STATS8: Introduction to Biostatistics

Exploring Relationships

Babak Shahbaba
UCI, Spring of 2012

Introduction

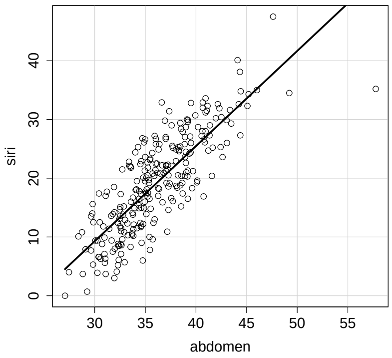
- So far, we have focused on using graphs and summary statistics to explore the distribution of individual variables.
- In this lecture we discuss using graphs and summary statistics to investigate relationships between two or more variables.
- We want to to develop a high-level understanding of the 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 numerica variables

- For illustration, we use the `bodyFat` data `http://lib.stat.cmu.edu/datasets/bodyfat`.
- Suppose that we are interested in examining the relationship between percent body fat and abdomen circumference among men.
- A simple way to visualize the relationship between two numerical variables is with a **scatterplot**.

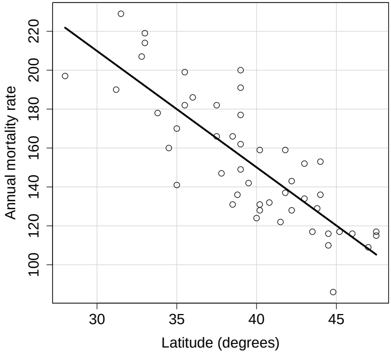
Scatterplot

- 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.



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.



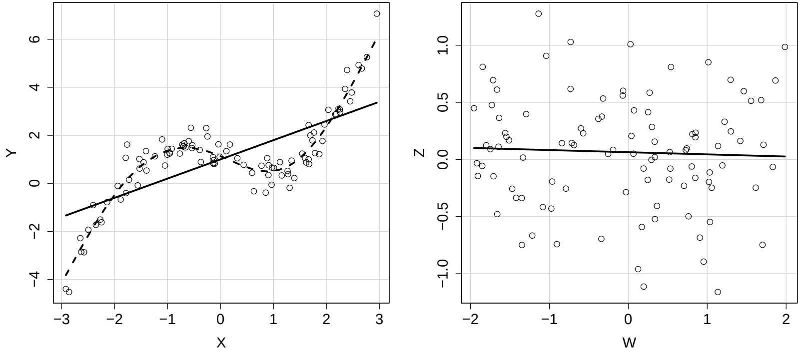
Scatterplot

- 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 (left plot).

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



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.

- For observed pairs of values, $(x_n, y_n), (x_n, y_n), \ldots, (x_n, y_n)$,

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

Correlation

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

Correlation

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

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

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.

- 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

Two categorical variables

- Each cell shows the frequency of one possible combination of disease status (heart attack or no heart attack) and experiment group (placebo or aspirin).

- 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 $p_2 = 104/11037 = 0.0094$.

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 $p_2 - p_1 = -0.0077$.
- The proportion of people suffered from heart attack reduces by 0.0077 in the aspirin group compared to the placebo group.

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 suffering from heart attack is $p_2/p_1 = 0.0094/0.0171 = 0.55$.

- This means that the risk of a heart attack in the aspirin group is 0.55 times of the risk in the placebo group.

Two categorical variables

- It is more common to compare the **sample odds**,

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

- The odds of a heart attack in the placebo group, o_1, and in the aspirin group, o_2, are

$$\begin{aligned} o_1 &= \frac{0.0171}{(1-0.0171)} = 0.0174, \\ o_2 &= \frac{0.0094}{(1-0.0094)} = 0.0095. \end{aligned}$$

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

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

Relationships Between Numerical and Categorical Variables

- Very often, we are interested in the relationship between a categorical variable and a numerical random variable.

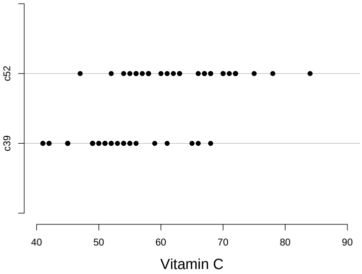


Figure: Dot plots of vitamin C content (numerical) by cultivar (categorical) for the cabbages data set from the MASS package.

Relationships Between Numerical and Categorical Variables

- A more common way of visualizing the relationship between a numerical variable and a categorical variable is to create boxplots.

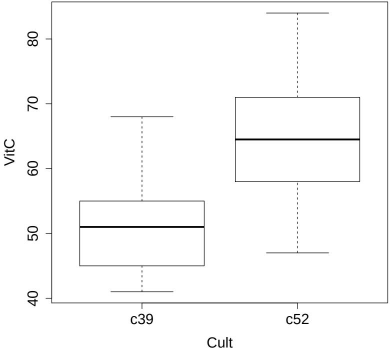


Figure: Boxplot of vitamin C content for different cultivars.

Relationships Between 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.

Relationships Between 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**.

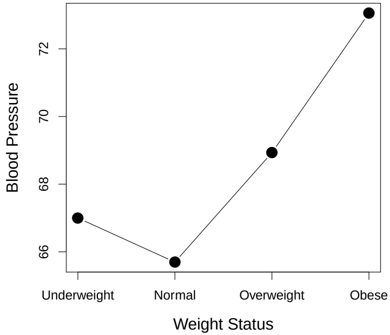


Figure: Plotting the means of `bp` for different weight group (which are defined based on BMI).