A Short Course on Biostatisitics

Part I: Fundamental Concepts & Exploratory Data Analysis

Babak Shahbaba, Ph.D.

Associate Professor, Department of Statistics
University of California, Irvine

July 30, 2018

Objective

- This course discusses some biostatistical methods, which involve applying statistical methods to biological problems.
- In statistics, we use **empirical evidence** to study **populations** and make informed **decisions**.
- We refer to this process as **statistical inference**.
- Here, we mainly focus on inference related to estimating unknown values and testing hypotheses.
- We first focus on summary statistics and visualization techniques for data exploration.
- Next, we will discuss some common methods to make formal statistical inference.

Overview

- The role of statistics in scientific studies
- Data exploration
- Summary statistics
- Data visualization techniques
- Exploring relationships
- Random variables and probability distributions

Statistics in Scientific Studies

The role of statistical analysis in science

- To study a population, we measure a set of characteristics, which we refer to as **variables**.
- The objective of many scientific studies is to learn about the **variation** of a specific characteristic (e.g., BMI, disease status) in the population of interest.
- In many studies, we are interested in possible **relationships** among different variables.
- We refer to the variables that are the main focus of our study as the **response** (or target) variables.
- In contrast, we call variables that explain or predict the variation in the response variable as **explanatory** variables or **predictors** depending on the role of these variables.
- Statistical analysis begins with a scientific problem usually presented in the form of a **hypothesis testing** or a **estimation** problem.

Study design

- Before we conduct our study, we need to design it in order to (Cox and Donnelly, 2011)
  - avoid systematic error arising from irrelevant sources that do not cancel out in the long term
  - reduce non-systematic error to a reasonable level by replication and other appropriate techniques
  - estimate realistically the extent of uncertainty in the final conclusion
  - ensure the scale of effort is appropriate; Goldilocks rule: not too limited, not too wasteful, just right!

Common study designs

- Survey studies
- Observational studies
  - Case-control studies
  - Cohort studies
- Experiments (Interventional; Clinical trials)
  - Placebo effect
  - Single-blind vs. double-blind studies

Study design– Methods

- **Observational studies**: Researchers are passive examiners, trying to have the least impact on the data collection process.
- **Experiments**, researchers attempt to control the process as much as possible.
  - Randomization
  - Replication
  - Blocking
- When studying the relationships between characteristics, it is important to distinguish between **association** and **causality**.
- In general, we use observational studies to discover association and use randomized experiments to establish causation.

Study design– Time frame

- Cross-sectional: Individuals are observed at one point in time, although some background information might also be used
- Prospective observational studies: following a group or cohort of individuals forward in time
- Retrospective observational studies: for each identified case, one or more controls are included in the sample, and explanatory features of all units of study are determined retrospectively
- Longitudinal: Typically, we have several observations per subject, for relatively large number of subjects
- Time series: Typically, we have many observations over time for relatively a small number of subjects

Confounding factors

- The relationship between the response variable and an explanatory variable could be influenced by a third variable variable.
- This usually happens when the **confounding** (lurking) variable influences both the response variable and the explanatory variable of interest.
- As a result, the relationship between the response variable and the explanatory variable of interest might seem stronger than, or weaker than, or even in the reverse direction of the true relationship.
- Factors such as diet, age, gender, ethnicity, and genetics are typical confounding factors in many scientific studies.

Sampling design

- We cannot of course observe the whole population of interest or conduct experiments on them.
- Instead, we select a **sample** of representative members from the population.
- Then with the methods of **statistical inference**, the conclusions based on the sample can cautiously be attributed to the whole population.

Sampling design

- The samples are selected **randomly** (i.e., with some probability) from the population.
- Unless stated otherwise, these randomly selected members of populations are assumed to be **independent**.
- The selected members (e.g., people, households, cells) are called **sampling units**.
- The individual entities from which we collect information are called **observation units**, or simply **observations**.
- Our sample must be representative of the population, and their environments should be comparable to that of the whole population.

Sampling design

- Some sampling schemes:
  - Simple random sampling
  - Stratified sampling
  - Cluster sampling
  - Multi-stage and temporal sampling
- We should always pay attention to whether we are dealing with independent or structured samples (e.g., repeated measures on each subject).
- We should alway respect the underlying structure of data and exploit it for better inference.

Data exploration

- After collecting data, the next step towards statistical inference and decision making is to perform **data exploration**, which involves visualizing and summarizing the data.
- The objective of data visualization is to obtain a high level understanding of the sample and their observed (measured) characteristics.
- To make the data more manageable, we need to further reduce the amount of information in some meaningful ways so that we can focus on the key aspects of the data. **Summary statistics** are used for this purpose.

Data exploration

- Using data exploration techniques, we can learn about the **distribution** of a variable.
- Informally, the distribution of a variable tells us the possible values it can take, the chance of observing those values, and how often we expect to see them in a random sample from the population.
- Through data exploration, we might detect previously unknown patterns and relationships that are worth further investigation.
- We can also identify possible data issues, such as unexpected or unusual measurements, known as **outliers**.

Statistical inference

- We collect data on a sample from the population in order to learn about the whole population.
- For example, Mackowiak, et al. (1992) wanted to find the average normal body temperature for the entire population and hypothesized that this average is less than $98.6\,^{\circ}\mathrm{F}$.
- That is, they wanted to **estimate** the unknown population average and perform **hypothesis testing**.
- To this end, they took a sample of 148 healthy subjects.

Statistical inference

- Note that in general the characteristics, relationships, and realities in the whole population always remain unknown.
- Therefore, there is always some **uncertainty** associated with our inference.
- In Statistics, the mathematical tool to address uncertainty is **probability**.
- The process of using the data to draw conclusions about the whole population, while acknowledging the extent of our uncertainty about our findings, is called **statistical inference**.
- The knowledge we acquire from data through statistical inference allows us to make decisions with respect to the scientific problem that motivated our study and our data analysis.

Data Exploration

Objective

- We start by focusing on data exploration techniques for one variable at a time.
- Our objective is to develop a high-level understanding of the data, learn about the possible values for each characteristic, and find out how a characteristic varies among individuals in our sample.
- In short, we want to learn about the **distribution** of variables.

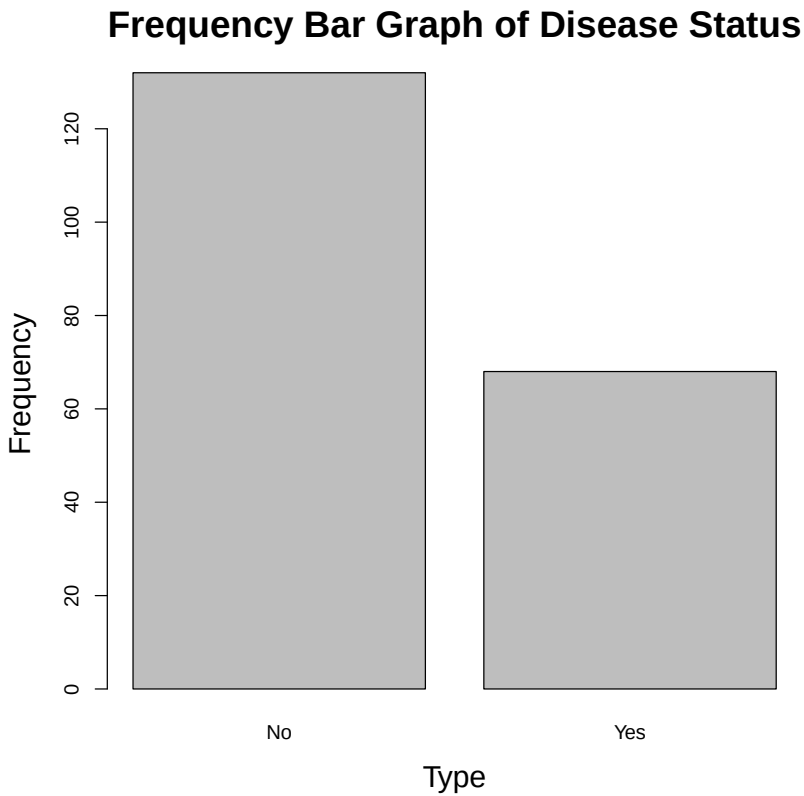
Variable types

- The visualization techniques and summary statistics we use for a variable depend on its type.
- Based on the values a variable can take, we can classify them into two general groups: **numerical** and **categorical**.
- Variables `age` and `bmi` are numerical variables since they take numerical values, and the numbers have their usual meaning.
- Variables `disease status` and `race` are categorical since their possible values consist of a finite number of categories.
- Sometimes we use numerical codings for categorical variables, but these numbers do not have their usual meaning.

Bar graph

- For categorical variables, **bar graphs** are one of the simplest ways of visualizing the data.
- Using a bar graph, we can visualize the possible values (categories) a categorical variable can take, as well as the number of times each category has been observed in our sample.
- The height of each bar in this graph shows the number of times the corresponding category has been observed.

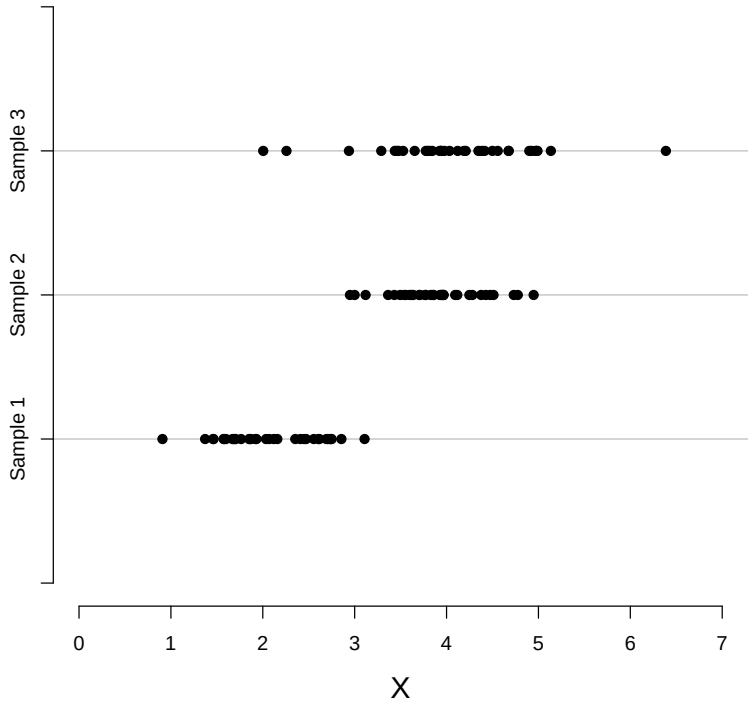
Bar graphs and frequencies



Exploring Numerical Variables

- For numerical variables, we are especially interested in two key aspects of the distribution: its **location** and its **spread**.
- The location of a distribution refers to the *central tendency* of values, that is, the point around which most values are gathered.
- The spread of a distribution refers to the *dispersion* of possible values, that is, how scattered the values are around the location.

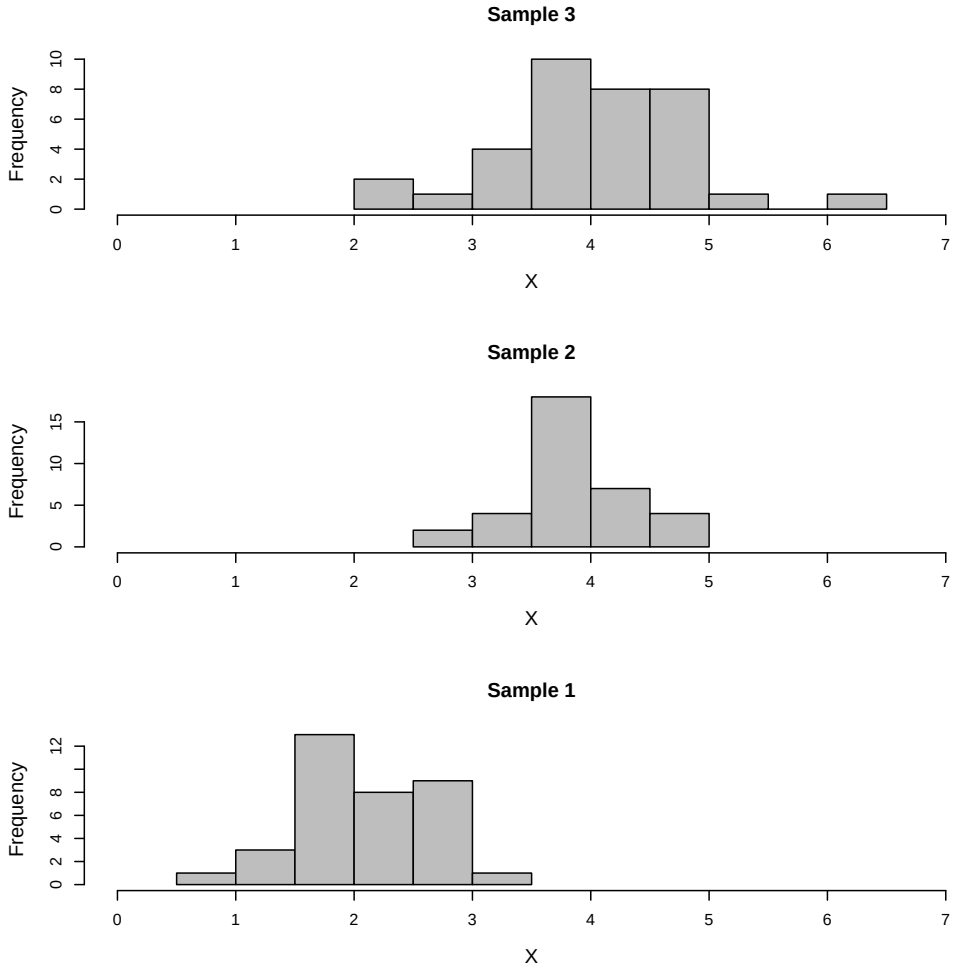
Exploring Numerical Variables



Histograms

- **Histograms** are commonly used to visualize numerical variables.
- A histogram is similar to a bar graph after the values of the variable are grouped (binned) into a finite number of intervals (bins).
- For each interval, the bar height corresponds to the frequency (count) of observation in that interval.

Histograms


Sample 3
Frequency
0 2 4 6 8 10
0 1 2 3 4 5 6 7
X
Sample 2
Frequency
0 5 10 15
0 1 2 3 4 5 6 7
X
Sample 1
Frequency
0 2 4 6 8 12
0 1 2 3 4 5 6 7
X

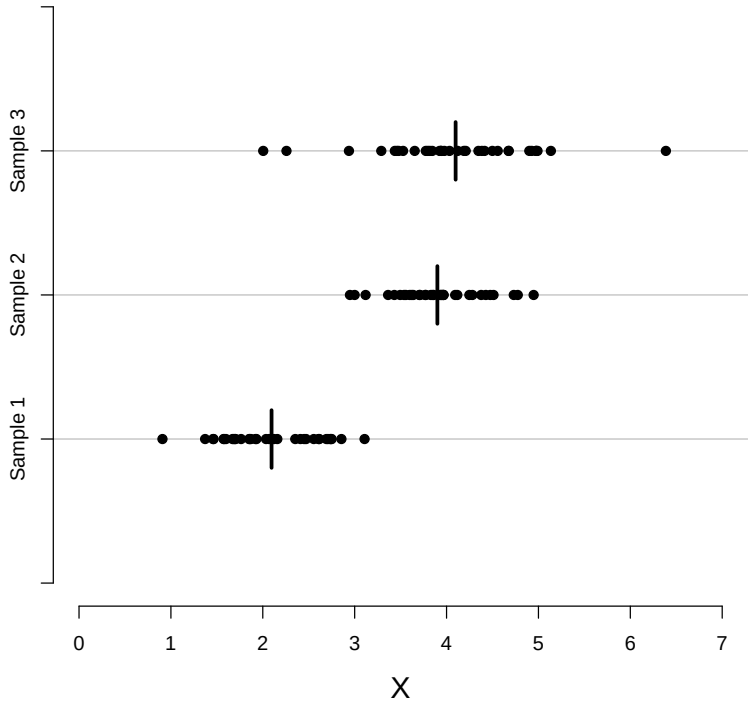
Sample mean

- Histograms are useful for visualizing numerical data and identifying their location and spread. However, we typically use summary statistics for more precise specification of the central tendency and dispersion of observed values.
- A common summary statistic for location is the **sample mean**.
- The sample mean is simply the average of the observed values. For observed values $x_1, \ldots, x_n$, we denote the sample mean as $\bar{x}$ and calculate it by

$$\bar{x} = \frac{\sum_i x_i}{n},$$

where x_i is the ith observed value of X, and n is the sample size.

Sample mean



Sample median

- The **sample median** is an alternative measure of location, which is less sensitive to outliers.
- For observed values $x_1, \ldots, x_n$, the median is denoted $\tilde{x}$ and is calculated by first sorting the observed values (i.e., ordering them from the lowest to the highest value) and selecting the middle one.
- If the sample size n is odd, the median is the number at the middle of the sorted observations. If the sample size is even, the median is the average of the two middle numbers.

Variance and standard deviation

- While summary statistics such as mean and median provide insights into the central tendency of values for a variable, they are rarely enough to fully describe a distribution.
- We need other summary statistics that capture the dispersion of the distribution.
- Consider the following measurements of blood pressure (in mmHg) for two patients:

$$\text{Patient A:} \quad x = \{95, 98, 96, 95, 96\}, \qquad \bar{x} = 96, \quad \tilde{x} = 96.$$
$$\text{Patient B:} \quad y = \{85, 106, 88, 105, 96\}, \quad \bar{y} = 96, \quad \tilde{y} = 96.$$

- While the mean and median for both patients are 96, the readings are more dispersed for Patient B.

Variance and standard deviation

- Two common summary statistics for measuring dispersion are the **sample variance** and **sample standard deviation**.
- These two summary statistics are based on the **deviation** of observed values from the mean as the center of the distribution.
- For each observation, the deviation from the mean is calculated as $x_i - \bar{x}$.

Variance and standard deviation

- The sample variance is a common measure of dispersion based on the squared deviations

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

- The square root of the variance is called the sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}},$$

Variance and standard deviation

Patient A			Patient B		
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
95	-1	1	85	-11	121
98	2	4	106	10	100
96	0	0	88	-8	65
95	-1	1	105	9	81
96	0	0	96	0	0
Σ	0	6	Σ	0	366

$$s^2 = 6/4 = 1.5$$
$$s = \sqrt{1.5} = 1.22$$

$$s^2 = 366/4 = 91.5$$
$$s = \sqrt{91.5} = 9.56$$

Quantiles

- Informally, the sample median could be interpreted as the point that divides the ordered values of the variable into two equal parts.
- That is, the median is the point that is greater than or equal to at least half of the values and smaller than or equal to at least half of the values.
- The median is called the 0.5 **quantile**.
- Similarly, the 0.25 quantile is the point that is greater than or equal to at least 25% of the values and smaller than or equal to at least 75% of the values.
- In general, the q quantile is the point that is greater than or equal to at least $100q\%$ of the values and smaller than or equal to at least $100(1-q)\%$ of the values.
- Sometimes, we refer to the q quantile as the $100q$th **percentile**.

Quartiles

- We can divide the ordered values of a variable into four equal parts using 0.25, 0.5, and 0.75 quantiles.
- The corresponding points are denoted Q_1, Q_2, and Q_3, respectively.
- We refer to these three points as **quartiles**, of which Q_1 is called the *first quartile* or the *lower quartile*, Q_2 (i.e., median) is called the *second* quartile, and Q_3 is called the *third quartile* or *upper quartile*.
- The interval from Q_1 (0.25 quantile) to Q_3 (0.75 quantile) covers the middle 50% of the ordered data.

Five-number summary and boxplot

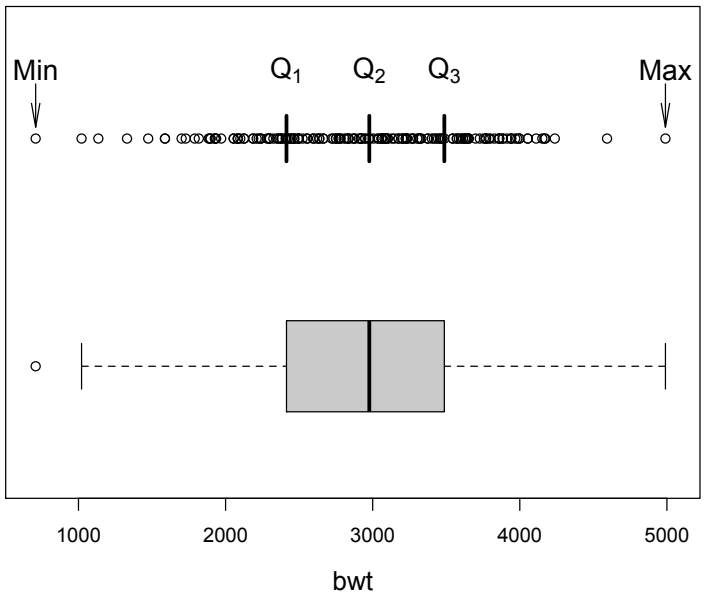
- The **minimum** (min), which is the smallest value of the variable in our sample, is in fact the 0 quantile.
- On the other hand, the **maximum** (max), which is the largest value of the variable in our sample, is the 1 quantile.
- The minimum and maximum along with quartiles (Q_1, Q_2, and Q_3) are known as **five-number summary**.
- These are usually presented in the increasing order: min, first quartile, median, third quartile, max.
- This way, the five-number summary provides 0, 0.25, 0.50, 0.75, and 1 quantiles.

Five-number summary and boxplot

- The five-number summary can be used to derive two measures of dispersion: the **range** and the **interquartile range**.
- The range is the difference between the maximum observed value and the minimum observed value.
- The interquartile range (IQR) is the difference between the third quartile (Q_3) and the first quartile (Q_1).

Five-number summary and boxplot

- To visualize the five-number summary, the range and the IQR, we often use a **boxplot** (a.k.a. **box and whisker** plot).



- Very often, boxplots are drawn vertically.

Five-number summary and boxplot

- The thick line at the middle of the "box" shows the median.
- The left side of the box shows the lower quartile.
- Likewise, the right side of the box is the upper quartile.
- The dashed lines are known as the **whiskers**.
- The whisker on the right of the box extends to the largest observed value or $Q_3 + 1.5 \times \text{IQR}$, whichever it reaches first.
- The whisker on the left extends to the lowest value or $Q_1 - 1.5 \times \text{IQR}$, whichever it reaches first.
- Data points beyond the whiskers are shown as circles and considered as possible outliers.

Exploration Relationships

Objective

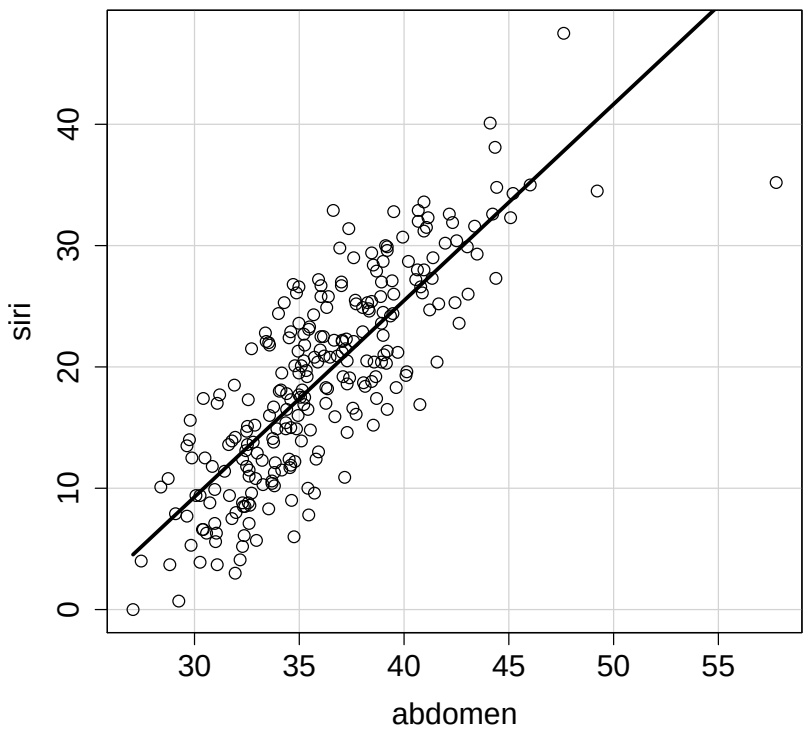
- Our objective is to develop a high-level understanding of the type and strength of relationships between variables.
- We start by exploring relationships between two numerical variables.
- We then look at the relationship between two categorical variables.
- Finally, we discuss the relationships between a categorical variable and a numerical variable.

Two numerical variables

- For illustration, we use the `bodyFat` data `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**.

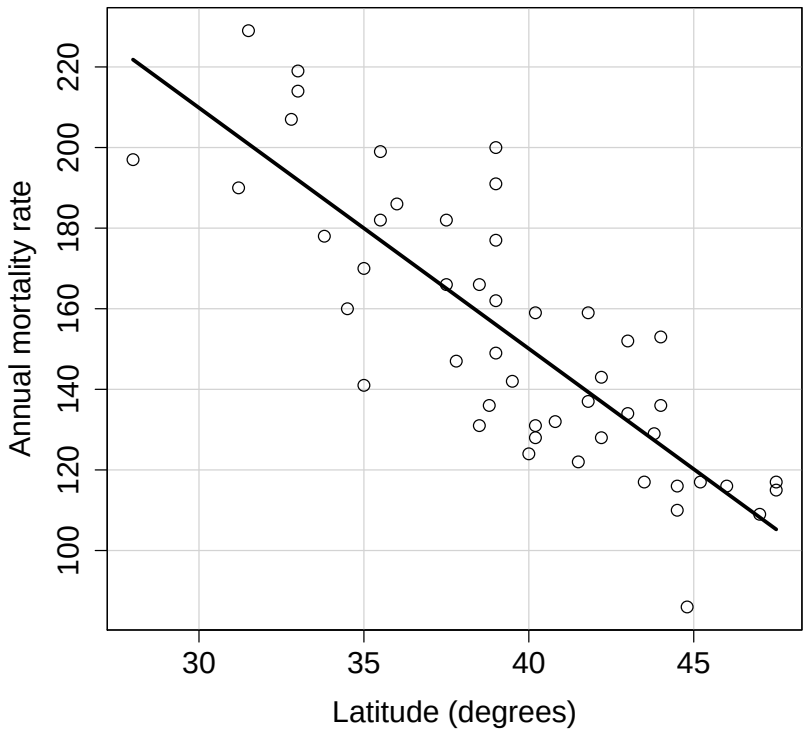
Scatterplot

- 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.

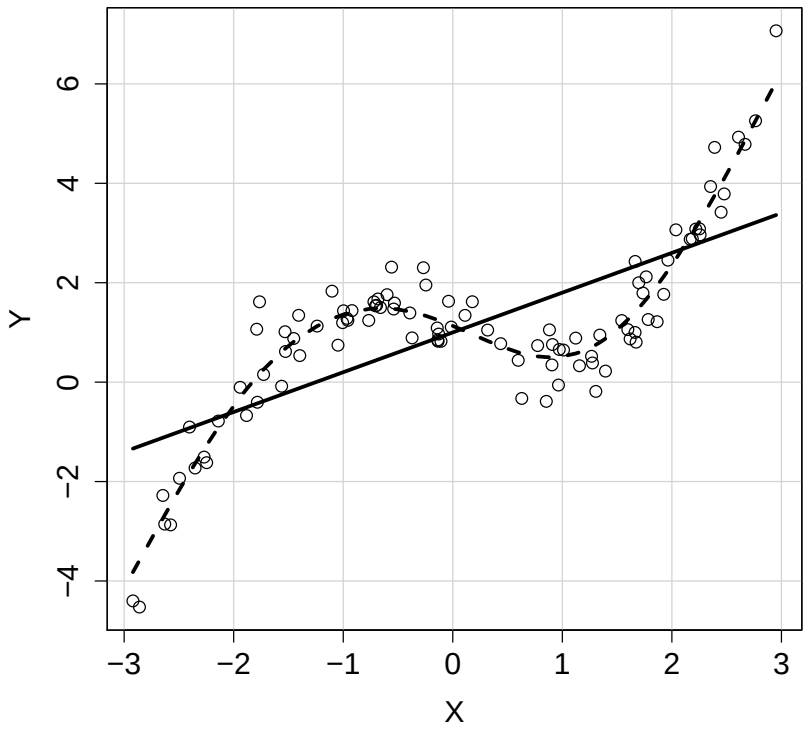


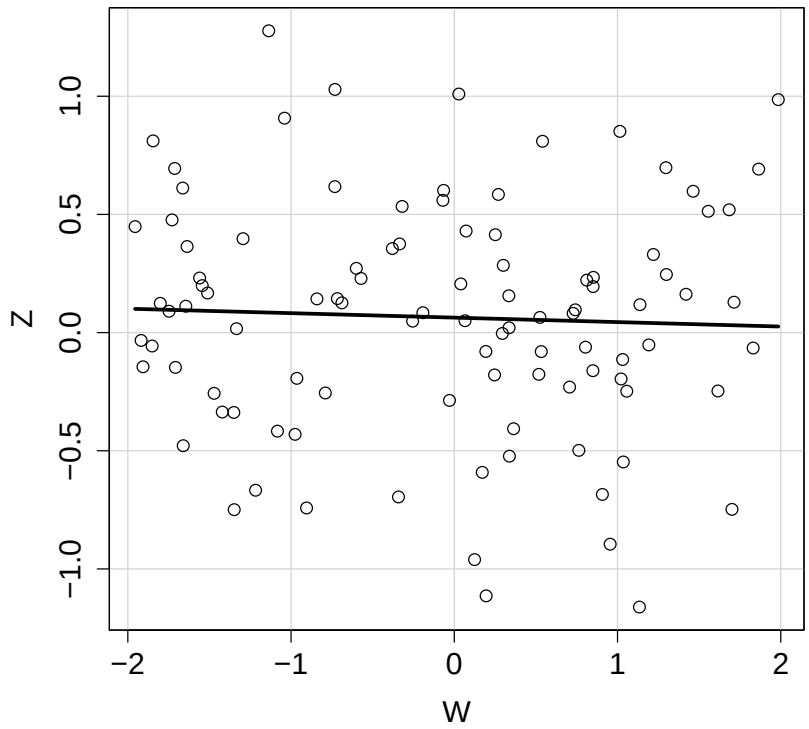
Scatterplot

- 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.

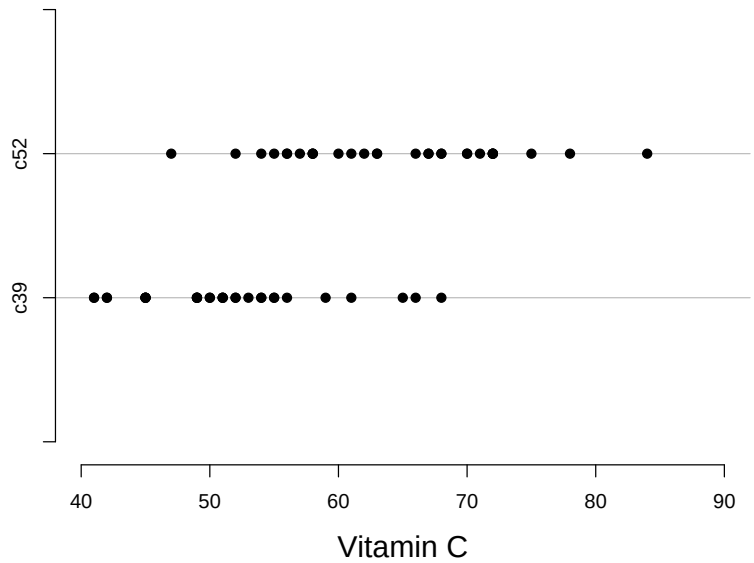


Figure: Dot plots of vitamin C content (numerical) by cultivar (categorical) for the `cabbages` data set from the `MASS` package.

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

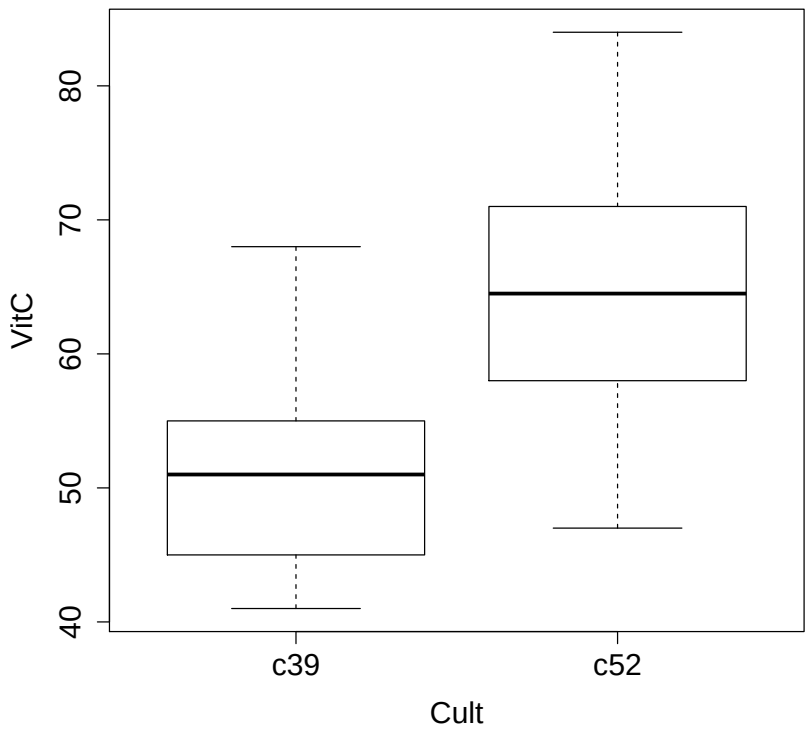


Figure: Boxplot of vitamin C content for different cultivars.

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.