Statistical Data Analysis

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

Exploring Relationships

Introduction

- So far, we have focused on using graphs and summary statistics to explore the distribution of individual variables.
- In this lecture we discuss using graphs and summary statistics to investigate relationships between two or more variables.
  - We want to develop a high-level understanding of the type and strength of relationships between variables.
- We start by exploring relationships between two numerical variables.
  - We then look at the relationship between two categorical variables.
- Finally, we discuss the relationships between a categorical variable and a numerical variable.

Two numerical variables

- For illustration, we use the *bodyfat* data
  - based on a study conducted by Dr. Fisher from Human Performance Research Center at Brigham Young University
    - The study involved measuring percent body fat as the target variable, along with several explanatory variables such as age, weight, height, and abdomen circumference for a sample of 252 men.
  - The collected data set *bodyfat* is available online at http://lib.stat.cmu.edu/datasets/bodyfat
  - You can also obtain this data set from the mfp package in R.
  - To install this package, enter the following command in R Console:
    - install.packages("mfp", dependencies=TRUE)

Two numerical variables

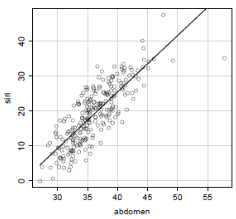
- Once the package is installed, it can be loaded into R using the following command:
  - library(mfp)
- Now you can access bodyfat by clicking
  - Data → Data in packages →Read data set from an attached package
- and selecting (doubleclicking) mfp under packages.
- You can learn more about this data set by looking at its accompanying help file.
  - In R-Commander, click
    - Data → Active data set→Help on active data set.

Two numerical variables

- Suppose that we are interested in examining the relationship between percent body fat and abdomen circumference among men.
  - Load the *bodyfat* set from the mfp package. Makesure *bodyfat* becomes the active data set and then view it.
  - For now, we are focusing on two variables, *siri* and *abdomen*.
    - The *siri* variable shows the percent body fat measurements derived based on body density using Siri's equation (percent body fat = 495/density−450).
    - The *abdomen* variable shows the abdomen circumference in centimeters.
- Both *siri* and *abdomen* are numerical variables.
  - A simple way to visualize the relationship between two numerical variables is with a scatterplot.

Scatterplot

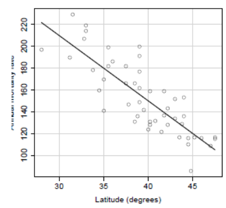
- In R-Commander, click
  - Graphs → Scatterplot and select *abdomen* for the x-variable and *siri* for the y-variable.
  - Under Options, uncheck Marginal boxplots and Smooth line.



- The plot suggests that the increase in percent body fat tends to coincide with the increase in abdomen circumference.
- The two variables seem to be related with each other.
  - The relationship is simply an association and should not be regarded as causation since the data come from an observational study.

Scatterplot

- As the second example, we examine the relationship between the annual mortality rate due to malignant melanoma for US states and the latitude of their geographical centers.



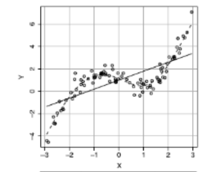
- The data are collected from the population of white males in the US during 1950–1969.
- You can obtain this data set, called *USmelanoma*, from the HSAUR2 package.
  - [Follow the above steps to install and load the package]
- The two variables are clearly associated since the increase in latitude tends to coincide with the decrease in mortality rate.

Scatterplot

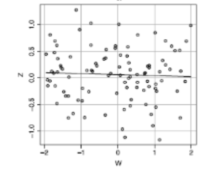
- Using scatterplots, we could detect possible relationships between two numerical variables.
  - In above examples, we can see that changes in one variable coincides with substantial systematic changes (increase or decrease) in the other variable.
- Since the overall relationship can be presented by a straight line, we say that the two variables have linear relationship.
  - We say that percent body fat and abdomen circumference have positive linear relationship.
  - In contrast, we say that annual mortality rate due to malignant melanoma and latitude have negative linear relationship.

Scatterplot

- In some cases, the two variables are related, but the relationship is not linear.

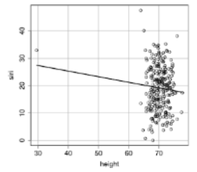


- In some cases, there is no relationship (linear or non-linear) between the two variables.

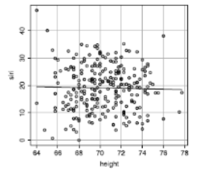


Scatterplot

- The scatterplot of percent body fat by height from the *bodyfat* data set.
  - The isolated point at the left of the graph is an outlier, which has a drastic influence on the overall pattern.



- The scatterplot of percent body fat by height after removing the outlier.
  - The two variables seem to be unrelated

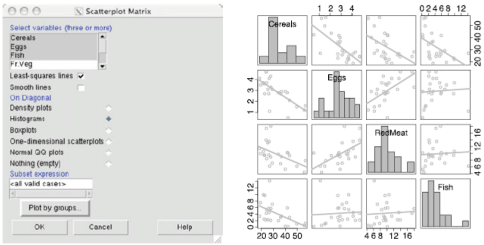


Scatterplot

- In practice, we should never remove an outlier just simply because it does not follow the overall pattern.
- Some outliers are due to rare events, which provide important information about the distribution of the corresponding variable.
- Even when we identify a data entry mistake, we should try to correct the mistake and keep the observation if possible.

Scatterplot Matrix

- Obtaining and viewing a *scatterplot matrix* in R-Commander.



  – The diagonal elements are histograms, and the off-diagonals are scatterplots with a trend line

Correlation

- Is a measure of similarities of two signals (cross-correlation)

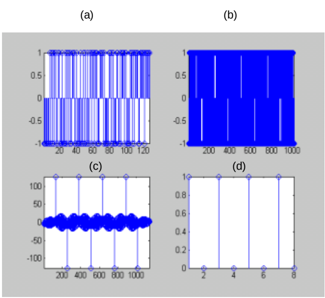
$$r_{xy}(k) = \sum_{n=0}^{N-1} x(n)y(k+n)$$

- Is a way to detect a known waveform in a noisy background (matched filter)

- Algorithm

```
for k=1:K+N-1
        for n=1:N
                y(k)=y(k)+a(n)*b(k+n-1);
        end
end
```


A correlation example



(a) A PN code
(b) A noisy binary signal (10101010) coded by the PN code in (a)
(c) Result of the correlation between (a) and (b)
(d) Recovered signal (10101010) after thresholding

Correlation

- To quantify the strength and direction of a linear relationship between two numerical variables,
  – we can use Pearson's correlation coefficient, r , as a summary statistic.
    - The values of r are always between -1 and +1.
    - The relationship is strong when r approaches -1 or +1.
    - The sign of r shows the direction (negative or positive) of the linear relationship.

Correlation

- Consider a set of observed pairs of values, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, for a sample of n observations.
- For these observed pairs of values, Pearson's correlation coefficient is calculated as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

  – For the two variable, s_x and s_y denote the sample standard deviations

Correlation

- Suppose that we have measured the height in inches and weight in pounds for five people.

Index	Height	Weight
1	62	160
2	71	198
3	65	173
4	73	182
5	60	143
Mean	66.2	171.2
Standard deviation	5.6	21.0

  – We denote height as X and weight as Y

Correlation

- Calculating Pearson's correlation coefficient for height and weight

Index	x	$x-\bar{x}$	y	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	62	-4.2	160	-11.2	47.04
2	71	4.8	198	26.8	128.64
3	65	-1.2	173	1.8	-2.16
4	73	6.8	182	10.8	73.44
5	60	-6.2	143	-28.2	174.84

$$r_{xy} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} = \frac{1}{4}\frac{421.8}{5.6\times 21.0} = 0.89$$

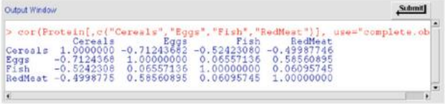
Correlation

- We can use R-Commander to calculate the sample correlation coefficient.
- To calculate r for percent body fat and abdomen circumference, make sure *bodyfat* is the active data set, then click
  - *Statistics → Summaries → Correlation matrix*
- Select both *abdomen* and *siri*. (You need to hold the *control* key.)
  - The output is in the form of a symmetric matrix called the *correlation matrix*, where the value in row i and column j is the correlation coefficient between the ith and j th variables.

Correlation

- Obtaining and viewing the correlation between percent body fat and abdomen circumference in R-Commander

- Correlation matrix for most of the numerical variables in the *Protein* data set



Sample Covariance

- If the standard deviations are removed from the denominator in Pearson's correlation coefficient, the statistic is called the sample covariance,

$$v_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

- Therefore

$$r_{xy} = \frac{v_{xy}}{s_x s_y}$$

Two categorical variables

- We now discuss techniques for exploring relationships between categorical variables.
- As an example, we consider the five-year study to investigate whether regular aspirin intake reduces the risk of cardiovascular disease.
    - ["Findings from the aspirin component of the ongoing Physicians' health study" in *New England Journal of Medicine* in 1988].
  - In this randomized experiment, 22071 physicians were randomly divided into two groups: 11037 physicians took an aspirin every other day, while 11034 physicians took a placebo. The investigators then recorded the number of people who suffered a heart attack within the five-year follow-up period.

Two categorical variables

- We usually use contingency tables to summarize such data.

	Heart attack	No heart attack	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

- Each cell shows
  - the frequency of one possible combination of disease status
    - heart attack or no heart attack
  - experiment group
    - placebo or aspirin
      - [A placebo is a substance or treatment with no active therapeutic effect. It may be given to a person in order to deceive the recipient into thinking that it is an active treatment]

Two categorical variables

- Using these frequencies, we can calculate the sample proportion of people who suffered from heart attack in each experiment group separately.
  - There were 11034 people in the placebo group, of which 189 had heart attack.
  - The proportion of people suffered from a heart attack in the placebo group is therefore

    $p_1 = 189/11034 = 0.0171.$
  - The proportion of people suffered from heart attack in the aspirin group is

    $p_2 = 104/11037 = 0.0094.$

25

Two categorical variables

- We refer to this as the risk (here, the sample proportion is used to measure risk) of heart attack.
- Substantial difference between the sample proportion of heart attack between the two experiment groups could lead us to believe that the treatment and disease status are related.
- One way of measuring the strength of the relationship is to calculate the difference of proportions, p_2-p_1.
  - Here, the difference of proportions is p_2-p_1 = -0.0077.

26

Two categorical variables

- The proportion of people suffered from heart attack reduces by 0.0077 in the aspirin group compared to the placebo group.
- We can present this difference as a percentage using the sample proportion (risk) in the placebo group as the baseline:

$$\frac{p_2 - p_1}{p_1} \times 100\% = \frac{-0.0077}{0.0171} \times 100\% = -45\%.$$

- This means that the risk of heart attack reduces by 45% in the aspirin group compared to the placebo group.

27

Two categorical variables

- Another common summary statistic for comparing sample proportions is the relative proportion p_2/p_1.
  - Since the sample proportions in this case are related to the risk of heart attack, we refer to the relative proportion as the relative risk.
- Here, the relative risk of sufering from heart attack is

  $p_2/p_1 = 0.0094/0.0171 = 0.55$

28

Two categorical variables

- This means that the risk of a heart attack in the aspirin group is 0.55 times of the risk in the placebo group.
- If the two sample proportions are equal, the relative proportion (risk) is equal to 1,
  - which is interpreted as no relationship between the two categorical variables.
- Values of the relative proportion away from 1 (either below 1 or above 1) indicate that the relationship is strong.

29

Two categorical variables

- It is more common to compare the sample odds,

$$o = \frac{p}{1-p}$$

  - where p is the sample proportion for the event of interest (e.g., heart attack).
- The odds of a heart attack in the placebo group, o_1, and in the aspirin group, o_2, are

$$o_1 = \frac{0.0171}{(1-0.0171)} = 0.0174, \quad o_2 = \frac{0.0094}{(1-0.0094)} = 0.0095.$$

30

Two categorical variables

- We usually compare the sample odds using the sample odds ratio

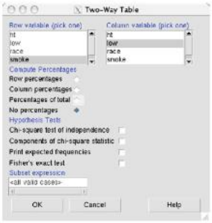
$$OR_{21} = \frac{o_2}{o_1} = \frac{0.0095}{0.0174} = 0.54.$$

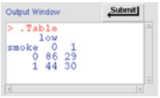
  - The index “21” shows that we are dividing the odds in the second group (here, the aspirin group) by the odds in the first group (here, the placebo group).
    - An odds ratio equal to 1 means that the odds are equal in both groups and is interpreted as no relationship between the two categorical variables.
    - Values of the odds ratio away from 1 (either greater than or less than 1) indicate that the relationship is strong.
  - Note that the odds ratio cannot be negative.
    - Therefore, its smallest possible value is zero.

31

Two categorical variables

- Contingency table for *smoke* and *low* in *birthwt* data set
  - For creating the contingency table for smoke and low, click
    - *Statistics → Contingency tables → Two-way table.*





32

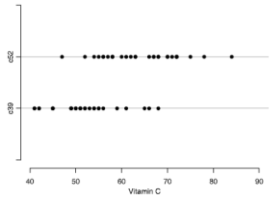
Numerical and Categorical Variables

- Very often, we are interested in the relationship between a categorical variable and a numerical random variable.
- When the sample size is small, we can visualize the relationship by simply creating dot plots of the numerical variable for different levels of the categorical variable.
- As an example, we use the *cabbages* data set available from the MASS package.

33

Numerical and Categorical Variables

- The dot plots of *ascorbic acid* (one form of vitamin C) *content* (numerical) by *cultivar* (categorical).



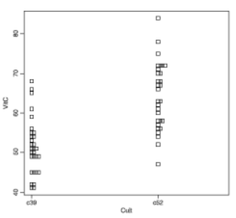
- The categorical variable has two possible categories: c39 and c52.
- It shows that the distribution of *vitamin C content* is different between the two *cultivars*.
- The central tendency for the observed values in the c39 group is around 50, whereas the central tendency for the c59 group is around 65.

- **In general, we say that two variables are related if the distribution of one of them changes as the other one varies.**

34

Numerical and Categorical Variables

- In the above example, the two variables, *vitamin C content* and *cultivar*, seem to be related.
- We can use R-Commander to create a dot plot (a.k.a. strip chart) similar to the one presented in previous slide.



- Strip chart for *vitamin C content* (*VitC*) by *cultivar* (*Cult*) from the *cabbages* data set
- Here, multiple observations with the same value of the numerical variable are stacked toward the right.
- Overall, vitamin C content tends to be higher in the c52 group compared to the c39 group.

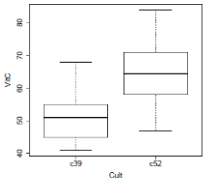
35

Numerical and Categorical Variables

- A more common way of visualizing the relationship between a numerical variable and a categorical variable is
  - to create boxplots of the numerical variable for different values of the categorical variable.
- This is especially useful when the sample size is large.
  - By focusing on some key aspects of the distributions, namely the five-number summaries, boxplots make the patterns easier to detect.
- In R-Commander, click
  - *Graphs→Boxplot*; select *VitC* as the Variable.
- Then click on
  - *Plot by groups* button and in the resulting window,
- Select
  - *Cult* as the *Groups variable.*

36

Numerical and Categorical Variables



- The resulting Boxplot of vitamin C content for different cultivars
- Summary statistics of *vitamin C content* by *cultivar* from the *cabbages* data set

```
Output Window
     mean        sd 0%  25%  50%   75% 100%  n
c39 51.5 7.123298 41   46 51.0 54.75   68 30
c52 64.4 8.455156 47   58 64.5 70.75   84 30
```

- This plot suggests that
  - vitamin C content tends to be higher in the c52 group compared to the c39 group.
    - This is indicative of a possible relationship between these two variables.

Numerical and Categorical Variables

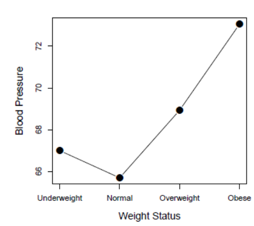
- In general, we say that two variables are related if the distribution of one of them changes as the other one varies.
- We can measure changes in the distribution of the numerical variable by obtaining its summary statistics for different levels of the categorical variable.
- It is common to use the difference of means when examining the relationship between a numerical variable and a categorical variable.
  - In the above example, the difference of means of vitamin C content is 64.4 -51.5 = 12.9 between the two cultivars.

Numerical and Categorical Variables

- When the categorical variable has multiple levels (categories), it is easier to compare the means across different levels using the plot of means.
- For example,
  - previously we created a categorical variable called *weight.status* based on *BMI* values in the *Pima.tr* data set.
  - This variable had four categories:
    - "Underweight", "Normal", "Overweight", and "Obese".
  - Here, we would like to investigate how blood pressure *bp* changes with *weight.status*, which is an *ordinal* variable

Numerical and Categorical Variables

- In R-Commander,
  - Click *Graphs* → *Plot of means* and
  - select *weight.status* as the *Factors* and *bp* as the *Response Variable*.
- For now, choose *no error bars*.



- The resulting graph shows that
  - compared to the Normal group, the average blood pressure increases for both Underweight and Overweight group.
  - The Obese group has the highest blood pressure average.
- Also, note that
  - as we move toward higher levels of weight group, average blood pressure first decreases and then increases.