Friedrich-Alexander-Universität Erlangen-Nürnberg

Lab Course

Statistical Methods for Audio Experiments

International Audio Laboratories Erlangen

Prof. Dr.-Ing. Jürgen Herre

Friedrich-Alexander Universität Erlangen-Nürnberg
International Audio Laboratories Erlangen
Lehrstuhl Semantic Audio Processing
Am Wolfsmantel 33, 91058 Erlangen

juergen.herre@audiolabs-erlangen.de

Authors:

   Fabian-Robert Stöter,
   Michael Schöffler

Tutors:

   Fabian-Robert Stöter,
   Michael Schöffler

Contact:

   Fabian-Robert Stöter, Michael Schöffler
   Friedrich-Alexander Universität Erlangen-Nürnberg
   International Audio Laboratories Erlangen
   Lehrstuhl Semantic Audio Processing
   Am Wolfsmantel 33, 91058 Erlangen
   `fabian-robert.stoeter@audiolabs-erlangen.de`
   `michael.schoeffler@audiolabs-erlangen.de`

Statistical Methods for Audio Experiments, © April 13, 2015

Statistical Methods for Audio Experiments

A part of this lab course material is taken from the *OpenIntro Statistics* Book [1] which is online for free at `www.openintro.org`. We encourage you to look into the book if you feel that you have problems understanding the basics of statistics.

1   Motivation

This course intends to teach students the basics of experimental statistics as it is used for evaluating auditory experiments. Listening tests or experiments are a crucial part of assessing the quality of audio systems. There is currently no system available to give researchers and developers the possibility to evaluate the quality of audio systems fully objectively. In fact the best evaluation instrument is the human ear. Therefore the success of audio coding systems such as *MP3* or *AAC* would not have been possible without hundreds of hours of music and speech content assessed by expert listeners using professional equipment. Since only fair and unbiased comparisons between codecs guarantee that new developments are more preferred than the previous system, it is important to bring fundamental knowledge of statistics into the evaluation process to address the main problems of experimental tests, such as uncontrolled environments, subpar headphones or loudspeaker reproduction systems, listeners who have no experience to listening tests and so on.

2   Elementary Descriptive Statistics

2.1   Box plots, quartiles, and the median

A *box plot* summarizes a data set using five statistics while also plotting unusual observations. Figure 1 provides a vertical dot plot alongside a box plot.
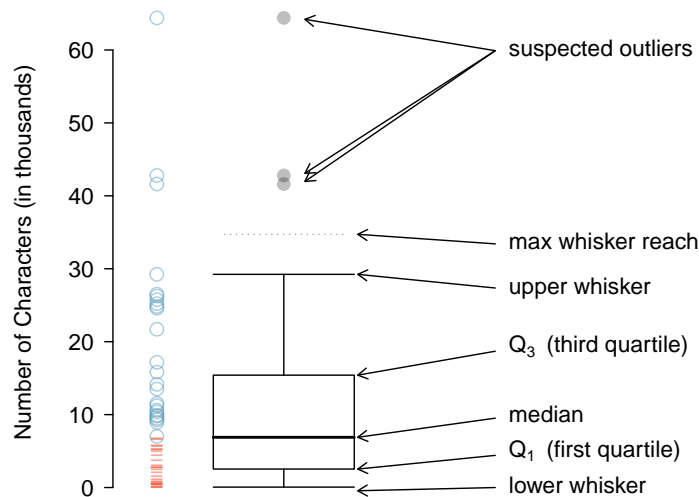

Figure 1: A vertical dot plot next to a labeled box plot for the number of characters in 50 emails. The median splits the data into the bottom 50% and the top 50%, marked in the dot plot by horizontal dashes and open circles, respectively.

The first step in building a box plot is drawing a dark line denoting the *median*, which splits the data in half. Figure 1 shows 50% of the data falling below the median (dashes) and other 50% falling above the median (open circles). There are 50 character counts in the data set (an even number) so the data are perfectly split into two groups of 25. We took the median in this case to be the average of the two observations closest to the 50th percentile. When there are an odd number of observations, there will be exactly one observation that splits the data into two halves, and in this case that observation is the median (no average needed).

Median: the number in the middle

If the data are ordered from smallest to largest, the *median* is the observation right in the middle. If there are an even number of observations, there will be two values in the middle, and the median is taken as their average.

The second step in building a box plot is drawing a rectangle to represent the middle 50% of the data. The total length of the box, shown vertically in Figure 1, is called the *interquartile range* (IQR, for short). Similar to the standard deviation, it is a measure of variability in data. The more variable the data, the larger the standard deviation and IQR. The two boundaries of the box are called the *first quartile* (the 25th percentile, i.e. 25% of the data fall below this value) and the *third quartile* (the 75th percentile), and these are often labeled Q_1 and Q_3, respectively.

Interquartile range (IQR)

The IQR is the length of the box in a box plot. It is computed as

$$IQR = Q_3 - Q_1$$

where Q_1 and Q_3 are the 25th and 75th percentiles.

Homework Excercise 1

What percent of the data fall between Q_1 and the median? What percent is between the median and Q_3?

Extending out from the box, the *whiskers* attempt to capture the data outside of the box, however, their reach is never allowed to be more than $1.5 \times IQR$. While the choice of exactly 1.5 is arbitrary, it is the most commonly used value for box plots. They capture everything within this reach. In Figure 1, the upper whisker does not extend to the last three points, which is beyond $Q_3 + 1.5 \times IQR$, and so it extends only to the last point below this limit. The lower whisker stops at the lowest value, 0, since there is no additional data to reach; the lower whisker's limit is not shown in the figure because the plot does not extend down to $Q_1 - 1.5 \times IQR$. In a sense, the box is like the body of the box plot and the whiskers are like its arms trying to reach the rest of the data.

Any observation that lies beyond the whiskers is labeled with a dot. The purpose of labeling these points – instead of just extending the whiskers to the minimum and maximum observed values – is to help identify any observations that appear to be unusually distant from the rest of the data. Unusually distant observations are called *outliers*.

Outliers are extreme

An *outlier* is an observation that appears extreme relative to the rest of the data.

Why is it important to look for outliers? Examining data for possible outliers serves many useful purposes, including

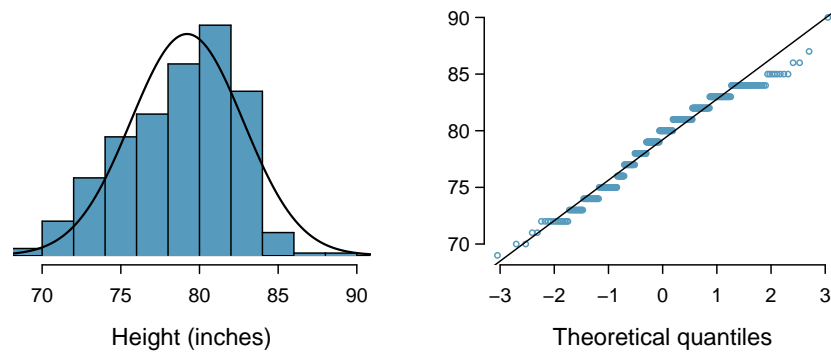1. Identifying strong skew in the distribution.

Figure 2: Histogram and normal probability plot for the NBA basketball **heights of players** from the 2008-9 season.

2. Identifying data collection or entry errors.

3. Providing insight into interesting properties of the data.

2.2 Distributions

2.2.1 Normal Distribution model

Among all the distributions, the normal distribution the most common.

> Many variables are nearly normal, but none are exactly normal. Thus the normal distribution, while not perfect for any single problem, is very useful for a variety of problems. We will use it in data exploration and to solve important problems in statistics.

The normal distribution model always describes a symmetric, unimodal, bell shaped curve. However, these curves can be different depending on the details of the model. Specifically, the normal distribution model can be adjusted using two parameters: mean and standard deviation.

2.2.2 Q-Q Plot

There are two visual methods for checking normality, which can be implemented and interpreted quickly. The first is a simple histogram with the best fitting normal curve overlaid on the plot, as shown in the left panel of Figure 2. The sample mean $\bar{x}$ and standard deviation s are used as the parameters of the best fitting normal curve. The closer this curve fits the histogram, the more reasonable the normal model assumption.

Another more common method is examining a *normal probability plot* or *quantile-quantile plot* shown in the right panel of Figure 2.

The Q-Q plot is a graphical method to determine if two variables follow the same statistical distribution. The plot evaluates the fit of sample data to the normal distribution or more generally, the for any theoretical distribution if you provide two variables. There is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal. If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. Because of the complexity of these calculations, Q-Q plots are generally created using statistical software.

3 Hypothesis testing

3.1 Experimental Variables

When designing experiments, three different types are considered differently. The independent variable is the variable which is controlled by the experimenter. E.g. when three different audio codecs are evaluated according to the audio quality, the audio codec is the independent variable. The second type of variable is the dependent variable. The dependent variable is the response of the participant. In the example used before, the reported subjective audio quality is the dependent variable. All other variables which might influence the experiment are of type "other". All "other" variables which might influence the experiment and are known by the experiment should be kept constant during the experiment. E.g. the headphones might influence the perceived audio quality. In such a situation, all participants must use the same type of headphones.

Variables of any type have different scales. The main scales are categorical, ordinal and quantitative. The levels of categorical variable can not be ordered naturally (e.g. the gender of the participants: "female" or "male"). Ordinal variables can be ordered naturally (e.g. age groups: "0-30 years old", "31 - 60 years old" and "61+ years old"). Quantitative variables are continuous variables. How many levels are needed for considering a variable as quantitative depends on the experimenter's point of view. Many experimenters consider a variable with 100 levels (e.g. 0-100 points) as quantitative. A mathematician would never do that.

When evaluating audio systems, the experimental design and planning plays an important role. Therefore it is needed to ensure that uncontrolled factors which can cause ambiguity in test results are minimized. As an example, if the actual sequences of audio items were identical for all the subjects in a listening test, then one could not be sure whether the judgements made by the subjects were due to that sequence rather than to the different levels of impairments that were presented. Accordingly, the test conditions must be arranged in a way that reveals the effects of the independent factors, and only of these factors.

Is your MP3 file actually of lower quality than the Original CD? It certainly depends on who you ask, what kind of equipment you are using, and etc.

The *null hypothesis* (H_0) often represents a skeptical perspective or a claim to be tested. The *alternative hypothesis* (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

The hypothesis testing framework is a very general tool, and we often use it without a second thought. If a person makes a somewhat unbelievable claim, we are initially skeptical. However, if there is sufficient evidence that supports the claim, we set aside our skepticism and reject the null hypothesis in favor of the alternative.

3.2 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

The *p-value* is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

If researchers are only interested in showing an increase or a decrease, but not both, use a one-sided test. If the researchers would be interested in any difference from the null value – an increase or decrease – then the test should be two-sided.

We evaluate the hypotheses by comparing the p-value to the significance level. Because the p-value is less than the significance level, we reject the null hypothesis. What we observed is so unusual with respect to the null hypothesis that it casts serious doubt on H_0 and provides strong evidence favoring the alternative H_A.

The following ideas below review the process of evaluating hypothesis tests with p-values:

- The null hypothesis represents a skeptic's position or a position of no difference. We reject this position only if the evidence strongly favors H_A.

- A small p-value means that if the null hypothesis is true, there is a low probability of seeing a point estimate at least as extreme as the one we saw. We interpret this as strong evidence in favor of the alternative.

- We reject the null hypothesis if the p-value is smaller than the significance level, α, which is usually 0.05. Otherwise, we fail to reject H_0.

- We should always state the conclusion of the hypothesis test in plain language so non-statisticians can also understand the results.

The p-value is constructed in such a way that we can directly compare it to the significance level (α) to determine whether or not to reject H_0.
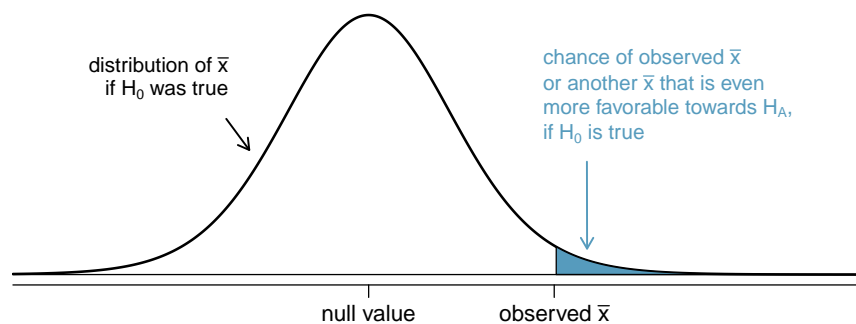

Figure 3: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_A under this distribution.

3.3   Two-sided hypothesis testing with p-values

We now consider how to compute a p-value for a two-sided test. In a two-sided test, *we shade two tails* since evidence in either direction is favorable to H_A.

Because the normal model is symmetric, the right tail will have the same area as the left tail. The p-value is found as the sum of the two shaded tails:

$$\text{p-value} = \text{left tail} + \text{right tail} = 2 \times (\text{left tail})$$

3.4 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test. A type 1 error is the incorrect rejection of a true null hypothesis. If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0. A type 2 error (or error of the second kind) is the failure to reject a false null hypothesis. If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

> The significance level selected for a test should reflect the consequences associated with Type 1 and Type 2 Errors.

3.5 Sample size and power

The Type 2 Error rate and the magnitude of the error for a point estimate are controlled by the sample size. Real differences from the null value, even large ones, may be difficult to detect with small samples. If we take a very large sample, we might find a statistically significant difference but the magnitude might be so small that it is of no practical value. In this section we describe techniques for selecting an appropriate sample size based on these considerations. Sample size computations are helpful in planning data collection, and they require careful forethought. The probability of rejecting the null hypothesis is called the *power*. The power varies depending on what we suppose the truth might be.

Cohen's d is a measure of effect size which indicates the amount of difference between two groups. Cohen's d is some sort of counter-point to significance tests and gives an indication of how big or small a significant difference is:

$$d = \frac{|mean_1 - mean_2|}{StandardDeviation}. \tag{1}$$

According to Cohen, Cohen $d = 0{,}2$ is defined as a *small effect*, $d = 0{,}5$ a *medium effect* and $d = 0{,}8$ is a considered to be a *large effect*.

3.6 Independent t-tests

3.6.1 One-sample means with the t distribution

We will see that the t distribution is a helpful substitute for the normal distribution when we model a sample mean $\bar{x}$ that comes from a small sample. While we emphasize the use of the t distribution for small samples, this distribution may also be used for means from large samples.

We use a special case of the Central Limit Theorem to ensure the distribution of the sample means will be nearly normal, regardless of sample size, provided the data come from a nearly normal distribution.

> **Central Limit Theorem for normal data**
> The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.
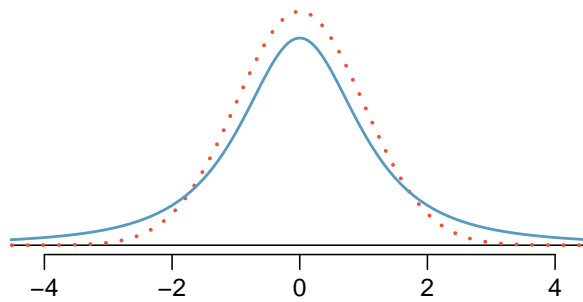
Figure 4: Comparison of a t distribution (solid line) and a normal distribution (dotted line).
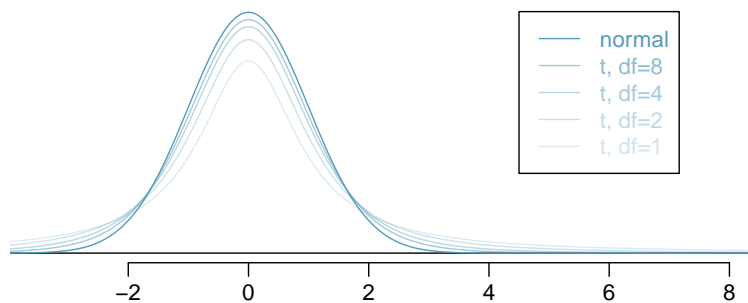

Figure 5: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

3.6.2 Introducing the t distribution

The second reason we previously required a large sample size was so that we could accurately estimate the standard error using the sample data. In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t distribution. A t distribution, shown as a solid line in Figure 4, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[2] These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The *degrees of freedom* describe the precise form of the bell shaped t distribution. Several t distributions are shown in Figure 5. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

> **Degrees of freedom (df)**
> The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

---

[2]The standard deviation of the t distribution is actually a little more than 1. However, it is useful to always think of the t distribution as having a standard deviation of 1 in all of our applications.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution.

3.7   The t distribution as a solution to the standard error problem

When estimating the mean and standard error from a small sample, the t distribution is a more accurate tool than the normal model. This is true for both small and large samples.

Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

To proceed with the t distribution for inference about a single mean, we must check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we should carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with $n-1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution with $df = 19 - 1 = 18$ degrees of freedom

3.8 t tests

In this section we use the t distribution for the difference in sample means. We will drop the minimum sample size condition and instead impose a strong condition on the distribution of the data.

It will be useful to extend the t distribution method to apply to a difference of means:

$$\bar{x}_1 - \bar{x}_2 \qquad \text{as a point estimate for} \qquad \mu_1 - \mu_2$$

First, we verify the small sample conditions (independence and nearly normal data) for each sample separately, then we verify that the samples are also independent. We can use the t distribution for inference on the point estimate $\bar{x}_1 - \bar{x}_2$.

The formula for the standard error of $\bar{x}_1 - \bar{x}_2$ also applies to small samples:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{2}$$

Because we use the t distribution, we need to identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we apply in the examples and exercises.[3]

Using the t distribution for a difference in means

The t distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t distribution and (2) the samples are independent. We estimate the standard error of the difference of two means using Equation (2).

---

[3]This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method.

3.9   Two sample t test

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t distribution. We are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B$. The standard error of the estimate can be calculated using Equation (2). Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

4   Analysis of Variance

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.

When these three conditions are met, we may perform an ANOVA to determine whether the data provide strong evidence against the null hypothesis that all the μ_i are equal.

Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means. We will soon learn that assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

Examine Figure 6. Compare groups I, II, and III. Can you visually determine if the differences in the group centers is due to chance or not? Now compare groups IV, V, and VI. Do these differences appear to be due to chance?

Any real difference in the means of groups I, II, and III is difficult to discern, because the data within each group are very volatile relative to any differences in the average outcome. On the other hand, it appears there are differences in the centers of groups IV, V, and VI. For instance, group V appears to have a higher mean than that of the other two groups. Investigating groups IV, V, and VI, we see the differences in the groups' centers are noticeable because those differences are large *relative to the variability in the individual observations within each group.*

The method of analysis of variance in this context focuses on answering one question: is the variability in the sample means so large that it seems unlikely to be from chance alone? This question is different from earlier testing procedures since we will *simultaneously* consider many groups, and evaluate whether their sample means differ more than we would expect from natural variation. We call this variability the *mean square between groups (MSG)*, and it has an associated degrees of freedom, $df_G = k - 1$ when there are k groups. The MSG can be thought of as a scaled variance formula for means. If the null hypothesis is true, any variation in the sample means is due to chance and shouldn't be too large.
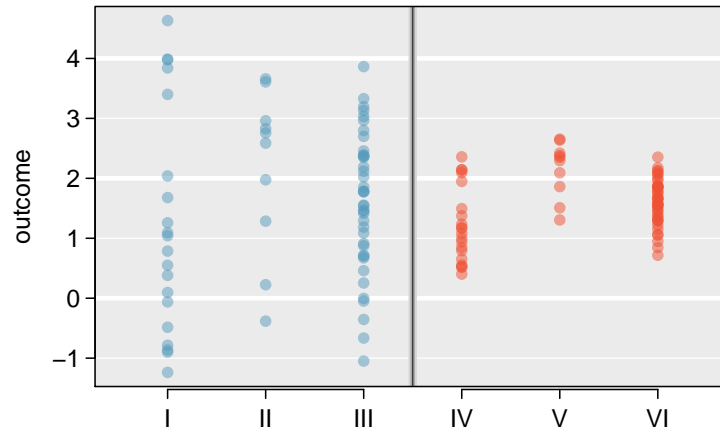
Figure 6: Side-by-side dot plot for the outcomes for six groups.

The mean square between the groups is, on its own, quite useless in a hypothesis test. We need a benchmark value for how much variability should be expected among the sample means if the null hypothesis is true. To this end, we compute a pooled variance estimate, often abbreviated as the *mean square error (MSE)*, which has an associated degrees of freedom value $df_E = n - k$. It is helpful to think of MSE as a measure of the variability within the groups.

When the null hypothesis is true, any differences among the sample means are only due to chance, and the MSG and MSE should be about equal. As a test statistic for ANOVA, we examine the fraction of MSG and MSE:

$$F = \frac{MSG}{MSE} \tag{3}$$

The MSG represents a measure of the between-group variability, and MSE measures the variability within each of the groups.

We can use the F statistic to evaluate the hypotheses in what is called an *F test*. A p-value can be computed from the F statistic using an F distribution, which has two associated parameters: df_1 and df_2. For the F statistic in ANOVA, $df_1 = df_G$ and $df_2 = df_E$.

The larger the observed variability in the sample means (MSG) relative to the within-group observations (MSE), the larger F will be and the stronger the evidence against the null hypothesis. Because larger values of F represent stronger evidence against the null hypothesis, we use the upper tail of the distribution to compute a p-value.

The F statistic and the F test Analysis of variance (ANOVA) is used to test whether the mean outcome differs across 2 or more groups. ANOVA uses a test statistic F, which represents a standardized ratio of variability in the sample means relative to the variability within the groups. If H_0 is true and the model assumptions are satisfied, the statistic F follows an F distribution with parameters $df_1 = k - 1$ and $df_2 = n - k$. The upper tail of the F distribution is used to represent the p-value.

Again, there are three conditions we must check for an ANOVA analysis: all observations must be independent, the data in each group must be nearly normal, and the variance within each group must be approximately equal.

4.1 Pairwise Comparisons

When we reject the null hypothesis in an ANOVA analysis, we might wonder, which of these groups have different means? To answer this question, we compare the means of each possible pair

of groups. For instance, if there are three groups and there is strong evidence that there are some differences in the group means, there are three comparisons to make: group 1 to group 2, group 1 to group 3, and group 2 to group 3. These comparisons can be accomplished using a two-sample t test, but we use a modified significance level and a pooled estimate of the standard deviation across groups.

There is strong evidence that the different means in each of the three classes is not simply due to chance. We might wonder, which of the classes are actually different? As discussed in earlier chapters, a two-sample t test could be used to test for differences in each possible pair of groups.

Multiple comparisons and the Bonferroni correction for α

The scenario of testing many pairs of groups is called *multiple comparisons*. The *Bonferroni correction* suggests that a more stringent significance level is more appropriate for these tests:

$$\alpha^* = \alpha/K$$

where K is the number of comparisons being considered (formally or informally). If there are k groups, then usually all possible pairs are compared and $K = \frac{k(k-1)}{2}$.

Caution: Sometimes an ANOVA will reject the null but no groups will have statistically significant differences. It is possible to reject the null hypothesis using ANOVA and then to not subsequently identify differences in the pairwise comparisons. However, *this does not invalidate the ANOVA conclusion*. It only means we have not been able to successfully identify which groups differ in their means.

The ANOVA procedure examines the big picture: it considers all groups simultaneously to decipher whether there is evidence that some difference exists. Even if the test indicates that there is strong evidence of differences in group means, identifying with high confidence a specific difference as statistically significant is more difficult.

5 MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA)

Subjective assessment methods, like listening tests, are an important tool when it comes to evaluate the quality of audio systems. One of these subjective assessment methods is MUSHRA which targets to evaluate intermediate quality levels of coding systems. The purpose of this chapter is to briefly describe the MUSHRA recommendation. For a comprehensive description of the MUSHRA methodology and the recommended statistical analysis, see [2].

In 2001, the ITU formally described in Recommendation BS.1534-0 the first version of MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) which is a test method for assessing intermediate audio quality [3]. A second revision of MUSHRA was just recently published in ITU-R Recommendation BS.1534-2 which introduced changes in the test design as well as in the analysis of the results [2].

In a MUSHRA test, assessors are presented with an open reference stimulus and a number of test stimuli (conditions). The conditions contain the hidden reference stimulus, at least two anchor stimuli (low quality anchor and mid quality anchor) and stimuli which were processed by the systems under test. MUSHRA allows to have a maximum of eleven conditions (eight systems under test, two anchors and the hidden reference) to be presented in one trial. Since the conditions are shown in random order, the assessor does not know which condition is a system under test, the hidden reference or an anchor. All conditions are rated relatively to the open reference stimulus. By adding the hidden reference stimulus to the conditions, an anchor for the highest rating is implicitly set. Without adding the hidden reference, the highest rating might also be given to a system under test which is the best system among all conditions but still causes some artifacts. With this design, it is more likely that only conditions are rated with the highest score which cannot be distinguished

from the open reference. Another purpose of the hidden reference is to find out whether an assessor would rate the hidden reference stimulus with a very high or the highest rating. If an assessor rates the hidden reference too low in too many trials, his or her ratings are excluded from the results by the post-screening process. BS.1534-2 recommends to add two anchor stimuli which are low-passed filtered versions of the reference stimulus. The low quality anchor has a cut-off frequency of 3.5 kHz and the mid quality anchor has a cut-off frequency of 7 kHz.

When assessing the conditions, the assessors are allowed to switch instantaneously between conditions and the open reference while listening. Besides switching between conditions, looping is a common strategy for assessing the conditions. Looping means that only an excerpt of the stimuli is marked and listened to repeatedly while assessing. For example, by marking a one-second critical part of a 10-second long stimulus, it is much easier to focus on the artifacts of this part caused by the system under test. The conditions are rated according to a continuous quality scale (CQS) ranging from 0 to 100. The scale is divided in five equal intervals, where each interval is labeled with an adjective (0-20: Bad, 20-40: Poor, 40-60: Fair, 60-80: Good, 80-100: Excellent). In most cases, the assessors are asked to rate the Basic Audio Quality (BAQ) of each condition. In ITU-R Recommendation BS.1534 Basic Audio Quality is defined as:

> "This single, global attribute is used to judge any and all detected differences between the reference and the object."

Although MUSHRA has been originally designed to evaluate the quality of audio coding systems, it is widely used for evaluating other types of audio systems.

5.1 Statistical Analysis

Statistical Analysis should always start with a visualization of the raw data. This may incorporate the use of histograms with a fitting curve for a normal distribution, boxplots (See Section 2.1), or quartile quartile plots (Q-Q-plots). Based on the visual inspection of these plots plus assumptions about the underlying population of the observed sample, it should be decided whether one may assume to have observed a normal distribution or not. If the fitting curve is clearly skewed, the histogram contains many outliers or the Q-Q-plot is not at all a straight line, one should not consider the sample as being normally distributed.The calculation of the median of normalized scores of all listeners that remain after post-screening will result in the median subjective scores.

5.2 Post-screening

Post-screening methods can be roughly separated into at least two classes:

- Stage 1 is based on the ability of the subject to make consistent repeated gradings

- Stage 2 relies on inconsistencies of an individual grading compared with the mean result of all subjects for a given item.

The Stage 1 post-screening method excludes subjects who do not consistently rate the hidden reference condition above 90. To quantify a post-hoc method that addresses this concern the following metric may be used: a listener should be excluded from the aggregated responses if he or she rates the hidden reference condition for >15% of the test items lower than a score of 90. Additionally, a subject should be excluded from the aggregated responses if he or she rates the mid-range anchor for more than 15% of the test items higher than a score of 90. If more than 25% of the subjects rate the mid-range anchor higher than a score of 90, this might indicate that the test item was not degraded significantly by the anchor processing. In this case assessors should not be excluded on the basis of scores for that item.

Stage 2 post-screening methods are additionally performed to collect more reliable aggregation. Stage 2 methods exclude subjects whose individual grades are inconsistent when compared with the

median result of all subjects for a given item and system. Box plot data visualization will provide indication of the existence and effect of outliers on the descriptive summaries of the data.

To remove outliers, it is recommended that those data points which are further apart than 1.5 times IQR from the first or third quartile, respectively, should be considered as outliers. The IQR shall be calculated and compared for each listener for each signal condition at test. If the listener's responses fall outside 1.5 * the upper or lower bound of the IQR of the aggregated listeners for 25% or more of the test conditions it suggests they are over-using the extremes of the rating scale. This would be a valid metric to consider exclusion across the testing group.

5.3   Main effect of conditon

For most cases, the main effect is of interest. If the ANOVA indicates a significant effect of condition, then the null hypothesis can be rejected that in the population the perceived audio quality is identical for all conditions (reference, coder 1 to k). In other words, the test indicates that in the population there are differences between the perceived audio quality of the audio systems. As a measure of effect size, it is not possible to use Cohen's (1988) d or one of its analogues, because d is not defined for a comparison of more than two means. In an ANOVA context, it is common to report a measure of association strength.

Following a significant test result for a main effect, it will then often be of interest to locate the origins of this effect. This can be achieved by computing paired comparison tests as described in section 4.1. For instance, one might be interested in whether the sound quality of a new coder differed from the sound quality of three established systems.

6   Regression

6.1   Linear Regression Models

Linear regression is a very powerful statistical technique. Many people have some familiarity with regression just from reading the news, where graphs with straight lines are overlaid on scatterplots. Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Figure 7 shows two variables whose relationship can be modeled perfectly with a straight line. The equation for the line is

$$y = 5 + 57.49x$$

Imagine what a perfect linear relationship would mean: you would know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process. For example, if we took family income x, this value would provide some useful information about how much financial support y a college may offer a prospective student. However, there would still be variability in financial support, even when comparing students whose families have similar financial backgrounds.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

$$y = \beta_0 + \beta_1 x \tag{4}$$

β_0, β_1
Linear model parameters

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). These parameters are estimated using data, and we write their point estimates as b_0 and b_1. When we use x to predict y, we usually call x the explanatory or *predictor* variable, and we call y the response.

It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in Figure 8. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The first plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The
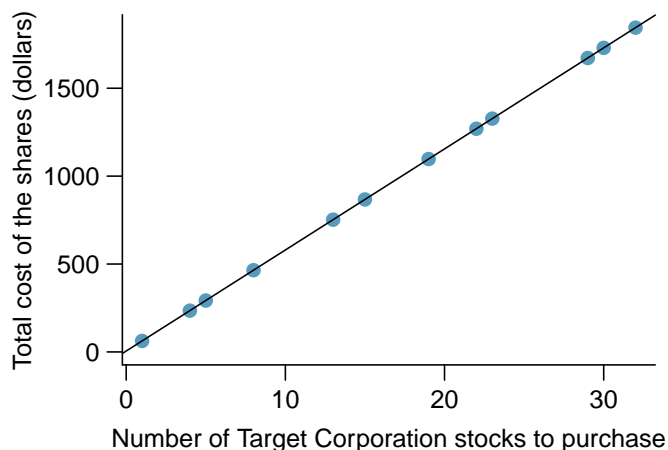
Figure 7: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker `TGT`, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear fit is perfect.
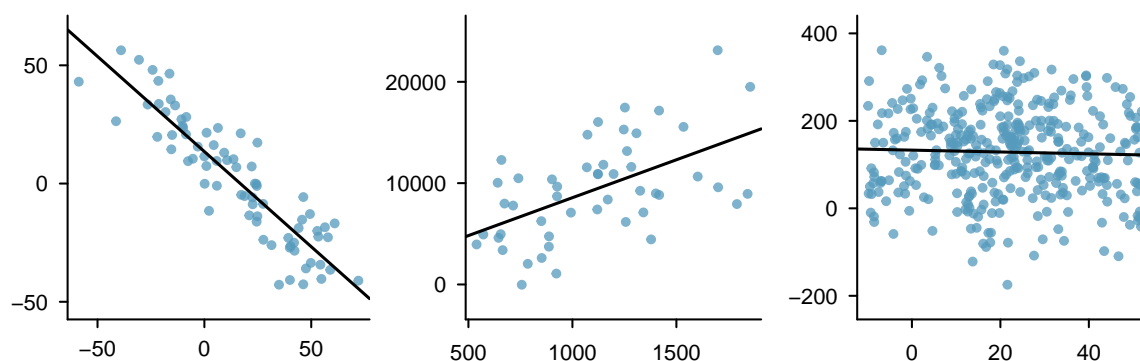

Figure 8: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

second plot shows an upward trend that, while evident, is not as strong as the first. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it. In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β_0 and β_1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less? As we move forward in this chapter, we will learn different criteria for line-fitting, and we will also learn about the uncertainty associated with estimates of model parameters.

We will also see examples in this chapter where fitting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 9 where there is a very strong relationship between the variables even though the trend is not linear.

7 Exercises

The actual lab course is set up in the R programming environment which perfect for statistical evaluation and includes a terrific number of build in statistics functions. The syntax is easy to learn but differs e.g. from Matlab.
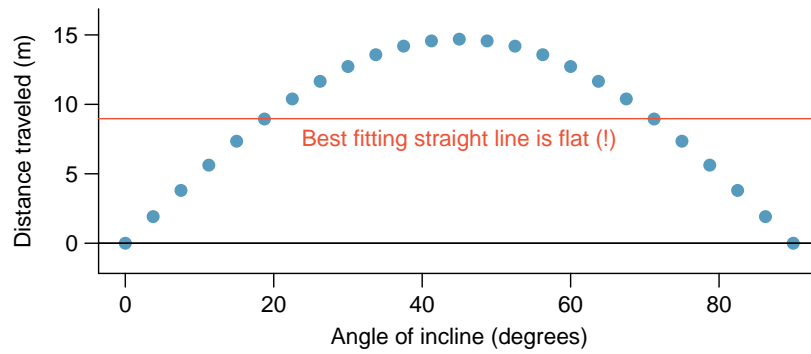
Figure 9: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

7.1 Additional Homework

Besides to the homework assignments that are marked in the script we ask you to go through the basics of R. For a quick guide we have attached a simple R tutorial in the lab course appendix. A more detailed course can be found at http://tryr.codeschool.com/

Homework Excercise 3

Read and work out the introduction to R and R-Studio attached to this document in section 10.

- Explain the use of the $-Operator in R.

- How does the following Matlab-Syntax translate in R:

  B = A(2:45,:)

- Write down a *hello world* example in R

- If *var*1 is the dependent variable of an experiment and *var*2 is the independent variable how does this translate in an R formula?

Part I

Exercises

8 Listening Test

Before you start with the programming exercises at least one student of each group should take part in a MUSHRA like listening test. The listening test can be conducted online through the browser at `http://work.audiolabs.uni-erlangen.de/aplab/`. Please use your headphones and follow the instructions. The results of the listening test will later be used in one of the exercises.

9 Programming

The software used for the lab course is R-Studio which is installed on the *AudioLabs*-Webserver and can be accessed from the web browser. Each group will have a separate workspace that they can use to assess the exercises. The exercises itself are directly stored in the main workspace folder and are named:

- `ex1_t.Rmd`

- `ex2_mushra.Rmd`

- `ex3_elem.Rmd` (**optional**)

Each of the exercises is a *R Markdown* file which is a mix between a plain text file formatted in Markdown syntax [4]. Within the text file you can place R code which is executed and the output is embedded at the same place. This even works for plots. In the end the exercise Rmd file should be a self contained document that includes both the lab course questions **and** your answers. The answers therefore should contain Text, R-Code and possible plots.

```{r}
plot(cars)
```

You can use the R-Console in the main R-Studio window to work with the data and the try out the different built-in functions for solving the exercises. Both the Rmd script and the console share the same global variable environment. But make sure that all variables are declared within the Rmd file[4] To render an HTML output you can use the following command on the console:

```
library(rmarkdown)
render("ex____.Rmd")
```

The R-Studio environment allows you to install additional packages from the CTAN repository.

```
install.packages("mypkg", dependencies = TRUE)
```

Additional (installed) libraries need to be loaded before they can be used:

```
library(mypkg)
```

---

[4]you can test this when you click on the **Knit HMTL** button in R-Studio. This will build the output without using the current Global Environment workspace.

What you should know after this lab

- Know the stastical grammar!

- What is significance, p-value, effect sizes?

- How to read, interpret and present the results of a statistical test?

Lab Experiment 1

R allows there text and code to coexist in one single "R Markdown"[a] file. We are using this ability to have file with both exercises and prototype code you can work on. A PDF version of the exercises can be found in the the Appendix. Please finish the two experiment documents `ex1_t.Rmd` and `ex2_mushra.Rmd` in the materials directory.

---

[a]`http://rmarkdown.rstudio.com`

Part II
Appendix

10   Introduction to R and R-Studio

Lab 1: Introduction to data

Some define Statistics as the field that focuses on turning information into knowledge. The first step in that process is to summarize and describe the raw information - the data. In this lab, you will gain insight into public health by generating simple graphical and numerical summaries of a data set collected by the Centers for Disease Control and Prevention (CDC). As this is a large data set, along the way you'll also learn the indispensable skills of data processing and subsetting.

Getting started

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. As its name implies, the BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet and weekly physical activity, their HIV/AIDS status, possible tobacco use, and even their level of healthcare coverage. The BRFSS Web site (*http://www.cdc.gov/brfss*) contains a complete description of the survey, including the research questions that motivate the study and many interesting results derived from the data.

We will focus on a random sample of 20,000 people from the BRFSS survey conducted in 2000. While there are over 200 variables in this data set, we will work with a small subset.

We begin by loading the data set of 20,000 observations into the R workspace. After launching RStudio, enter the following command.

```
source ( "http://www.openintro.org/stat/data/cdc.R" )
```

The data set `cdc` that shows up in your workspace is a *data matrix*, with each row representing a *case* and each column representing a *variable*. R calls this data format a *data frame*, which is a term that will be used throughout the labs.

To view the names of the variables, type the command

```
names ( cdc )
```

This returns the names `genhlth`, `exerany`, `hlthplan`, `smoke100`, `height`, `weight`, `wtdesire`, `age`, and `gender`. Each one of these variables corresponds to a question that was asked in the survey. For example, for `genhlth`, respondents were asked to evaluate their general health, responding either excellent, very good, good, fair or poor. The `exerany` variable indicates whether the respondent exercised in the past month (1) or did not (0). Likewise, `hlthplan` indicates whether the respondent had some form of health coverage (1) or did not (0). The `smoke100` variable indicates whether the respondent had smoked at least 100 cigarettes in her lifetime. The other variables record the respondent's `height` in inches, `weight` in pounds as well as their desired weight, `wtdesire`, `age` in years, and `gender`.

> **Exercise 1** How many cases are there in this data set? How many variables? For each variable, identify its data type (e.g. categorical, discrete).

We can have a look at the first few entries (rows) of our data with the command

```
head ( cdc )
```

and similarly we can look at the last few by typing

```
tail ( cdc )
```

You could also look at *all* of the data frame at once by typing its name into the console, but that might be unwise here. We know cdc has 20,000 rows, so viewing the entire data set would mean flooding your screen. It's better to take small peeks at the data with head, tail or the subsetting techniques that you'll learn in a moment.

Summaries and tables

The BRFSS questionnaire is a massive trove of information. A good first step in any analysis is to distill all of that information into a few summary statistics and graphics. As a simple example, the function summary returns a numerical summary: minimum, first quartile, median, mean, second quartile, and maximum. For weight this is

```
summary ( cdc $ weight )
```

R also functions like a very fancy calculator. If you wanted to compute the interquartile range for the respondents' weight, you would look at the output from the summary command above and then enter

```
190  -  140
```

R also has built-in functions to compute summary statistics one by one. For instance, to calculate the mean, median, and variance of weight, type

```
mean ( cdc $ weight )

var ( cdc $ weight )

median ( cdc $ weight )
```

While it makes sense to describe a quantitative variable like weight in terms of these statistics, what about categorical data? We would instead consider the sample frequency or relative frequency distribution. The function table does this for you by counting the number of times each kind of response was given. For example, to see the number of people who have smoked 100 cigarettes in their lifetime, type

```
table ( cdc $ smoke100 )
```

or instead look at the relative frequency distribution by typing

```
table ( cdc $ smoke100 ) / 20000
```

Notice how R automatically divides all entries in the table by 20,000 in the command above. This is similar to something we observed in the last lab; when we multiplied or divided a vector with a number, R applied that action across entries in the vectors. As we see above, this also works for tables. Next, we make a bar plot of the entries in the table by putting the table inside the barplot command.

```
barplot ( table ( cdc $ smoke100 ) )
```

Notice what we've done here! We've computed the table of `cdc$smoke100` and then immediately applied the graphical function, `barplot`. This is an important idea: R commands can be nested. You could also break this into two steps by typing the following:

```
smoke  <-  table ( cdc $ smoke100 )

barplot ( smoke )
```

Here, we've made a new object, a table, called `smoke` (the contents of which we can see by typing `smoke` into the console) and then used it in as the input for `barplot`. The special symbol `<-` performs an *assignment*, taking the output of one line of code and saving it into an object in your workspace. This is another important idea that we'll return to later.

> **Exercise 2** Create a numerical summary for `height` and `age`, and compute the interquartile range for each. Compute the relative frequency distribution for `gender` and `exerany`. How many males are in the sample? What proportion of the sample reports being in excellent health?

The `table` command can be used to tabulate any number of variables that you provide. For example, to examine which participants have smoked across each gender, we could use the following.

```
table ( cdc $ gender ,  cdc $ smoke100 )
```

Here, we see column labels of 0 and 1. Recall that 1 indicates a respondent has smoked at least 100 cigarettes. The rows refer to gender. To create a mosaic plot of this table, we would enter the following command.

```
mosaicplot ( table ( cdc $ gender ,  cdc $ smoke100 ) )
```

We could have accomplished this in two steps by saving the table in one line and applying `mosaicplot` in the next (see the table/barplot example above).

> **Exercise 3** What does the mosaic plot reveal about smoking habits and gender?

Interlude: How R thinks about data

We mentioned that R stores data in data frames, which you might think of as a type of spreadsheet. Each row is a different observation (a different respondent) and each column is a different variable (the first is `genhlth`, the second `exerany` and so on). We can see the size of the data frame next to the object name in the workspace or we can type

```
dim ( cdc )
```

which will return the number of rows and columns. Now, if we want to access a subset of the full data frame, we can use row-and-column notation. For example, to see the sixth variable of the 567th respondent, use the format

```
cdc [ 567 ,  6 ]
```

which means we want the element of our data set that is in the 567th row (meaning the 567th person or observation) and the 6th column (in this case, weight). We know that `weight` is the 6th variable because it is the 6th entry in the list of variable names

```
names ( cdc )
```

To see the weights for the first 10 respondents we can type

```
cdc [ 1 : 10 ,  6 ]
```

In this expression, we have asked just for rows in the range 1 through 10. R uses the ":" to create a range of values, so 1:10 expands to 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. You can see this by entering

```
1 : 10
```

Finally, if we want all of the data for the first 10 respondents, type

```
cdc [ 1 : 10 ,  ]
```

By leaving out an index or a range (we didn't type anything between the comma and the square bracket), we get all the columns. When starting out in R, this is a bit counterintuitive. As a rule, we omit the column number to see all columns in a data frame. Similarly, if we leave out an index or range for the rows, we would access all the observations, not just the 567th, or rows 1 through 10. Try the following to see the weights for all 20,000 respondents fly by on your screen

```
cdc [ ,  6 ]
```

Recall that column 6 represents respondents' weight, so the command above reported all of the weights in the data set. An alternative method to access the weight data is by referring to the name. Previously, we typed `names(cdc)` to see all the variables contained in the cdc data set. We can use any of the variable names to select items in our data set.

```
cdc $ weight
```

The dollar-sign tells R to look in data frame `cdc` for the column called `weight`. Since that's a single vector, we can subset it with just a single index inside square brackets. We see the weight for the 567th respondent by typing

```
cdc $ weight [ 567 ]
```

Similarly, for just the first 10 respondents

```
cdc $ weight [ 1 : 10 ]
```

The command above returns the same result as the `cdc[1:10,6]` command. Both row-and-column notation and dollar-sign notation are widely used, which one you choose to use depends on your personal preference.

A little more on subsetting

It's often useful to extract all individuals (cases) in a data set that have specific characteristics. We accomplish this through *conditioning* commands. First, consider expressions like

```
cdc $ gender  ==  "m"
```

or

```
cdc $ age  >  30
```

These commands produce a series of TRUE and FALSE values. There is one value for each respondent, where TRUE indicates that the person was male (via the first command) or older than 30 (second command).

Suppose we want to extract just the data for the men in the sample, or just for those over 30. We can use the R function `subset` to do that for us. For example, the command

```
mdata  <-  subset ( cdc ,  cdc $ gender  ==  "m" )
```

will create a new data set called `mdata` that contains only the men from the `cdc` data set. In addition to finding it in your workspace alongside its dimensions, you can take a peek at the first several rows as usual

```
head ( mdata )
```

This new data set contains all the same variables but just under half the rows. It is also possible to tell R to keep only specific variables, which is a topic we'll discuss in a future lab. For now, the important thing is that we can carve up the data based on values of one or more variables.

As an aside, you can use several of these conditions together with `&` and `|`. The `&` is read "and" so that

```
m_and_over30  <-  subset ( cdc ,  cdc $ gender  ==  "m"  &  cdc $ age  >  30 )
```

will give you the data for men over the age of 30. The `|` character is read "or" so that

```
m_or_over30  <-  subset ( cdc ,  cdc $ gender  ==  "m"  |  cdc $ age  >  30 )
```

will take people who are men or over the age of 30 (why that's an interesting group is hard to say, but right now the mechanics of this are the important thing). In principle, you may use as many "and" and "or" clauses as you like when forming a subset.

> **Exercise 4** Create a new object called `under23_and_smoke` that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime. Write the command you used to create the new object as the answer to this exercise.

Quantitative data

With our subsetting tools in hand, we'll now return to the task of the day: making basic summaries of the BRFSS questionnaire. We've already looked at categorical data such as `smoke` and `gender` so now let's turn our attention to quantitative data. Two common ways to visualize quantitative data are with box plots and histograms. We can construct a box plot for a single variable with the following command.

```
boxplot ( cdc $ height )
```

You can compare the locations of the components of the box by examining the summary statistics.

```
summary ( cdc $ height )
```

Confirm that the median and upper and lower quartiles reported in the numerical summary match those in the graph. The purpose of a boxplot is to provide a thumbnail sketch of a variable for the purpose of comparing across several categories. So we can, for example, compare the heights of men and women with

```
boxplot ( cdc $ height  ~  cdc $ gender )
```

The notation here is new. The ~ character can be read "versus" or "as a function of". So we're asking R to give us a box plots of heights where the groups are defined by gender.

Next let's consider a new variable that doesn't show up directly in this data set: Body Mass Index (BMI). BMI is a weight to height ratio and can be calculated as.

$$BMI = \frac{weight\ (lb)}{height\ (in)^2} * 703^{\dagger}$$

The following two lines first make a new object called `bmi` and then creates box plots of these values, defining groups by the variable `cdc$genhlth`.

```
bmi  <-  ( cdc $ weight / cdc $ height ^ 2 )  *  703

boxplot ( bmi  ~  cdc $ genhlth )
```

---

†703 is the approximate conversion factor to change units from metric (meters and kilograms) to imperial (inches and pounds)

Notice that the first line above is just some arithmetic, but it's applied to all 20,000 numbers in the `cdc` data set. That is, for each of the 20,000 participants, we take their weight, divide by their height-squared and then multiply by 703. The result is 20,000 BMI values, one for each respondent. This is one reason why we like R: it lets us perform computations like this using very simple expressions.

> **Exercise 5** What does this box plot show? Pick another categorical variable from the data set and see how it relates to BMI. List the variable you chose, why you might think it would have a relationship to BMI, and indicate what the figure seems to suggest.

Finally, let's make some histograms. We can look at the histogram for the age of our respondents with the command

```
hist ( cdc $ age )
```

Histograms are generally a very good way to see the shape of a single distribution, but that shape can change depending on how the data is split between the different bins. You can control the number of bins by adding an argument to the command. In the next two lines, we first make a default histogram of `bmi` and then one with 50 breaks.

```
hist ( bmi )

hist ( bmi ,  breaks  =  50 )
```

Note that you can flip between plots that you've created by clicking the forward and backward arrows in the lower right region of RStudio, just above the plots. How do these two histograms compare?

At this point, we've done a good first pass at analyzing the information in the BRFSS questionnaire. We've found an interesting association between smoking and gender, and we can say something about the relationship between people's assessment of their general health and their own BMI. We've also picked up essential computing tools – summary statistics, subsetting, and plots – that will serve us well throughout this course.

References

[1] OpenIntro. Openintro statistics: Second edition. [Online]. Available: http://www.openintro.org

[2] International Telecommunications Union, "ITU-R BS.1534 (Method for the subjective assessment of intermediate quality levels of coding systems)," July 2014.

[3] ——, "ITU-R BS.1534-0 (Method for the subjective assessment of intermediate quality levels of coding systems)," June 2001.

[4] Wikipedia. Markdown. [Online]. Available: http://en.wikipedia.org/wiki/Markdown