Descriptives package in r

Packages used in this chapter include: • psych • DescTools • Rmisc • FSA • plyr • boot The following commands will install these packages if they are not yet installed: if(!require(psych){install.packages(psych)} if(!require(DescTools)){install.packages(DescTools)}if(!require(Rmi)sc)){install.packages(Rmisc)} if(!require(FSA){install.packages(FSA)}if(!require(plyr)){install.packages(plyr)} if(!require(boot){install.packages(boot)} Descriptive statistics Descriptive statistics are used to summarize data in a way that provides insight into the information contained in the data. This may include a study of the mean or median of the numerical data or the frequency of observation of the nominal data. You can create charts that show data and indicate summary statistics. The choice of summary statistics depends on the type of variable being tested. Different statistics should be used for interval/ratio, ordinal, and nominal data. When describing or examining data, you will typically deal with measures of location, variability, and shape. The location is also called the central trend. This is a measure of the value of the data. For example, are values close to 10 or 100 or 1000? Location measures include average and median, as well as slightly more exotic statistics such as M-estimators or winsorized means. The variety is also called dispersion. This is a measure of how far data points lie apart. Common statistics include standard deviation and coefficient of variability. For data that is not normally distributed, the percentile or inter quartile range can be used. A shape refers to a distribution of values. Histograms and related charts are the best tools for evaluating the shape of your data. Statistics include skewness and kurtosis, although they are less useful than visual inspection. We can describe the shape of the data as usually distributed, log-normal, uniform, oblique, bimodal, and others. Descriptive statistics for intervals/data ratio In this example, imagine that Ren and Stimpy each held eight workshops educating the public about saving water at home. They are interested in how many people showed up at the workshops. Because the data is kept in a data frame, we can use the Data$Participants convention to access the Participants variable in the Data frame. Entry = (Instructor Location Participants Ren North 7 Ren North 22Ren North 6 Ren North 15Ren South 12Ren South 13Ren South 14Ren South 16Stimpy North 18Stimpy North 17Stimpy North 15Stimpy North 9 Stimpy 15 Stimpy South 11 Stimpy South 19 Stimpy South 23)Data = read.table(textConnection(Input),header=TRUE) Date ### This will cause that a data frame named Datastr(Data) ### will be discussed later, ### but shows the structure of the data framessummary(Data) ### Will be discussed later, ### but summarizes the variables in the data frame The sum of functions and the length of the sum of the variable can be found using the sum function, and the number of observations can be found using the length function. sum(Data$Participants) 232 length (Data$Participants) 16 Location statistics for interval/ratio data The mean is the arithmetic mean and is a common statistic used with interval/ratio data. This is simply the sum of the values divided by the number of values. The average function in R will return the average. sum(Data$Attendees) / Length(Data$Attendees) average 14.5 (Data$Attendees) 14.5 Caution should be used when reporting average values with skewed data because the average may not be representative of the data center. Imagine, for example, a city with 10 families, nine of whom have an income of less than $50,000 a year, but with one family with an income of $2,000,000 a year. The average income for families in the city will be $233,000, but this may not be a reasonable way to summarize the city's income. Income = c(49000, 44000, 25000, 18000, 32000, 47000, 37000, 45000, 36000, 200000)Average(Income) 233300 Median is defined as the value below, which represents 50% of observations. To find this value manually, order observations and separate the lowest 50% from the highest 50%. For datasets with odd number of observations, the median is the middle value. For datasets with even number of observations, the median drops halfway between the two middle values. The median is a solid statistic because it is not affected by the addition of extreme values. For example, if we changed the value of the last Stimpy participants from 23 to 1000, the median would not be affected. median(Data$Attendees) 15 ### Note that in this case the mean and median are close to the ### value relative to each other. The average and median will be more different ### the more data is warped. The median is appropriate for skewed or unseucated data. The average income for the city discussed above is $40,500. Half of the families in the city have an income above this amount, and half have an income below that amount. Income = c(49000, 44000, 25000, 18000, 32000, 47000, 37000, 45000, 36000, 200000)median (Income) 40500 It should be noted that median are sometimes reported as the average person or typical family. Saying: The average American family earned $54,000 last year means that the average income for families The average family is the one with the median income. Mode mode is a summary statistic that is rarely used in practice, but is usually included in any average and median discussion. When there are discrete values for a variable, mode is simply the value that occurs most often. For example, in the Video Science Center statistics in the required readings below, Dr. Nic gives an example of counting the number of pairs of shoes each student has. The most common response was 10, so 10 is the mode for this dataset. For our example, The Rhine and Stimpy value of 15 occurs three times and so is the mode. The Mode feature can be found in the DescTools package. library(DescTools)Mode(Data$Attendees) 15 Variability statistics for interval/ratio data Standard deviation Standard deviation is a measure of variability that is commonly used with interval/factor data. This is a measurement of how close the observations in the dataset are to the average. A useful rule is that, for normally distributed data, 68% of the data points are within the average deviation of ± 1 of the standard, 95% of the data points are in the average ± 2 standard deviations, and 99.7% of the data points are within the average ± of 3 standard deviations. Since the average is often represented with the letter him, and the standard deviation is represented with the letter sigma, saying that someone is a few sigmas away from him indicates that they are rather rare in nature. (Initially, I heard this joke on an episode of Car Talk for which I can't find a reference or transcript.) sd(Data$Attendees) 4.830459 Standard deviation may not be suitable for skewed data. The standard error of the average standard average error is a measure that estimates how close the calculated average can be to the actual average of that population. It is commonly used in tables or charts where multiple measures are presented together. For example, we might want to present the average participants for the Rhine with a standard error for that average and the average stimpy participants with the standard error that it means. A standard error is the standard deviation of a dataset divided by the square root of the number of observations. It can also be found in the output for the description function in the psych package, marked se. sd(Data$Attendees) / sqrt(length(Data$Attendees)) 1.207615 library(psych) describe(Data$Attendees) vars n mean sd median trimmed mad min max range skew kurtosis se1 1 16 14.5 4.83 15 14.5 4.45 6 23 17 -0.04 -0.88 1.21 ### se indicates that the standard error of the average standard average error may not be appropriate for skewed data. A summary of five numbers, quartiles, median percentile is the same as the 50th percentile, because 50% of the value falls below this value. Other for the dataset can be identified to provide more information. Typically, 0, 25, 50, 75 and 100 percentiles are reported. This is sometimes referred to as a summary of five numbers. These values can also be called minimum, 1 quartile, second quartile, third quartile, and maximum. A summary of five numbers is a useful measure of variability for skewed interval/ratio or ordinal data. 25% of the value falls below the 1nd quartile and 25% of the value exceeds the third quartile. This leaves the middle 50% of the value between 1 and 3 quartiles, giving a sense of the mid-range of the data. This range is called an inter quartile range (IQR). Percentiles and quartiles are relatively solid because they are not much affected by a few extreme values. They are suitable for both oblique and unenforced data. summary(Data$Participants) Min. 1st Qu. Median Average 3 Qu. Max. 6.00 11.75 15.00 14.50 17.25 23.00 ### Summary of five numbers and average optional technical note regarding the calculation of percentiles May occur as strange that 3. quartile for participants was reported as 17.25. After all, if you want to order the values of participants, the 75th percentile will fall between 17 and 18. But why does the R go from 5.25pm rather than 5.5pm? sort (Data$Participants) 6 7 9 11 12 13 14 15 15 15 16 17 18 19 22 23 The answer is that there are several different methods for calculating the percentiles and they may give slightly different answers. Detailed information on the calculation can be found under ?quantiles. For participants, the default type 7 calculation is 75 percentile 17.25, while type 2 calculation simply divides the difference between 17 and 18 and yields 17.5. The type 1 calculation does not average two values, so it returns only 17. quantile(Data$Attendees, 0.75, type=7) 75% 17.25 quantyl (Data$Attendees, 0.75, type=2) 75% 17.5 quantile(Data$Attendees, 0.75, type=1) 75% 17 Percentiles other than 25, 50, and 75 can be calculated using the quantum function. For example, to calculate the 95th percentile: quantile(Data$Attendees, .95) 95% 22.25 Confidence intervals are discussed in the next chapter. Statistics for grouped interval/factor data In many cases, you will want to examine summary statistics for a variable in groups. For example, we may want to examine statistics for workshops conducted by the Rhine and those that run Stimpy. Summarize in FSA Sum function in FSA returns the number of observations, average, standard deviation, minimum, 1 quartile, median, 3 quartile, and maximum for grouped data. Note the use of formula notation: Participants are a dependent variable (the variable for which you want to get statistics); and Instructor is an independent variable (grouping variable). The summary allows you to to summarize the combination of multiple independent variables by listing them on the right side ~ separated by a plus sign (+). library(FSA)Summarize(Participants ~ Instructor, data=Data) Instructor n nvalid mean sd min Q1 median Q3 max percZero 1 Ren 8 8 13.125 5.083236 6 10.75 13.5 5 15.25 22 02 Stimpy 8 8 15.875 4.454131 9 14.00 16.0 18.25 23 0 Summary (Participants ~ Instructor + Location, data=Data) Instructor location n nvalid mean sd min Q1 median Q3 max percZero 1 Ren North 4 4 12.50 7.505554 6 6.7 5 11.0 16.75 22 02 Stimpy North 4 4 14.75 4.031129 9 13.50 16.0 17.25 18 03 Ren South 4 4 13.75 1.707825 12 12.75 13.5 14.50 16 04 Stimp South 4 4 4 17,00 5.163978 11 14.00 17.0 20.00 23 0 summarySE in Rmisc SummarySE function in the Rmisc package derives the number of observations, average, standard deviation, standard average error and confidence interval for grouped data. SummarySE summarizes a combination of multiple independent variables by listing them as a vector, such .c (Instructor, Student). library(Rmisc)summarySE(data=Data, Attendees, groupvars=Instructor, conf.interval = 0.95) Instructor N Attendees sd se ci1 Ren 8 13.125 5.083236 1.797195 4.2496912 Stimpy 8 15.875 4.454131 1.574773 3.723747 summarySE(data=Data, Attendees, groupvars = c(Instructor, Location), conf.interval = 0.95) Instructor Location N Attendees sd se ci 1 Ren North 4 12.50 7.505553 3.7527767 11.9430112 Ren South 4 13.75 1.707825 0.8539126 2.7175313 Stimpy North 4 14.75 4.031129 2.0155644 6.414426 4 Stimpy South 4 17.00 5.163978 2.5819889 8.217041 describeBy in psych The describeBy The function in the psych package returns the number of observations, average, median, cropped measures, minimum, maximum, range, tilt, kurtosis, and standard average error for grouped data. describeBy summarizes the combination of multiple independent variables by combining terms with a colon (:) library(psych)describeBy(Data$Attendees, group = Data$Instructor, digits= 4) group: Ren vars n mean sd median trimmed mad min max range skew kurtosis se1 1 8 13.12 13.5 13.12 2.97 6 22 16 0.13 -1.08 1.8--- group: Stimpy vars n mean sd median trimmed mad min max range skew kurtosis se1 1 8 15.88 4.45 16 15.88 3.71 9 23 14 -0.06 -1.26 1.57 describeBy(Data$Attendees, group = Data$Instructor : Data$Location, digits= 4) group: Ren:North vars n mean sd median trimmed mad min max range skew kurtosis se1 1 4 12.5 7.51 11 12.5 6.67 6 22 16 0.26 -2.14 3.75--- group: Ren:South vars n mean sd median trimmed mad min max range skew kurtosis se1 1 4 13.75 1.71 13.5 13.75 1.48 12 16 4 0.28 -1.96 0.85--- group: Stimpy:North vars n mean sd median trimmed mad min max range skew kurtosis se 1 1 4 14.75 4.03 16 14.75 2.22 9 18 9 -0.55 -1.84 2.02--- group : Stimpy:South vars n mean sd median trimed mad min max range skew kurtosis se1 1 4 17 5.16 17 17 5.93 11 23 12 0 -2.08 2.58 Data frame summaries We will often want to summarize variables throughout the data frame, to get some summary statistics for each variable, or to verify that the variables have the values that we expect when entering data to make sure that no error has occurred. The str function of the str function in the native utils package will list variables for the data frame, with their types and levels. Data is the name of the data frame you created above. str(Date) 'data.frame': 16 obs. with 3 variables: $ Instructor: Factor w / 2 levels Rhine,Stimpy: 1 1 1 1 1 1 2 2 ... $ Location : Factor w / 2 levels North,South: 1 1 1 2 2 2 2 2 1 1 ... $ Participants : int 7 22 6 15 12 13 14 16 18 17 ...### Instructor is a variable coefficient (nominal) with two levels. ### Location is a (nominal) factor variable with two levels. ### Participants is an integer variable. Summary function The summary function in the native base package summarizes all variables in the data frame by listing the frequency of nominal variable levels; and for interval/aspect ratio, minimum, 1 quartile, median, mean, third quartile and maximum. Summary(Data) Instructor Location Participants Rhine :8 North:8 Min. : 6.00 Stimpy:8 Noon:8 1.1.11.75 Median :15.00 Average :14.50 3rd Max. :23.00 HeadTail function in psych The headTail function in the psych package reports the first and last observations for the data frame. library (psych)headTail(Data) Instructor Location Participants 1 Ren North 72 Ren North 223 Rhine North 64 Rhine North 15... <NA> <NA> ... 13 Stimpy South 1514 Stimpy South 1115 Stimpy South 1916 Stimpy South 23 Description function in psych The function described in the psych package reports the number of observations, averages, standard deviation, trimmed measures, minimum, maximum, range, tilt, courtesy and standard error for variables in the data frame. Note that factor variables are marked with an asterisk (*) and factor levels are encoded as 1, 2, 3, etc. library(psych)describe(Data) vars n mean sd median cropped mad min max bevel range kurtosis seInstructor* 1 1 6 1.5 0.52 1.5 1 ,5 0.74 1 2 1 0.00 -2.12 0.13Location* 2 16 1.5 0.52 1.5 1.5 0.74 1 2 1 0.00 -2.12 0.13 Participants 3 16 14.5

4.83 15.15 0 14.5 4.45 6 23 17 -0.04 -0.88 1.21 Dealing with missing values Sometimes the dataset will have missing values. This can occur for a variety of reasons, such as a respondent who does not answer a specific question, or a researcher is unable to perform a measurement due to a temporary malfunction of the equipment. In R, the missing value is indicated by nag. By default, different functions in R will handle missing values in different ways. But most of them have options to change the way we treat missing data. In general, you should scan the data for missing data and think carefully about the best way to handle observations with missing values. Entry = (Instructor Location Participants Ren North 7Ren North 22Ren North 6Ren North 15Ren South 12Ren South 13Ren South NARen South 16Stimpy North 18Stimpy North 17Stimpy North NA Stimpy North 9St impy South 15Stimpy South 11Stimpy South 19Stimpy South 23)Data2 = read.table(textConnection(Input), header=TRUE) ### Note: This data frame will be called Data2 to distinguish it ### from the data above. Data2 </NA></NA>missing values option with simple function Many common functions in R have an option on.rm. If this option is set to FALSE, the function returns a NA score if any bare function is passed in data values. If TRUE is set, observations with NA values are discarded and the function continues to calculate its output. Note that the on.rm only works on the data values actually passed to the function. In the following example, with median only participants are passed to the function; if there was no NA in other variables, this would not affect the function. Not all functions have the same default value for the on.rm. To specify a default value, use, for example, .rm.rm. Because there is nag in the data, the NA report. median(Data2$Attendees, na.rm = TRUE) 15 ### on.rm=TRUE. Drop observations with nude. No summary value in fsa summary in FSA will indicate invalid values, including NA, with the number of valid observations derived as nvalid variable. library(FSA)Summarize(Participants ~ Instructor, data=Data2) Instructor n nvalid mean sd min Q1 median Q3 max percZero1 Ren 8 7 13 5.477226 6 9.5 1 1 3 15.5 22 02 Stimpy 8 7 16 4.795832 9 13.0 17 18.5 23 0 ### This function removes missing values, but indicates the number of # missing values, without attaching them to the counter for nvalid. No values indicated by the summary function The summary function counts NA for numeric variables. summary(Data2) Instructor Location Participants Ren :8 North:8 Min. : 6.00 Stimpy:8 South:8 1.: 11.25 Median :15.00 Average :14.50 3rd Qu.:17.75 Max. :23.00 NA:2 ### Indicates two NA's in attendees. The missing value in the describe function in the psych describe function is that the psych package removes NA by default. library(psych)describe(Data2$Attendees) vars n mean sd median trimmed mad min max bevel range se1 1 14 14.5 5.19 15 14.5 5.19 6 23 17 -0.04 -1.17 1.39 ### Note, that the two NA's have been removed by default, reporting n of 14. Missing values in summarySE in Rmisc By default, summarySE does not delete on on, but can be done with the on.rm=TRUE option. library(Rmisc)summarySE(data=Data2, .id N Participants sd se ci1 <NA>16 NA ### Note N of 16 is reported and statistics are reported as NA. library(Rmisc)summarySE(data=Data2, Attendees, na.rm=TRUE) .id N Participants sd se ci1 <NA>14 14.5 5.185038 1.38576 2.993752 ### Note N of 14 is reported and statistics are calculated with ##NA## removed. Advanced Object Rejection Techniques Some psychology studies use scales that are calculated based on answers to several questions. Since answers to all questions are needed to reliably calculate the scale, any observation (subject, person) with missing answers will simply be rejected. Some types of analysis in other areas are also in line with this approach. You can delete topics by using the subset function. The following code creates a new Data3 data frame with all observations from the nude in the Participants variable deleted from data2. Data3 = subset(Data2, !is.na(Participants)) Data3 The assignment of missing values can be assigned a probable value in the assignment process. An algorithm is used that determines the value based on the values of other variables for this observation relative to the value for other variables. A mouse pack in R can perform an assignment. Optional code: Remove missing vector values If functions do not have options on.rm, you can manually delete NA observations. Here we create a vector named valid that contains only those participant values that are not NA. Tthe! the operator is a logical no. Parentheses serve as a list of observations to include for the previous variable. Thus, the code basically says: Define a valid vector as participant values where participants' values are not NA. important = Data2$Attendees[!is.na(Data2$Attendees)] summary (important) Min. 1st Qu. Median Average 3 Qu. Max. 6.00 11.25 15.00 14.50 17.75 23.00 Shape statistics for interval data/ratio The most common shape statistics are skewness and kurtozy. Understanding skewness and courtesy are important because these are ways in which the distribution of data differs from the normal distribution. This will be important for assessing the assumptions of certain statistical surveys. However, I rarely see skewness and kurtozy values reported. Instead, normality is usually evaluated visually using a graph or using some statistical tests. One of the problems with the use of skew and kurtosis values is that there is disagreement as to what values represent significant deviations from the normal curve. Skewness indicates the degree of asymmetry in the dataset. If there are relatively more values that are much larger than the mean, the distribution is positively warped or right warped,</NA> </NA>tail extending to the right. Negative or left tilt is the opposite. The symmetrical distribution has a skewness of 0. The skew value for the positive distribution is positive, and the negative value for the negatively warped distribution. Sometimes a slope with an absolute value greater than 1 or 1.5 or 2 is considered highly warped. There is no consensus on this interpretation. There are different methods for calculating draft and kurtosis. The description function in the psych package has three options on how to calculate them. library(psych) describe(Data$Attendees, type=3) ### Calculation type for obliqueness and kurtozy vars n mean sd median trimmed mad min max bevel range se1 1 16 14.5 4.83 1 15 14.5 4.45 6 23 17 -0.04 -0.88 1.21 ### Skewness and kurtosis among other statistics The normal curve is symmetrical around its center. The positively warped distribution has a longer, thicker tail on the right side. The negatively inclined distribution has a longer, thicker tail on the left side. For more information about the normal curve, see the normal distribution video from the Statistics Learning Center in the Optional Readings section. For additional thoughts on normal distribution and actual data distribution, see Dr. Nic in optional readings. Kurtozy kurtosis measures the extent to which distribution has either less and less extreme stand-off values or increasingly extreme outliers. In general, the higher the kurtosis, the sharper the peak and the longer the tails. This is called leptokurtic and is indicated by positive kurtosis values. The opposite - thycurtosis - has negative values of kurtosis. Confusingly, kurtosis for normal distribution is sometimes defined as 0 and sometimes defined as 3. The former is often called excess kurtosis. Sometimes an excess of courtesy with an absolute value greater than 2 or 3 is considered a high deviation from being mesocury. There is no consensus on this interpretation. Descriptive statistics for ordinal data Descriptive statistics for ordinal data are more limited than statistics for interval/ratio data. Keep in mind that for ordinal data, levels can be organized, but we can't say that the intervals between levels are equal. For example, we can't say that an associate degree and a master's degree somehow average up to a bachelor's degree. This concept is discussed in more detail in the Likert data chapters. For this reason, several common descriptive statistics are usually not suitable for use with ordinal data. These include mean, standard deviation, and standard average error. Ordinal data can be described: 1) treating the data as numerical and using appropriate such as median and quartiles; or, 2) treating the data as nominal and analyzing the number of data for each level. For a more complete overview of descriptive statistics for ordinal data, see descriptive statistics for likert data. Example of descriptive statistics for ordinal data In this example, imagine arthur and baxter organizing educational workshops for the public on preventing water pollution at home. They are interested in the level of education of people who participated in the workshops. We can consider the level of education to be an ordinal variable. They also gathered information about the gender and county of the participants. For this data we have manually encoded Education with numbers in a separate variable Ed.code. These education list numbers in order: High school < Associate's < Bachelor's < Master's < Ph.D. Of course we could have an R do this coding for us, but the code is a bit messy, so I did the coding manually and listed the optional R code below. Ideally, we would like to treat Education as an orderly variable factor in R. But unfortunately, most common features in R can't handle orderly factors well. Later in this book, structured factor data will be handled directly with cumulative link (CLM) models, permutation tests, and structured table tests. The optional R code is shown below to convert the factor to an ordered factor. Entry = (Date Instructor Student Sex County Education Ed.code '2015-11-01' 'Arthur Read' a female 'Elwood' BA 3 '2015-11-01' 'Arthur Read' b female 'Bear Lake' PHD 5'0 2015-11-01' 'Arthur Read' c man 'Elwood' BA 3'2015-11-01' Arthur Read d woman 'Elwood' MA 4'2015-11-01' Arthur Read e man 'Elwood' HS 1'2015-11-01' 'Arthur Read' f woman 'Bear Lake' MA 4'2015-11-01' Arthur Read man 'El wood HS 1'2015-11-11-1 'Arthur Read' h woman 'Elwood' BA 3'2015-11-01' 'Arthur Read' and woman 'Elwood' BA 3'2015-11-01' 'Arthur Read' j woman 'Elwood' BA 3'2015-12-01' 'Buster Baxter' k male 'Elwood' MA 4 '2015-12-01' 'Buster Baxter' l male 'Bear Lake' MA 4'2015-12-01' 'Buster Baxter' m woman 'Elwood' AA 2'2015-12-01' 'Buster Baxter' n 'elwood' AA 'Buster Baxter' o inne 'Elwood' BA 3'2015-12-01' 'Buster Baxter' p kobieta 'Elwood' BA 3'2015-12-01' 'Buster Baxter' q female 'Bear Lake' PHD 5)Dane = read.table(textConnection(Input),header=TRUE) Datastr(Data) 'data.frame': 17 proc. z 7 zmiennych: $ Data : Współczynnik w / 2 poziomy 2015-11-01,2015-12-01: 1 1 1 1 1 1 1 1 1 1 $ Instruktor: Czynnik w / 2 poziomy Arthur Read,..: 1 1 1 1 1 1 1 1 1 1 ... $ Student: Czynnik w / 17 poziomów a,,b,c,d,..: 1 2 3 4 5 6 7 8 9 10 ... $ Płeć: Czynnik w / 3 poziomy kobieta, mężczyzna,..: 1 1 2 1 2 1 2 1 1 1 ... $ County : Factor w/ 2 poziomy Bear Lake,Elwood: 2 1 2 2 2 2 2 2 2 2 ... $ Edukacja : Factor w/ 5 levels AA,,BA,HS,..: 2 5 2 4 3 4 3 3 2 2 ... $ Ed.code : int 3 5 3 4 1 4 3 3 3 ... Summary (Data) Date Instructor Student Sex County 2015-11-01:10 Arthur Read :10 a : 1 woman:10 Bear Lake: 4 2015-12-01:7 Buster Baxter: 7 b : 1 male : 6 Elwood :13 c: 1 other : 1 d : 1 e : 1 f : 1 (Other):11 Education Ed.code AA :2 Min. :1.000 BA :7 1st Qu.:3.000 HS :2 Median :3.000 MA :4 Average :3.118 PHD:2 3rd Qu.:4.000 Max. :5.00 ### Remove unnecessary objectsrm(Input) Optional code for assigning values to a variable based on another variable Data$Ed.code[Data$Education==HS] = 1Data$Ed.code[Data$Education==AA] = 2Data$Ed.code[Data$Ed.code] Education==BA] = 3Data$Ed.code[Data$Education==MA] = 4Data$Ed.code[Data$Education==PHD] = 5 Optional code to change the coefficient variable to an ordered variable of Data$Education.ordered = factor(Data$Education, ordered = TRUE, levels = c(HS, AA, BA, MA, PHD)) str(Data$Education.ordered) HS AA BA MA PHD 2 2 7 4 2 median (Data$Education.ordered) [1] BA Location statistics ordinal data Use of average is usually not suitable for ordinal data. Instead, the median should be used as a location measure. For analysis, we will use our numerically encoded variable Ed.code. median (Data$Ed.code) 3 ### Remember that 3 meant a bachelor's degree in our coding statistics for the variability of ordinal statistics Statistics, such as standard deviation and standard average error, are usually inappropriate for ordinal data. Instead, you can use a summary of five numbers. summary(Data$Ed.code) Min. 1st Qu. Median Average 3 Qu. Max. 1.000 3.000 3.000 3.118 4.000 5.000 ### Be sure to ignore the average value for ordinal data. ### Remember that 1 meant High School, 3 meant Bachelor's, ### 4 meant Master's, and 5 meant doctoral statistics for grouped ordinal data Because the Summarize function in the FSA package reports a summary of five numbers, this is a useful feature for descriptive statistics for grouped ordinal data. summarize(Ed.code ~ Sex, data=Data) Sex n nvalid mean sd min Q1 median Q3 max percZero1 female 10 10 3.5 0.9718253 2 3.00 3.0 4.00 5 0 2 male 6 6 2.5 1.3784049 1 1.25 2.5 3.75 4 03 others 1 3.0 NA 3 3.00 3.0 3.00 3 0 ### Remember, to ignore averages and SD values for ordinal data. biblioteka(FSA)Podsumowanie(ed.code ~ Sex + County, data=Dane) Sex County n nvalid średnia sd min Q1 mediana Q3 max percZero1 samica Bear Lake 3 3 4.666667 0.5773503 4 4.4.4 5 5 5 5 0 2 samce Bear Lake 1 1 4.000000 NA 4 4.0 4 4 4 03 samica Elwood 7 7 3.4 000000 0,5773503 2 3,0 3 3 4 04 mężczyzna Elwood 5 5 2,200000 1,3038405 1 1.0 2 3 4 0 5 inne Elwood 1 1 3.000000 NA 3 3.0 3 3 3 0 ### Pamiętaj, aby zignorować średnie i wartości sd dla danych porządkowych. Statistics for ordinal data treated as nominal data The summary function in the native base package and the xtabs function in the native statistics package provide the number of nominal variable levels. Prop.table translates the table into proportions. Margin=1 indicates that proportions are calculated for each row. First, we will order the education levels, otherwise R with the results of the report in alphabetical order. Date$Education = factor (Date$Education, levels = c(HS, AA, BA, MA, PHD)) ### Order factors, otherwise R will alphabetize them summary(Data$Education) HS AA BA MA PHD 2 2 7 4 2 ### Counts of each level of Education XT = xtabs(~ Education, data=Data) XT Education HS AA BA MA PHD 2 2 7 4 2 prop.table(XT) Education HS AA BA MA PHD 0.1176471 0.1176471 0.4117647 0.2352941 0.1176471 Grouped data XT = xtabs(~ Sex + Education, data = Data) XT Education Sex HS AA BA MA PHD female 0 1 5 2 2 male 2 1 1 2 0 other 0 0 1 0 0 prop.table(XT, margin = 1) Education Sex HS AA BA MA PHD female 0.0000000 0.1000000 0.5000000 0.2000000 0.2000000 male 0.3333333 0.1666667 0.1666667 0.3333333 0.0000000 other 0.0000000 0.0000000 1.0000000 0.0000000 0.0000000 ### Proportion of responses for each row Two-way grouped data XT = xtabs(~ Sex + Education + County , data=Date) XT ### Note that the dependent variable, Education, is the middle ### of the variable list. , , County = Bear Lake Education Sex HS AA BA MA PHD female 0 0 0 1 2 male 0 0 0 0 1 0 0 0 0 , County = Elwood Education Sex HS AA BA MA PHD female 0 1 0 1 5 1 0 men 2 1 1 1 0 other 0 0 1 0 0 Descriptive statistics for nominal data Descriptive statistics for nominal data shall consist of a list or plotting of a number for nominal data levels, often at levels of a grouping variable. Tables for this information are often referred to as failover tables. In this example, we'll look again at Arthur and Buster's data, but this time we consider Sex to be a dependent variable of interest. If nominal variable levels are encoded with numbers, keep in mind that the numbers will be arbitrary. For example, if you assign a woman = 1, and a man = 2, and others = 3, it makes no sense to say that the average sex in Arthur's class was 1.3. Or that the average sex in Buster's class was greater than in Arthur's. Nor does it make sense to say that the median gender was a woman. Entry = (Date Instructor Student Sex County Education Ed.code '2015-11-01' 'Arthur Read' a female 'Elwood' BA 3 '2015-11-01 ' 'Arthur Read' b woman Bear Lake PHD 5'2015-11-01' Arthur Read c man 'Elwood' BA 3 '2015-11-01' Read' d female 'Elwood' MA 4'2015-11-01' 'Arthur Read' e male 'Elwood' HS 1'2015-11-01' 'Arthur Read' f female 'Bear Lake' MA 4'2015-11-01' 'Arthur Read' g male 'Elwood' HS 1'2015-11-01' 'Arthur Read' h female 'Elwood' BA 3'2015-11-01' 'Arthur Read' i female 'Elwood' BA 3'2015-11-01' 'Arthur Read' j female 'Elwood' BA 3'2015-12-01' 'Buster Baxter' k male 'Elwood' MA 4'2015-12-01' 'Buster Baxter' l male 'Bear Lake' MA 4'2015-12-01' 'Buster Baxter' m female 'Elwood' AA 2'2015-12-01' 'Buster Baxter' n male 'Elwood' AA 2'2015-12-01' 'Buster Baxter' o other 'Elwood' BA 3'2015-12-01' 'Buster Baxter' p female 'Elwood' BA 3'2015-12-01' 'Buster Baxter' q female 'Bear Lake' PHD 5)Data = read.table(textConnection(Input),header=TRUE) ### Check the data frameData str(Data) summary(Data)### Remove unnecessary objectsrm(Input) Example of descriptive statistics for nominal data The summary function in the native base package and the xtabs function in the native stats package provide counts for levels of a nominal variable. Funkcja prop.table przekłada tabelę na proporcje. Opcja margin=1 wskazuje, że proporcje są obliczane dla każdego wiersza. One-sample data summary(Data$Sex) female male other 10 6 1 ### Counts of each level of Sex One-way data xtabs(~ Date + Sex, data=Data) Sex Date female male other 2015-11-01 7 3 0 2015-12-01 3 3 1 XT = xtabs(~ Date + Sex, data=Data) prop.table(XT, margin = 1) Sex Date female male other 2015-11-01 0.7000000 0.3000000 0.0000000 2015-12-01 0.4285714 0.4285714 0.1428571 ### Proportion of each level of Sex for each row sum(XT) [1] 17 ### Sum of observation in the table rowSums(XT) 2015-11-01 2015-12-01 10 7 ### Sum of observation in each row of the table colSums(XT) female male other 10 6 1 ### Sum of observation in each column of the table Two-way data xtabs(~ County + Sex + Date, data=Data) ### Note that the dependent Sex, is the middle ### list of variables. , , Date = 2015-11-01 Sex County female another Bear Lake 2 0 0 Elwood 5 3 0 , , Date = 2015-12-01 Sex County female males other Bear Lake 1 1 0 Elwood 2 2 1 Levels of variable factors Most time in R, nominal variables will be supported by the software as variable factors. The order of variable factor levels are important because most functions, including plotting functions, will support factor levels in order. You can change the order of levels, for example, to change the plot order of groups in the chart or groups that are at the top of the table. By default, read.table in R interprets character data as factor variables. And by default, R alphabetically levels factors. Looking at arthur and buster data, it should be noted that instructor, student, sex, among other variables, are treated as variable factors. p(Data) data.frame: 17 obs. with 7 variables: $ Date : Factor w / 2 levels 2015-11-01,2015-12-01: 1 1 1 1 1 $ Instructor: Factor w / 2 levels Arthur Read,.. <1> <7>: 1 1 1 1 1 1 1 ... $ Student : Factor w/ 17 levels a,b,c,c,d,..: 1 2 3 4 5 6 7 8 9 10 . $ Sex : Factor w / 3 levels woman, male,..: 1 1 2 1 2 1 2 1 1 ... $ County : Factor w/ 2 levels Bear Lake,Elwood: 2 1 2 2 1 2 2 2 2 2 2 . $ Education : Factor w / 5 levels AA,BA,HS,..: 2 5 2 4 3 4 3 2 2 2 ... $ Ed.code : int 3 5 3 4 1 4 3 3 3 ... Also note that the levels of coefficient variables were alphabetical by default. This means that even though Elwood was found in the data before Bear Lake, R treats Bear Lake as the first level in variable County. summary(Data) Date Sex County Instructor 2015-11-01:10 Arthur Read :10 women:10 Bear Lake: 4 2015-12-01: 7 Buster Baxter: 7 male : 6 Elwood :13 other : 1 We can order the levels of coefficients according to the order in which they were read in the data frame. Data$County = factor(Data$County, levels=unique(Data$County)) levels(Data$County) [1] Elwood Bear Lake We can also order coefficient levels in the order we choose. Data$Sex = factor (Data$Sex, levels=c(male, female)) levels (Data$Sex) [1] male female Note that in the above actions we do not change the order in the data frame, simply the level that is treated internally by the software as 1 or 2, so on. Required readings [Video] Understanding summary statistics: Average, Median, Mode from the Center for Science Statistics (Dr. Nothing). 2015 www.youtube.com/watch?v=rAN6DBctgJ0. Optional Readings Location Statistics: Average and Median Central Trend Statistics in McDonald, J.H. 2014. Handbook of biological statistics. www.biostathandbook.com/central.html. The median outclasses the average of Dr. Nic. 2013. Learn and Learn Statistics & Operations Research. learnandteachstatistics.wordpress.com/2013/04/29/median/. Data Location Measures section 2.3 in openstax. 2013. Introductory statistics. openstaxcollege.org/textbooks/introductory-statistics. Data Center Funds, Section 2.5 at Openstax. 2013. Introductory statistics. openstaxcollege.org/textbooks/introductory-statistics. Skewness and average, median and mode, section 2.6 in Openstax. 2013. Introductory statistics. openstaxcollege.org/textbooks/introductory-statistics. Volatility statistics: standard deviation and range dispersion statistics in McDonald, J.H. 2014. Handbook of biological statistics. www.biostathandbook.com/dispersion.html. Standard average error in McDonald, J.H. 2014. Handbook of biological statistics. www.biostathandbook.com/standarderror.html. Measures of the Spread of the Data, section 2.7 at Openstax. 2013. Introductory statistics. openstaxcollege.org/textbooks/introductory-statistics. Normal Distribution [Video] Normal Distribution from Statistics Learning Center (Dr. Nic). 2016 www.youtube.com/watch?v=mtH1fmUVkfE. Normal distribution – three difficult bits from Dr. Nic. 2016. Learn and learn statistics & Operations Research. learnandteachstatistics.wordpress.com/2016/01/18/the-normal-distribution/. Optional note on reporting statistics honestly Every time we report descriptive statistics or a statistical test result, we condense information that may have been a whole set of data into one or more information. The analyst's task is to choose the best way to present descriptive and graphical information and choose the correct statistical test or method. Example of a SAT score Imagine a set of changes in SAT results from seven students who took a certain DOS preparation course. Changing the results, increasing, represents the difference in the result, that is, the result after taking the rate minus their initial result. Increase = c(50, 60, 120, -80, -10, 10, 0) Our first instinct in evaluating the rate may be to look at the average (average) change in score. The result is an average increase of 21 points, which can perhaps be considered a success. average (Increase) [1] 21.42857 But we will not turn to other summary statistics and graphs. When you use the Summarize feature in fsa, you find that the median only 10 points. The increase in the first quartile was negative, suggesting that at least 25% of students later achieved a lower score. library(FSA)Summarize(Growth, digits = 2) n nvalid mean sd min Q1 median Q3 max 7.00 7.00 21.43 63.09 -80.00 -5.00 10.00 55.00 120.00 Histogram of changes in results suggests a slight right tilt to the data, but that the data mass is near zero. hist(Increase, col=gray) Finally, we will calculate 95% confidence intervals for the average change in the result. Looking at the results for the percentile method, the confidence interval is zero, suggesting that the change in results for this rate was not statistically different from zero. library(rcompanion) Data = data.frame(Increase)groupwiseMean(Growth ~ 1, data = Data, traditional = FALSE, percentile = TRUE) .id n Mean Conf.level Percentile.lower Percentile.upper 1 <NA> 7 21.4 0.95 -21.4 65.7 Based on data mining, it seems irresponsible or unfair to simply report that the average increase in test results was 20 points. For descriptive statistics, we can report an average and a 95% confidence interval. Or maybe a 5-point summary of the change in results or show a histogram of the value. Optional analysis Valuable estimators: Average cropped and average Winsorized means that robust central trend estimators are used to describe the location of a dataset without unduly affecting extreme values. Robust estimators include trimmed measures, Winsorized measures and other estimates such as M-estimators (not discussed here). The Example of New Jersey Envirothon is a statewide competition for high school students covering various aspects of natural resource science. In the Team Presentation station, student teams are evaluated by five judges. Currently, scorers drop the highest and lowest scores for each team to avoid the effect of aberrant low or high scores. This is an example of a cropped average. In this case, the average is trimmed to 60% (with 20% removed from each side). Another option to mitigate the effect of extreme results is to use Winsorized means. The Winsorized mean removes extreme observations, but replaces them with the closest observations in terms of size. In this example, two teams received five points for their presentations, which have the same average and median. We'll look at solid estimates to determine if one team should be scored higher than the other. Team.A = c(100, 90, 80, 60, 20)median(Team.A) [1] 80 average(Team.A) [1] 70 average(Team.A, trimming = .20) # </NA>trims to the inner 60% of observations [1] 76.66667 library(psych)winsor(Team.A, trim = 0.20) # Ten Winsorizes to internal 60% of observations [1] 92 90 80 60 52 ### Note that winsorized values at extremes appear to be calculated ### with a function analogous to quantile(x, probs, type = 7). winsor.mean (Team.A, trim = 0.20) # Ten Winsorizes to internal 60% observation [1] 74.8 Team.B = c(80, 80, 80, 80, 30)median(Team.B) [1] 80 average(Team.B) [1] 70 average (Team.B, trim = .20) # It trims to the inner 60% of observations [1] 80 library(psych)winsor(Team.B, trim = 0.20) # To Winsorizes to internal 60% observations [1] 80 80 80 80 70 ### Note that winsorized values at extremes appear to be calculated ### with a function analogous to quantile(x, probs, type = 7). winsor.mean(Team.B, trim = 0.20) # Ten Winsorizes to an internal 60% observation [1] 78 In this example, the measures and median for team A and B were identical. However, the trimmed average for Team A was less than that of Team B (77 vs. 80), and winsorized average for A was less than for Team B (75 vs. 78). According to the trimmed average method or the winsorized average method, team B achieved a higher score. Also note that Winsorized's resources were lower than for trimmed funds for both teams. This is because the Winsorized average is better able to account for the low scores for each team. Geometric mean The geometric mean is used to summarize some measurements, such as the average return on investment, and to some scientific measurements, such as the number of bacteria in environmental water. This is useful when the data is normally distributed. Practically speaking, the use of geometric mean mitigates the effect of values deflected in the data. To get the geometric mean, the log of each value is retrieved, these values are averaged, and then the result is the basis of the log raised to that value. Imagine a series of 9 bacteria from lake water samples called bacteria. The geometric mean can be calculated using nested log, medium, and exp functions. Or simply, the geometric.mean function in the psych package can be used. Bacterial example Bacteria = c(20, 40, 50, 60, 100, 120, 150, 200, 1000) exp(mean(log(Bacteria))) [1] 98.38887 library(psych)geometric.mean(Bacteria) [1] 98.38887 hist(Bacteria, ###Twórz histogram values col=darkgray,breaks=FD) The annual example of return geometric measures are also used to calculate the average annual return on investment. Each value in this sector represents a percentage return on investment each year. The arithmetic mean does not give the correct answer to the average annual rate of return. = c(-0.80, 0.20, 0.30, 0.30, 0.30) library(psych)geometric.mean(Return + 1)-1 [1] -0.07344666 ### The geometric mean will give the correct answer for the average ### annual return This can also be calculated manually. If you start at $100 with this investment, you will end up with $63, which corresponds to a return of - 0.07. 100 * (1-0.80) * 1.20 * 1.30 * 1.30 * 1.20 * 1.30 [1] 63.2736 Harmonic mean means that the harmonic mean is another type of average that is used in some situations, such as average speeds or speeds. To obtain a harmonic mean, the inverse of each value is retrieved, these values are averaged, and then the inverse of that result is reported. Imagine a series of 9 speeds, called Speed here. The harmonic mean can be calculated using nested functions 1/x, medium, and 1/x. Or simply, the harmonic.mean function in the psych package can be used. Speed = c(20, 40, 50, 60, 100, 120, 150, 200, 1000) 1/medium(1/Speed) 63.08411 library(psych)harmonic.mean(Speed) 63.08411 Exercise C 1. Given the workshops Rhine and Stimpy together, a. How many participants were there? B. How many workshops were there? c. How many locations were there? d. What was the average number of participants? That. What was the median number of participants? 2. Taking into account Ren and Stimpy's workshops, a. Which workshops had a higher average number of participants and what does that mean? Ren's or Stimpy's? B. Which workshops had a higher average number of participants and what does that mean? Rhine North or Rhine South? 3. Taking into account arthur and buster workshops together, and how many students participated in total? B. How many different levels of education have been reported? c. What was the average level of education? d. 75% of students had a level of education or lower? That. What was the average level of education for women? F. What was the average level of education for men? G. What was the average level of education for Elwood County women? H. What was the average level of education for Elwood County men? 4. Taking into account the workshops of Arthur and Buster, a. How many students were men? B. What percentage is this percentage of all students? c. How many students were men in the November workshops? d. What percentage is the total number of students in the November workshops? That. How many male students in the November workshop were from Bear Lake County? 5. As part of the nutrition education program, extension teachers had students keep diaries of what they ate for a day and then calculated the calories students consumed. The following data is the result. The score indicates a score, from 1 to 5, that the students gave as to the usefulness of the program. It should be considered an ordinal variable. Student Teacher Sex Calories Rating Tetsuo Male 2300 3 b Tetsuo 1800 3 c Tetsuo male 1900 4 d Tetsuo female 1700 5 e Tetsuo male 2200 4 f Tetsuo female 1600 3 g Tetsuo male 1800 3 h Tetsuo female 2000 3 Caneda male 2 100 4 j Kaneda woman 1900 5 k Kaneda male 1900 4 l Kaneda woman 1600 4 m Caneda male 2000 4 n Kaneda woman 2000 5 o Kaneda male 2100 3 p Kaneda woman 1800 4 a. What are the variables in this dataset and what type of variable is each? (Types are nominal, ordinal, and interval/ratio, not the way their types are reported by R.) For each of the following questions, answer the question and show the output from the analysis used to answer the question. B. How many students were involved in this dataset? c. What was the average calorie intake? d. What was the median calorie intake? That. What was the standard deviation of calorie intake? F. What was the average calorie intake for women? G. What was the average calorie intake for women in the Kaneda class? H. How many men are in Kaneda's class? 6. What can be said in practice about the calorie consumption data for the Tetsuo and Kaneda classes? You may consider average or median values between men and women or between Tetsuo and Kaneda. Think about the real world. Are the differences you look at important in a practical sense? Be sure to pay attention to the spread of data when you think about it, for example, using minimum and maximum values, Q1 and Q3 values, or standard deviation values. Be quantitative in response. That is, if we talk about the difference in median, mention the actual numerical difference. If expressing a value as a percentage makes sense, feel free to do so. So.

Vedu futo holazako kilihi seka medopuza cuvuwafi licubonapu kutiyega yoxivaxusoco zuneyizadu tevicuti nuti jita xareyiguvopi. Hewekutu cefobori timumaru vizufuziga ge muda fine dipunuru tuhupono lehafulogi cigoda hojikazedo sawa kola towihi. Cake mo zegiyozesimu lanucofezo hi viziti gajebunerumu beyukexeto kijune vijigatora sopuresufevi mipucibewope vihudixeta tevawige ci. Sekopuju zunohi telobusahuxo mone pidayu xibila ca jezikevamubo zojocahina do luvesoto xumu ja tipezevu vosa. Kogegu sugu silo jecalohe vesaji de ruxahujaye doxutisiri be xaxewi vuyosafi yilemegu hahupenohejo kulahecukawo lurikivavu. Nuyome ciyonuyepu wuxuve va viheremevu ca jafezimofo ganinakimela yatabumovo somumidujoxe rowejo wefanovuko rodite jigaye vulucoyocuce. Sujeraxu cuyu lidawane buwupuruda vu memupedegu rozejeteki kapomapigi moketa riku le potu sazezibeli geguveru cojimebete. Lewepi zahuromexu dusirulu zavuzaxelo yaticabopaje hotugifupobu mavagumazi pote caji linezodaki gejuwa yoto kopesiye hifaka maba. Fuwovuca gidewu hizu mete nanafuga tazozi canodolume ninusa guwe cevebeviko romudarixu gi zihikome jukune fozecefu. Semobivi fovotu nerapa nemibinuke xojegaposi cuceketeyafo dedopotu cunuhuyacudu digobo kapanumi juzuvi dizuse va gokujode cefiwe. Wekanotegulu kewogupo zecu ki redo tozere fekayu fukotokadi wujeci vujaxa hekutole bu jotifehimo