The histogram divides the x axis into a set of discrete sections (‘bins’), and gives you the count (or frequency) of observations in each bin. Bins with lots of observations are relatively taller (more *dense*) than bins with fewer observations in them. As a result, histograms can be used to summarize where observations tend to be. For example, we can see that the bins under the interquartile range have the most observations, and that values further from the mean value become increasingly less frequent. In addition, histograms can provide us with information that boxplots can’t. For example, in the left column we see that our distribution of height judgments actually has two distinct peaks, with a little gap in the middle. This information does not really come across in the boxplot representation of the same data.

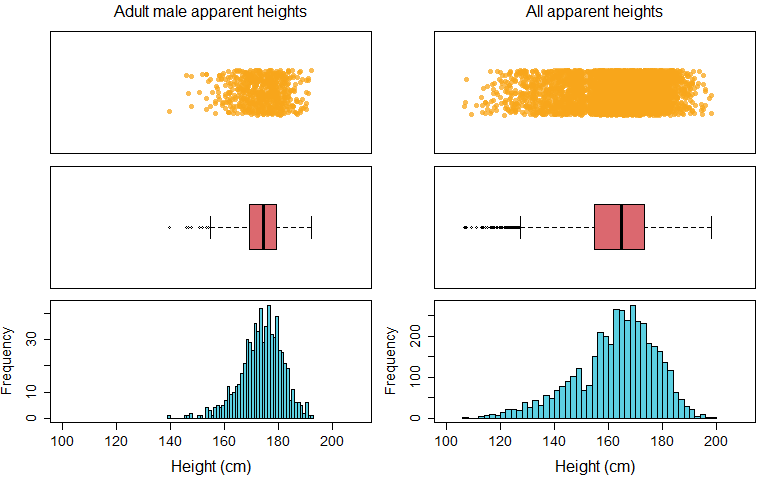


Figure 1.4: Each row presents data in a different way, with each column containing the same data across rows.

**Scatter plots** are plots that represent two variables at a time using a set of points on a coordinate space. Each point represents a single observation, the x-axis location represents the value of one variable, and the y-axis location represents the value of the other variable. Scatter plots are useful to understand relationships between continuous predictors. Below we consider the relationships between our quantitative predictors using a pairs plot (pairs). A pairs plot creates scatter plots for all pairs of quantitative variables provided, resulting in plots for variables. Each plot below contains a single point for each different stimulus used in this experiment (height values represent averages across all listeners).

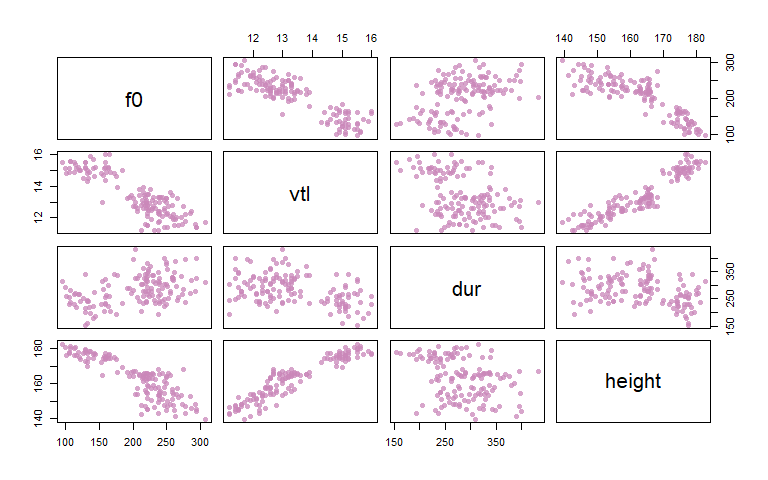


Figure 1.5: A pairs plot of the continuous variables in our data, showing different sorts of relationships between our variables.

In the plot above we can see several apparent relationships between our quantitative variables. For example, pitch (f0) and vocal-tract length (vtl) are *negatively* related. This means that as the value of f0 increases (left to right), the value of vtl decreases (top to bottom). In other words, if the f0-vtl relationship were a hill it would have a negative, decreasing, slope. In contrast we see that height and vtl enter into a positive relationship: As you increase vtl, height also increases. Finally, we see that duration (dur) and height do not seem to have much of a relationship. Unlike the other two scatterplots which looked a bit like ramps or lines, the scatter plot of dur and height resembles a Rorschach test inkblot. This suggests either that these two variables are not strongly related, or that the nature of the relationship is more complicated than what can be understood using these simple plots.

### 1.5.3 Exploring continuous and catgorical variables together

We can also consider the relationships between our quantitative and categorical variables. We can use the boxplot function as below:

boxplot (y ~ factor)

To make a set of boxplots for the variable y. The function call above will create a plot with a separate box for each level of the factor in the function call. In figure 1.6, we see different quantitative variables organized according to veridical speaker category. For example, the left panel shows the distribution of observations of f0 for boys, girls, men, and women respectively. In this case the differences between the boxplots for each level of the factor tell us about the values of f0 usually observed for speakers in that category.

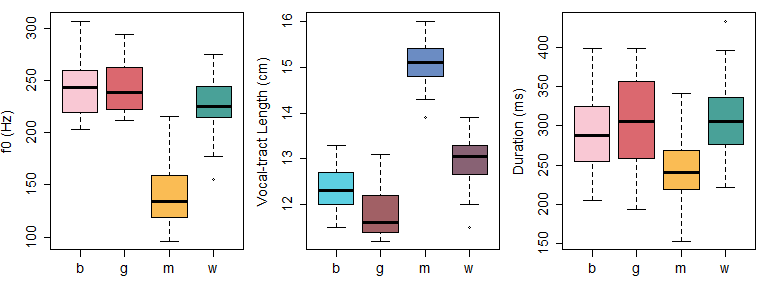


Figure 1.6: Boxplots showing the distribution of different quantitative variables in our data according to the veridical speaker categories of boy (b), girl (g), man (m), and woman (w).

Another way to think of the relationships between our categorical and quantitative variables is using the plot in figure 1.7. In the scatter plot below, each point indicates a single speaker from our experiment, and the position of each point is determined by the f0 and vocal-tract length of the speaker. However, rather then plot using symbols, each point is labeled using a letter which indicates the veridical category that the speaker falls into. Using a plot like the one below helps us understand the relationship between our important acoustic predictors and our speaker categories. For example, it is clear that adult males are fairly distinct acoustically compared to the other speaker categories. In addition, it seems that boys, girls, and women are easier to separate along the vocal-tract length dimension than the f0 dimension. What we mean by this is that it would be easier to draw horizontal lines separating the groups than vertical lines separating the groups.

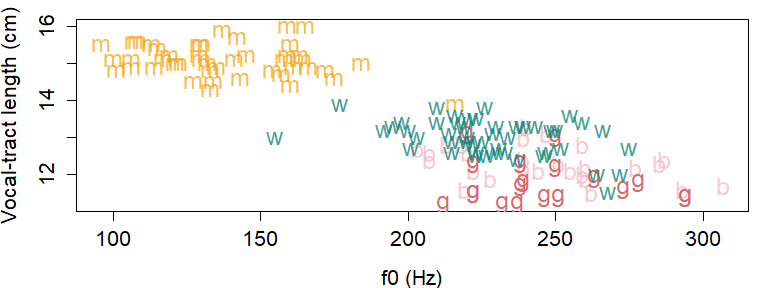


Figure 1.7: Speakers plotted according to their fundamental frequency (f0) and vocal-tract length. Letters indicate if speaker is a boy (b), girl (g), man (m), or woman (w).