Homework 2: 45pts

Assigned: September 6, 2021

Due: September 13, 2021, 8:30am

1. (12pts) Below are 4 datasets with exposure (or “predictor”, or “covariate”…) , and outcome (or “response”) . Spoiler alert! This is a well-known phenomenon called Anscombe’s quartet.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | y |  | x | y |  | x | y |  | x | y |
| 10.0 | 8.04 |  | 10.0 | 9.14 |  | 10.0 | 7.46 |  | 8.0 | 6.58 |
| 8.0 | 6.95 |  | 8.0 | 8.14 |  | 8.0 | 6.77 |  | 8.0 | 5.76 |
| 13.0 | 7.58 |  | 13.0 | 8.74 |  | 13.0 | 12.74 |  | 8.0 | 7.71 |
| 9.0 | 8.81 |  | 9.0 | 8.77 |  | 9.0 | 7.11 |  | 8.0 | 8.84 |
| 11.0 | 8.33 |  | 11.0 | 9.26 |  | 11.0 | 7.81 |  | 8.0 | 8.47 |
| 14.0 | 9.96 |  | 14.0 | 8.10 |  | 14.0 | 8.84 |  | 8.0 | 7.04 |
| 6.0 | 7.24 |  | 6.0 | 6.13 |  | 6.0 | 6.08 |  | 8.0 | 5.25 |
| 4.0 | 4.26 |  | 4.0 | 3.10 |  | 4.0 | 5.39 |  | 19.0 | 12.50 |
| 12.0 | 10.84 |  | 12.0 | 9.13 |  | 12.0 | 8.15 |  | 8.0 | 5.56 |
| 7.0 | 4.82 |  | 7.0 | 7.26 |  | 7.0 | 6.42 |  | 8.0 | 7.91 |
| 5.0 | 5.68 |  | 5.0 | 4.74 |  | 5.0 | 5.73 |  | 8.0 | 6.89 |

1. (6pts) Compute the mean, variance, and standard deviation of and for each dataset. Compute the correlation between and for each dataset. (Note: correlation is a measure of how related a pair of variables are, where |correlation|=1 is a perfect linear relationship. See ?cor or ?cor.test) What do you notice about the values computed?

Df1: xmean = 9, xvar = 11, xsd = 3.32, ymean = 7.5, yvar = 4.13, ysd = 2.03, cor = 0.82

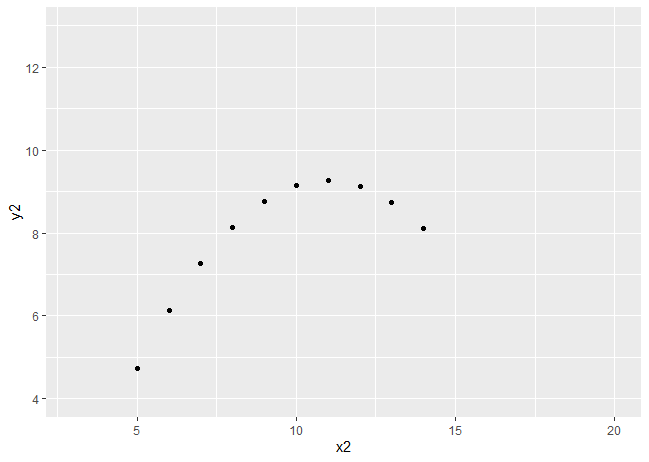
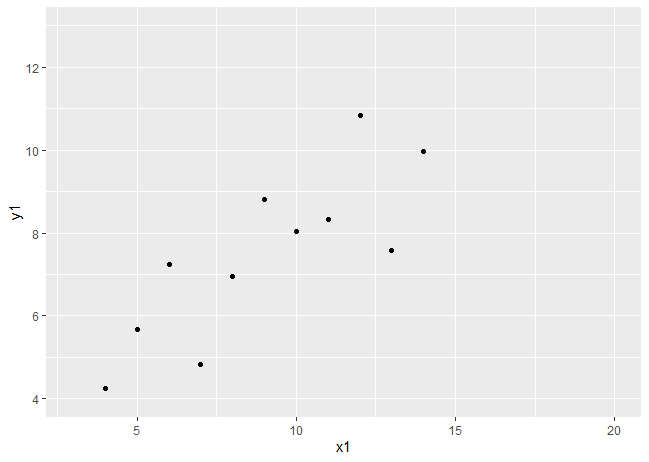
Df2: xmean = 9, xvar = 11, xsd = 3.32, ymean = 7.5, yvar = 4.13, ysd = 2.03, cor = 0.82

Df3: xmean = 9, xvar = 11, xsd = 3.32, ymean = 7.5, yvar = 4.12, ysd = 2.03, cor = 0.82

Df4: xmean = 9, xvar = 11, xsd = 3.32, ymean = 7.5, yvar = 4.12, ysd = 2.03, cor = 0.82

They are all the same, with little variance in the variance of the y column.

1. (4pts) Create a scatterplot of each of the four datasets. Use the xlim and ylim arguments in the plot function (or whatever equivalent in the software you choose) so that all four plots are on the same scale. (See ?plot.default for help)





1. (2pts) Based on the results of part (b) and (c), comment on why it is important to graph the data, and not just rely on summary statistics.

Though they all have the same summary statistics, they each have very different distributions and have their own patterns/outliers.

The file “NHANES702.csv” (located in Resources → Final Project on Sakai) contains data for the Final Project described in detail during class. You may review this in the recording on 9/1/2021, starting around the 4:25 mark.

1. (3pts) What type of study is the Final Project, where we are using 2017-2018 NHANES data? Choose all that apply from experimental, observational, prospective, retrospective, case-control, cohort, cross-sectional, longitudinal, randomized, etc.

Retrospective, cross-sectional, randomized, observational, case-control

1. (10pts) Read in the NHANES data into the software of your choice. In R, you may use the following code, but replacing the file path as appropriate. See ?read.csv for argument details.

dat <- read.csv(file="[path]/NHANES702.csv", header=TRUE, na.strings="",

stringsAsFactors=TRUE)

The table below describes SOME of the variables in the dataset. In the specifications column, fill in the data types for each variable based on your exploration of the csv file and the written descriptions given. List the categories for the categorical variables. Use the ?table function to get quick crosstabs in order to fill in specifications for variables that have 2 versions like marital status.

table(dat$Marital, dat$Married2cat, exclude=NULL)

|  |  |  |
| --- | --- | --- |
| Variable | Description | Specifications |
| Age | Age (years)  Ascertained at the time of screening. Individuals 80 and over were topcoded at 80 years.  A valid range is 18 ≤ Age ≤ 80 | continuous |
| Sex | Sex of the participant | A dichotomous variable where  F= female  M= Male |
| Race | Reported race and Hispanic origin information | A nominal variable where  1= Mexican American  2= Other Hispanic  3 = Non-Hispanic White  4 = Non-Hispanic Black  5 =  6 = Non-Hispanic Asian  7 = Other / Multi-Racial |
| Marital  Married2cat | Marital status | Marital is a categorical variable where  1= Married  2= Widowed  3= Divorced  4= Separated  5= Never married  6= Living with partner  Married2cat is a dichotomous variable where  0= Not married or living with partner (2, 3, 4, 5)  1= Married or living with partner (1, 6) |
| Educ  Educ2cat | Education  Ascertained by the question “What is the highest grade or level of school you have completed or the highest degree you have received?” | Educ is an ordinal variable where  1 = Less then 9th grade  2 = 9-12th grade (no diploma)  3 = High school graduate/GED or equivalent  4 = Some college or AA degree  5 = College graduate or above  Educ2cat is a binary variable where 0 = HS or less (1,2,3) and 1 = Some college or more (4,5) |
| BMI | Body mass index (kg/m2)  A valid range is 15 ≤ BMI < 70 | Continuous variable |
| Smoking | Smoking status  Derived from two questions: 1) “Have you smoked at least 100 cigarettes in your entire life?” and 2) “Do you now smoke cigarettes?”.  Current smoker is defined as yes/yes. Former smoker is defined as yes/no. Never smoker is defined as no/no. | A nominal variable where  0 = never  1 = former  2 = current |
| Alcohol12m | Alcohol use  Ascertained by the question “During the past 12 months, about how often did you drink any type of alcoholic beverage?” | An ordinal variable where  0 = none  1 = a few times  2 = monthly  3 - weekly |
| SBP | Systolic blood pressure (mm Hg)  Computed as the average of 4 readings.  A valid range is 70 ≤ SBP < 180 | A continuous variable |
| ACR | Urinary albumin to creatinine ratio (mg/g)  A valid range is 0 ≤ ACR < 6000 | A continuous variable |
| A1C | Glycohemoglobin (%)  A valid range is 4 ≤ A1C < 11 | A continuous variable |
| DM | Diabetes status  Glucose control is defined as  no self-report of DM/preDM AND (FPR < 100 or non-FPG < 200) AND A1C < 5.7 AND no DM medication  Prediabetes is defined as  self-report prediabetes OR 100 ≤ FPG < 126 OR 5.7 ≤ A1C < 6.5  Diabetes is defined as  self-reported diabetes OR self-reported borderline diabetes OR FPG ≥ 126 mg/dL OR non-FPG > 200 OR A1C ≥ 6.5% or diabetes medication | An ordinal variable where  0= GC/IND  1 = preDm  2 = DM |
| Sleep | Sleep (hours)  Number of hours usually sleep on weekdays or workdays | A continuous variable |
| PHQ9  PHQ5cat | Depression score  Derived from a 9-item instrument called the Patient Health Questionnaire (PHQ) to determine the frequency of depression symptoms over the past 2 weeks.  A valid range is 0 ≤ PHQ ≤27, where a higher score means more depression. PHQ score can be categorized into minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19), and severe (20-27). | PHQ9 is a discrete, quantitative variable  PHQ5cat is an ordinal variable where  0 = none-minimal  1 = mild  2 = moderate  3 = mod severe  4 = severe |

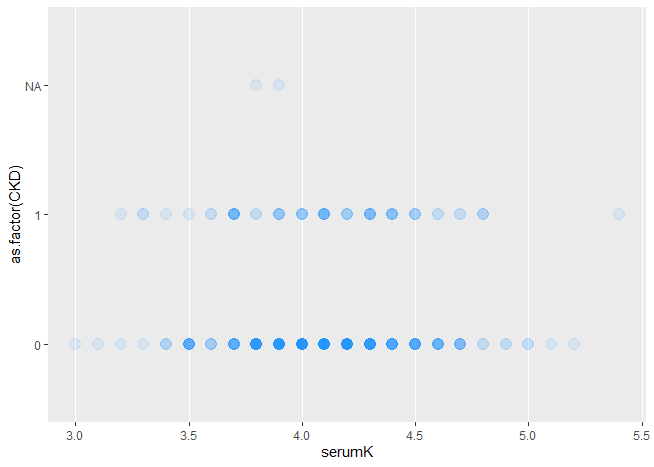
1. (4pts) What variables in the table above have values outside of valid ranges? (Note that you will need to check this for ALL variables in the Final Project!) Set invalid values to missing.

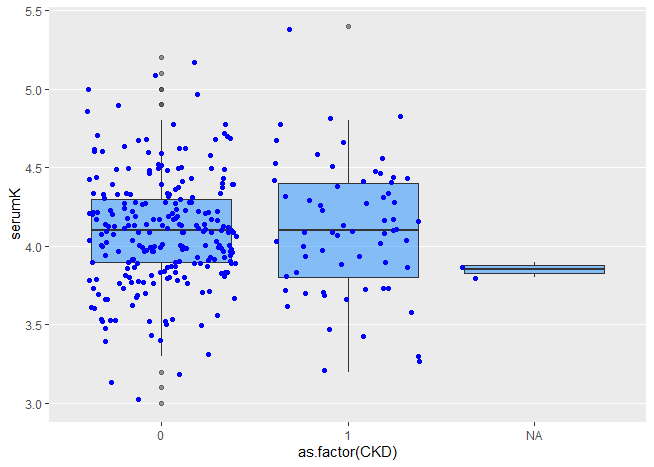
The 4 variables in the table that have invalid values are BMI, SBP, ACR, & A1C. These invalid values have all been set to NA.

1. (8pts) For each variable in the table, determine which measures of central location and spread are meaningful and most appropriate and what plots would be appropriate visualizations. Justify your answer. Hint: “none” is a possible answer.

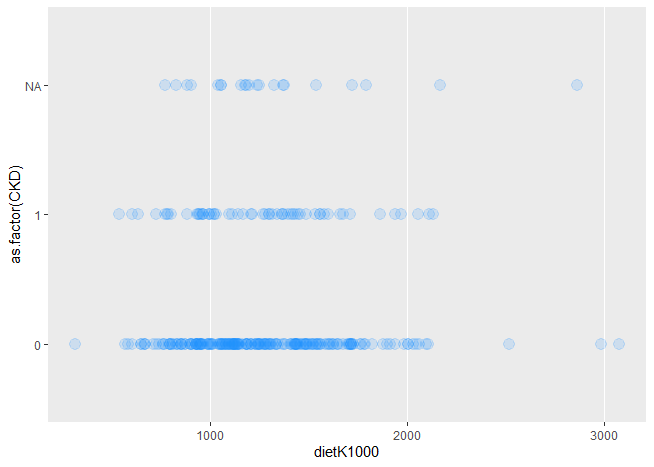
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Measure of Center | Measure of Spread | Plot | Justification |
| CKD | None | None | Barplot | Though the mode could be used, I elected to say none would be more fit, given there are only 2 outcomes excluding missing values. |
| eGFR | Mean | Standard deviation | Histogram, boxplot | The data is not very skewed, so the mean is an accurate measure of center, and is actually very close to the median. It is also continuous |
| SerumK | Mean | Standard deviation | Histogram, boxplot | Though values only range from 3-5.4, it is not really a categorical variable and still follows an approximately normal distribution. |
| dietK1000 | Mean | Standard Deviation | Histogram, boxplot | Continuous variable, not skewed |
| SEQN | NA | NA | NA | NA-Unique Identifier |
| Age | Median | IQR | Histogram, boxplot | Median because there is a bit of a skew, which pulls the mean from the median |
| Marital | Mode | None | Barplot | This is a categorical variable so it makes sense to show the counts of each of the categories compared to each other. A stacked barplot could also present benefits. |
| PHQ | Median | IQR | Histogram, boxplot, violinplot | Also categorical, so it does not make sense to say a 2 is half of 4. Also very right skewed. |

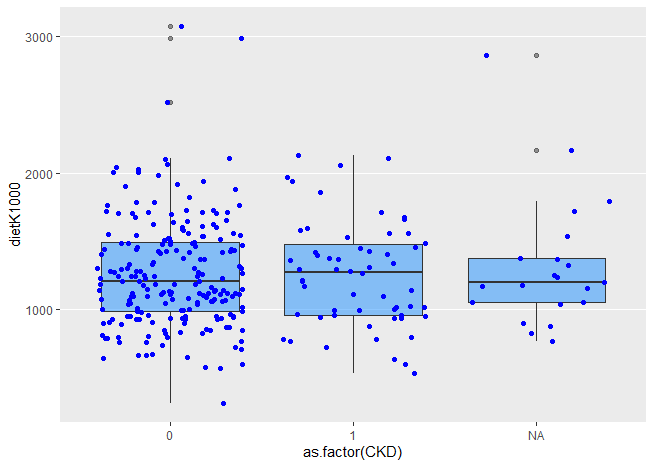
1. (5pts) Visually, does there appear to be a difference in potassium (serum or diet separately) by CKD status? Present plots of the data and describe them to answer this question.





These plots above do seem to show a relationship between CKD status and serumK. The darkness of the points represent how many observations are at that point in the first plot, and the second shows width-restrictive jitter point overlay on the boxplots for each level of CKD.





These plots also appear to show a relationship, though I would say not as strong. It is possible that this plot is deceiving though, because there are many more observations with no CKD (CKD == 0) than with. It appears that each of these distributions are similar.

May need later:

|  |  |  |  |
| --- | --- | --- | --- |
|  | No CKD | CKD | Total |
| Normal Serum K (K > 4) |  |  |  |
| Low-Normal Serum K (4 >= K >= 3) |  |  |  |
| Total |  |  |  |