Statistics

Leonard Vincent Ramil

Bachelor Degree in Applied Computer Science and Artificial Intelligence
Sapienza University of Rome

Contents

1 Statistics **4**
1.1 Statistical Science: Description and Inference . . . 4
1.1.1 Definition: Statistics and Statistical Science . . . 4
1.1.2 Statistical Methods: Design, Description and Inference . . . 4
1.1.3 Why Study Statistics? . . . 5
1.1.4 What Do Statisticians Do? . . . 5
1.1.5 Descriptive Statistics and Inferential Statistics . . . 5
1.1.6 Population vs Sample . . . 5
1.1.7 Sample Statistics and Population Parameters . . . 6
1.1.8 Randomness, Variability and Margin of Error . . . 6
1.2 Types of Data and Variables . . . 7
1.2.1 Variables . . . 7
1.2.2 Categorical and Quantiative Variables . . . 7
1.2.3 Categorical Variables Are Nominal or Ordinal or Binary . . . 8
1.2.4 Quantitative Variables Are Discrete or Continuous . . . 8
1.2.5 Distribution of a Variable . . . 8
1.2.6 Frequency Table and Relative Frequencies . . . 9
1.3 Good and Poor Ways to Sample . . . 9
1.3.1 Remark: Population, Census, Sample . . . 9
1.3.2 Sampling Frame and Sampling Design . . . 10
1.3.3 Simple Random Sampling . . . 10
1.3.4 Methods of Collecting Data in Sample Surveys . . . 10
1.3.5 Sources of Potential Bias in Sample Surveys . . . 11
1.3.6 Observational and Experimental Data . . . 11
1.4 Sampling Distribution . . . 12
1.4.1 Statistic vs Parameter . . . 12
1.4.2 Sampling Distributions: Probability Distributions for Statistics . . . 12
1.4.3 How Sample Proportions Vary Around the Population Proportion . . . 13
1.4.4 How Sample Means Vary Around the Population Mean . . . 13
1.4.5 Law of Large numbers and Central Limit Theorem . . . 14

2 Descriptive Statistics **17**
2.1 Graphical Summaries of Data . . . 17
2.1.1 Graphs for Categorical Data . . . 17
2.1.2 Graphs for Quantitative Data . . . 19
2.1.3 Shape of Distribution . . . 20
2.1.4 Time Plots: Displaying Data over Time . . . 22
2.2 Measuring the Center of Quantitative Data . . . 22
2.2.1 Calculating the Mean . . . 23
2.2.2 Calculating the Median . . . 24
2.2.3 Calculating the Mode . . . 24
2.2.4 Comparing the Mean and Median . . . 25
2.3 Measuring the Variability of Quantitative Data . . . 25
2.3.1 The Range . . . 26

2.3.2 The Standard Deviation . . . 26
2.4 Using Measures of Position to Describe Variability . . . 27
2.4.1 Measures of Position: The Quartiles and Percentiles . . . 27
2.4.2 Measuring Variability: The Interquartile Range . . . 29
2.4.3 Detecting Potential Outliers . . . 29
2.4.4 The Box Plot: Graphing a Five-Number Summary of Positions . . . 30
2.4.5 The Z-Score . . . 32
2.5 Concept of Association/Relationship . . . 32
2.5.1 Response and Explanatory Variables . . . 32
2.5.2 Association Between Two Variables . . . 33
2.6 The Association Between Two Categorical Variables . . . 33
2.6.1 Contingency Tables . . . 33
2.6.2 Calculate Conditional Proportions or Percentages . . . 34
2.7 The Relationship Between Two Quantitative Variables . . . 34
2.7.1 Looking for a Trend: The Scatterplot . . . 34
2.7.2 How to Examine a Scatterplot: Positive, Negative and Linear Relationship 35
2.7.3 The Correlation Coefficient . . . 36
2.8 Cautions in Analyzing Associations . . . 37
2.8.1 Association Does Not Imply Causation . . . 37
2.8.2 Extrapolation could be Dangerous . . . 37

3 Inferential Statistics **38**
3.1 Point Estimates and Confidence Intervals . . . 38
3.1.1 Overview: Point Estimate and Confidence Intervals . . . 38
3.1.2 Properties of Point Estimates (or/and Estimator) . . . 38
3.1.3 Interval Estimator: Confidence Intervals for Parameters . . . 40
3.2 Constructing Confidence Intervals . . . 42
3.2.1 Concept of Confidence Interval . . . 42
3.2.2 Confidence Interval to Estimate a Population Proportion . . . 43
3.2.3 Confidence Interval to Estimate a Population Mean . . . 44
3.3 Significance Test . . . 46
3.3.1 What is Significance Test? . . . 46
3.3.2 The Steps of a Significance Test . . . 46
3.3.3 Significance Test About Proportions . . . 48
3.3.4 Significance Test About Means . . . 49
3.3.5 Decisions and Types of Errors in Significance Tests . . . 50
3.3.6 Optimal Test . . . 50
3.4 Comparing Two Population Means or Proportions . . . 51
3.4.1 Sampling Distribution of the Difference Between Two Sample Proportions 51
3.4.2 ... with different Proportion . . . 51
3.4.3 Confidence intervals for the difference in two population proportions (different Proportion) . . . 51
3.4.4 ... with common Proportion . . . 52
3.4.5 Sampling Distribution for the Difference of Two Sample Means . . . 52
3.4.6 ... with unequal Variances . . . 52
3.4.7 Confidence Interval for the Difference Between Two Population Means (with unequal Variances) . . . 52
3.4.8 ... with common Variances . . . 53
3.4.9 Confidence Interval for the Difference Between Two Population Means (with common Variances) . . . 53

4 Statistical Models **54**
4.1 Predicting the Outcome of a Variable . . . 54
4.1.1 Definition of a Regression Line . . . 54
4.1.2 Interpreting the y-Intercept and Slope . . . 54
4.1.3 Concept of Error . . . 54
4.1.4 Concept of Residual . . . 55
4.1.5 Residual Sum of Squares (RSS) . . . 55
4.1.6 The Method of Least Squares Yields the Regression Line . . . 56
4.1.7 r-Squared (r^2) . . . 56
4.2 Concept of Regression . . . 56
4.2.1 Definition: Regression . . . 56
4.2.2 Two Distinct Goals . . . 57
4.2.3 Sample vs Population . . . 57
4.2.4 Notations . . . 57
4.3 Model for Simple Linear Regression . . . 58
4.3.1 The Linear Model Describes a Conditional Expectation . . . 58
4.3.2 Sample Estimates of Model . . . 59
4.3.3 How to decide the best estimates based on the sample? . . . 59
4.4 Sampling Distribution of b_0 and b_1 in Linear Regression . . . 61
4.4.1 Remark Normal Distribution . . . 61
4.4.2 Sampling Distribution of b_0 and b_1 . . . 61
4.4.3 Standardized b_0, b_1 . . . 63
4.5 Statistical Inference Tools . . . 63
4.5.1 Hypothesis Testing for slope β_1 . . . 64
4.5.2 Hypothesis Testing for intercept β_0 . . . 65
4.5.3 Confidence Interval for slope β_1 . . . 65
4.5.4 Hypothesis Testing v.s. Confidence Interval . . . 65
4.6 Confidence Interval and Prediction Inteval . . . 65
4.6.1 Confidence Interval for $E(Y)$ . . . 65
4.6.2 Prediction Interval for y . . . 66
4.6.3 CI for $E(Y)$ and PI for y : Comparison . . . 67
4.6.4 Important Notes . . . 67
4.7 Multiple Regression Model . . . 67
4.7.1 Linear Model . . . 67
4.7.2 Notation for Multiple Regression Model . . . 68
4.7.3 Matrix Notation for Multiple Regression Model . . . 68
4.7.4 Estimates (Least Square Criterion) . . . 69
4.7.5 Two Notation of Estimation . . . 70
4.8 Sampling Distribution of b in Multiple Regression . . . 70
4.8.1 Sampling Distribution of b . . . 70
4.8.2 Multicollinearity . . . 70
4.8.3 Model fitting . . . 71
4.8.4 Validate Model Assumptions . . . 71
4.8.5 Inference: Testing Significance of Each β Coefficient . . . 72
4.9 Analysis of Variance (ANOVA) . . . 72
4.9.1 ANOVA Table . . . 72
4.9.2 A decomposition of variance in y . . . 72
4.9.3 A Significance Test: F Test . . . 73

Chapter 1

Statistics

1.1 Statistical Science: Description and Inference

1.1.1 Definition: Statistics and Statistical Science

A common concept of statistic is that is merely a number calculated from data, the observations that provide information about the subject matter. But the field of statistical science has a much broader sense: as a field that gives us a way of gathering and analyzing the data in an objective manner.

Definition: Statistical Science

Statistical science is the science of developing and applying methods for collecting, analyzing, and interpreting data.

Definition: Statistics

Statistics is the art and science of collecting, presenting, and analyzing data to answer an investigative question. Its ultimate goal is translating data into knowledge and an understanding of the world around us. In short, statistics is the art and science of learning from data

Statistics and Probability

Many methods of statistical science incorporate reasoning using tools of probability. The methods enable us to deal with uncertainty and variability in virtually all scientific fields. With statistical methods, we learn from the data while measuring, controlling, and communicating uncertainty.

1.1.2 Statistical Methods: Design, Description and Inference

Statisticsl sciende has three aspects:

1. **Design:** Planning how to gather relevant data for the subject matter of interest.
2. **Description:** Summarizing the data.
3. **Inference:** Making evaluations, such as estimations and predictions, based on the data.

Design

- Stating the goal and/or statistical question of interest and planning how to obtain data that will address them
- Design refers to planning how to obtain data that will efficiently shed light on the statistical question of interest.

Description

- Description refers to summarizing the data, to mine the information that the data provide. For any study, the raw data are a complete listing of observations that can be overwhelming for comprehension.
- To present the results, we reduce the data to simpler and more understandable form without distorting or losing much information.
- Graphs, tables, and numerical summaries such as averages and percentages are called descriptive statistics.

Inference

- Making decisions and predictions based on the data for answering the statistical question
- Inference refers to using the data to make estimations and other sorts of evaluations, such as predictions. These evaluations take into account random variability that occurs with the characteristics measured and the resulting uncertainty in decision-making.
- Data-based evaluations such as estimations and predictions are called statistical inferences.

1.1.3 Why Study Statistics?

- To evaluate numerical facts
- To perform statistical data analysis or to interpret the results of sampling
- To make inference about the population through the sample

1.1.4 What Do Statisticians Do?

1. **Gather Data:** Draw a random sample of students. The sample size depends on how accurate you need your inference to be and the margin of error you can tolerate.
2. **Summarize data:** Summarize data from the sample
3. **Analyze data:** Analyze the data through statistical techniques and make inference through confidence interval and hypothesis testing.
4. **Draw conclusions and report the results:** Write reports and support your conclusions including plots, tabular and numerical displays.

1.1.5 Descriptive Statistics and Inferential Statistics

Description in Statistical Analyses

Descriptive statistics refers to methods for summarizing the collected data (where the data constitute either a sample or a population). The summaries usually consist of graphs and numbers such as averages and percentages.

Inference in Statistical Analyses

Inferential statistics refers to methods of making decisions or predictions about a population, based on data obtained from a sample of that population.

1.1.6 Population vs Sample

The entities on which a study makes observations are called the sample subjects. Usually the subjects are individual people, such as in a survey, but they need not be. Although we obtain data for the sample subjects, our ultimate interest is on the population from which the sample is taken.

Definition: Population and Sample

- **Population:** The population is the total set of subjects of interest.
- **Sample:** A sample is the set of subjects from the population for which data are available. in essence, a subset of the population from which information is actually collected

Remind: We observe samples but are interested in populations

1.1.7 Sample Statistics and Population Parameters

To distinguish between a descriptive statistic calculated for a sample and the corresponding characteristic of the population, we use the term parameter for the population characteristic.

Definition: Parameter and Statistics

A parameter is a numerical summary of the population. A statistic is a numerical summary of a sample taken from the population.

Remark

In practice, our primary interest is in the values of parameters rather than sample descriptive statistics. Parameter values are almost always unknown. The sample and statistics describing it help us to make inferences about the unknown parameter values.

A key aspect of statistical inference involves reporting the precision of the sample statistic that estimates the population parameter.

1.1.8 Randomness, Variability and Margin of Error

Randomness

Random is often thought to mean chaotic or haphazard, but randomness is an extremely powerful tool for obtaining good samples and conducting experiments. A sample tends to be a good reflection of a population when each subject in the population has the same chance of being included in that sample: This is the basis of random sampling.

- **Simple Random Sampling**
  Random sampling allows us to make powerful inferences about populations, which is designed to make the sample representative of the population. Each subject in the population has the same chance of being included in the sample
- **Randomness**
  Randomness is crucial to insuring that the sample is representative of the population so that powerful inferences can be made

Variability

An essential goal of Statistics is to study and describe variability: How observations vary from person to person within a sample, and how statistics computed from samples vary from one sample to the next, in essence, between samples. So we have to face with two different sources of variability:

1. **Within Sample Variability**
   People are different from each other, so, not surprisingly, the measurements we make on them vary from person to person. So, measurements may vary from subject to subject

2. **Between Sample Variability**
   Just as observations on people vary, so do results computed from samples. So, measurements may vary from sample to sample

A very interesting and useful, yet surprising, fact with random sampling is that the amount of variability in the results from one sample to another is actually quite predictable.

1.2 Types of Data and Variables

1.2.1 Variables

For the characteristics we measure in a study, variability occurs naturally among subjects in a sample or population. Any characteristic that we can measure for the subjects is called a variable. The term reflects that values of the characteristic vary among subjects.

Definition: Variable and Measurement Scale

- **Variable**
  A variable is a characteristic that can vary in value among subjects in a sample or population.

- **Measurement Scale**
  The values the variable can take form a measurement scale. The valid statistical methods for a variable depend on its measurement scale.

1.2.2 Categorical and Quantiative Variables

What is the Observations?

The data values that we observe for a variable are called observations. Each observation can be a number, such as the number of centimeters of precipitation in a day. Or each observation belongs to a category, such as "yes" or "no" for whether it rained.

Definition: Categorical (Qualitative) Variables

A variable is called categorical when the measurement scale is a set of categories so if each observation belongs to one of set of distinct categories

Definition: Quantitative Variables

A variable is called quantitative when the measurement scale has numerical values that represent different magnitudes of the variable

Key Features

- **For categorical variables:** a key feature is the percentage of observations in each of the categories
- **For quantitative variables:** a key features are the center (location) and spread (variability)

1.2.3 Categorical Variables Are Nominal or Ordinal or Binary

Qualitative (Categorical) variable each observation belongs to one of a set of categories order between them. Categorical values may be:

- **Binary:** where there are two choices
- **Nominal:** A categorical variable is nominal if its categories do not have an inherent ordering, in essence where no hierarchy is implied
- **Ordinal:** A categorical variable is ordinal if its categories can be ordered, in essence where the categories imply levels with hierarchy or order of preference

1.2.4 Quantitative Variables Are Discrete or Continuous

Quantitative variable each observations on it take numerical values that represent different magnitudes of the variable. Quantitative values can be:

- **Discrete:** A quantitative variable is descrete if they are possible values form a set of separate numbers, such as 0,1,2,3,... The set of possible values is not dense.
- **Continuous:** A quantitative variable is continuos if they are possible values from an interval. The set of possible values is dense.

1.2.5 Distribution of a Variable

The first step in analyzing data collected on a variable is to look at the observed values by using graphs and numerical summaries. The goal is to describe key features of the distribution of a variable.

Definition: Distribution

The distribution of a variable describes how the observations fall (are distributed) across the range of possible values.

Categorical and Quantitative Distribution

- The distribution for a categorical variable then simply shows all possible categories and the number (or proportion) of observations falling into each category
- The distribution for a quantitative variable, the entire range of possible values is split up into separate intervals, and the number (or proportion) of observations falling in each interval is given.

Features to Look

- **Categorical Distribution**
  Features to look for in the distribution of a categorical variable are the category with the largest frequency, called the modal category, and more generally how frequently each category was observed.

- **Quantitative Distribution**
  Features to look for in the distribution of a quantitative variable are its
  - shape (do observations cluster in certain intervals and/or are they spread thin in others?)
  - center (where does a typical observation fall?)
  - variability (how tightly are the observations clustering around a center?)

1.2.6 Frequency Table and Relative Frequencies

A frequency table displays:

- the distribution of a variable numerically
- the proportion or percentages of the number of observations falling in each category.

Definition: Frequency Table

A frequency table is a listing of possible values for a variable, together with the number of observations for each value.

Raw data

Code	Gender
000001	F
000002	M
...	...
100000	F

Frequency table

Gender	n_i	f_i	p_i
F	1000	0.01	1
M	99000	0.99	99
	100000	1.00	100

Definition: Proportion and Pecentage

- **Proportion**
  The proportion of observations falling in a certain category is the number of observations in that category divided by the total number of observations.

$$\text{Proportion} = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}} \tag{1.1}$$

- **Percentage**
  The percentage is the proportion multiplied by 100.
- **Relative Frequencies**
  Proportions and percentages are also called relative frequencies and serve as a way to summarize the distribution of a categorical variable numerically.

1.3 Good and Poor Ways to Sample

1.3.1 Remark: Population, Census, Sample

Population

A population is any entire collection of people, animals, plants or things from which we collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

Census

A census is a survey that attempts to count the number of subjects (persons, firms etc) in the population and to measure certain characteristics about them.

Sample

A sample survey selects a sample(=subset) of people (subjects) from a population and interviews (measures some variables on) them to collect data. We consider only a sample instead of the entire population

1.3.2 Sampling Frame and Sampling Design

Sampling Frame

The sampling frame is the list of subjects in the population from which the sample is taken, ideally it lists the entire population of interest.

Sampling Design

The sampling design determines how the sample is selected.

Type of sampling

There exists more type of sampling schemes:

- **Simple Random Sampling:** that is the simpliest
- **Stratified Random Sampling:** divide the population into separate groups ("strata") and then select a simple random sample from each stratum.
- **Cluster random sampling:** divide the population into a large number of clusters, such as city blocks, and select a simple random sample of the clusters.

1.3.3 Simple Random Sampling

Definition: Random Sampling

Random Sampling is the best way of obtaining a sample that is representative of the population.

Definition: A Simple Random Sample

A simple random sample of 'n' subjects from a population is one in which each possible sample of that size has the same chance of being selected.

Random Sample

A simple random sample is often just called a random sample. The "simple" adjective distinguishes this type of sampling from more complex random sampling designs that will be presented in the sequel.

1.3.4 Methods of Collecting Data in Sample Surveys

Sample Surveys

Some studies sample people from a population and interview them to collect data. This method of data collection is called a sample survey.

Methods of Collecting Data

The interview could be a personal interview, telephone interview, or self-administered questionnaire. Implementing a sample survey requires finding or constructing a valid population list for selecting the sample.

- **Personal interview**
  An interviewer asks prepared questions and records the subject's responses.
  - Advantage: subjects are more likely to participate.
  - Disadvantage: cost.
- **Telephone interview**
  like a personal interview but conducted over the phone.
  - Advantage: lower cost.
  - Disadvantage: the interview might have to be short. Every subject in the population must have a telephone.
- **Self-administered questionnaire**
  subjects are requested to fill out a questionnaire mailed to them.
  - Advantage: cheaper.
  - Disadvantage: many subjects may fail to participate.

1.3.5 Sources of Potential Bias in Sample Surveys

Sampling Bias

Sampling bias occurs from using nonrandom samples or having undercoverage (incomplete sampling frame).

Non-Response Bias

Nonresponse bias occurs when some sampled subjects cannot be reached or refuse to participate or fail to answer some questions.

Response Bias

Response bias occurs when the subject gives an incorrect response (perhaps lying) or the way the interviewer asks the questions (or wording of a question in print) is confusing or misleading.

1.3.6 Observational and Experimental Data

When we are interested in studying the association between two variables, we can gather the data in two different ways.

Experiment Study

A researcher conducts an experiment by assigning subjects to certain experimental conditions (treatment, explanatory variable) and then observing outcomes on the response variable.

Observational Study

In an observational study, the researcher observes values of the response variable and explanatory variables for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).

Summary

- In short, an observational study merely observes rather than experiments with the study subjects.
- An experimental study assigns to each subject a treatment and then observes the outcome on the response variable.

1.4 Sampling Distribution

1.4.1 Statistic vs Parameter

- **Statistic**
  A statistic is a numerical summary of sample data such as a sample proportion or sample mean.
- **Parameter**
  A parameter is a numerical summary of a population such as a population proportion or population mean.

In Practice

In practice, we seldom know the values of parameters. Inference on population parameters is made by using sample data (i.e. statistics like sample proportions, sample means etc).

Remark

In what follows, we will consider only simple random samples, i.e. the observations are extracted from the population by using a method similar to that of extracting a ball from an urn with replacement.

1.4.2 Sampling Distributions: Probability Distributions for Statistics

The value of any particular statistic varies according to the sample chosen. With randomization, a probability distribution for the possible values of a statistic enables us to evaluate how far the actual sample outcome of that statistic is likely to fall from the population parameter that it estimates.

Sampling Distribution: Definition

- The sampling distribution of a statistic is the probability distribution that specifies probabilities for the possible values the statistic can take.
- The sampling distribution of a statistic describes all potential values that could be observed for the statistic when taking a random sample of size n from the population.

Summary: Population Distribution, Data Distribution, Sampling Distribution

- **Population Distribution**
  This is the distribution that describes the possible values of the variable of interest in the population. It depends on unknown parameters, such as the population proportion p for a binary variable. By taking a sample from the population, we hope to learn and make predictions about the unknown population parameter(s).
- **Data Distribution**
  This is the distribution of the variable of interest based on the observations in the sample (the data). Its properties are described by statistics such as the sample proportion or the sample mean. With random sampling, the larger the sample size n, the more closely the data distribution resembles the population distribution.

- **Sampling Distribution**
  This is the distribution of a statistic computed from observations in the sample, such as the sample proportion or the sample mean. Because the numerical value of a statistic varies from sample to sample, we get an entire distribution of possible values for it, which we call the sampling distribution.

1.4.3 How Sample Proportions Vary Around the Population Proportion

We can describe the sampling distribution of a sample proportion by focusing on the key features of shape center, and variability. We typically use the mean to describe center and the standard deviation to describe variability.

Mean and Standard Deviation of the Sampling Distribution of the Sample Proportion

For a random sample of size n from a population with proportion p of outcomes in a particular category, the sampling distribution of the sample proportion in that category has:

$$\text{Mean } = \mu = p \tag{1.2}$$

$$\text{Standard deviation} = \sigma = \sqrt{\frac{p(1-p)}{n}} \tag{1.3}$$

$$\tag{1.4}$$

Remarks

- It is important to note that the distribution is centered around the value of the population proportion and its variability decreases when sample size increases.
- If n is sufficiently large so that the expected numbers of outcomes of the two types:
  - np in the category of interest
  - $n(1-p)$ not in that category

  are both at least 15, then the sampling distribution of a sample proportion is approximately normal.

1.4.4 How Sample Means Vary Around the Population Mean

As with the sampling distribution of the sample proportion, there are two main results about the sampling distribution of the sample mean:

- One result provides formulas for the mean and standard deviation of the sampling distribution.
- The other indicates that for large n its shape is approximately a normal distribution, as we observed in the previous section for the sample proportion.

Now, the variable of interest is a quantitative variable (for which we want to find the mean), not a binary one (for which we found a proportion).

The Sampling Distribution of the Sample Mean

- The Sample Mean $\bar{X}$ is a random variable that varies from sample to sample.
- By contrast, the population mean, μ, is a single fixed number.

- The center of its distribution is the population mean

$$\text{mean} = \mu \tag{1.5}$$

- the standard deviation equals the population standard deviation divided by the square root of the sample size:

$$\text{standard deviation} = \frac{\sigma}{\sqrt{n}} \tag{1.6}$$

- Even in this case the standard error (se) coincides with the standard deviation. It is important to note that when n increases the se decreases.
- Its distribution is normal if the distribution of the population is normal.

Mean and Standard Deviation of the Sampling Distribution of the Sample Mean

To find the mean and variance of the sampling distribution of the sample mean $\bar{X}$, we use the properties of Expectation and Variance:

$$E\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} E\left(Y_i\right) = n\mu, \quad \text{var}\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} \text{var}\left(Y_i\right) = n\sigma^2. \tag{1.7}$$

Where the n random variables are said to be independent and identically distributed, abbreviated by iid because with independent observations from the same distribution, such as with simple random sampling or a randomized experiment, all $\{\mu_i\}$ are identical and all $\left\{\sigma_i^2\right\}$ are identical, that is:

$$\{\mu_i = \mu\} \text{ and } \left\{\sigma_i^2 = \sigma\right\} \tag{1.8}$$

Then, the mean and variance of the sampling distribution of $\bar{X}$ are:

$$\begin{aligned} E(\bar{X}) &= E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}(n\mu) = \mu, \\ \text{var}(\bar{X}) &= \text{var}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{1}{n^2}\left[\text{var}\left(\sum_{i=1}^{n} X_i\right)\right] = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}. \end{aligned} \tag{1.9}$$

Standard Error

The spread of a sampling distribution is described by its standard deviation, which is called the standard error. The standard deviation of the sampling distribution of $\bar{Y}$ is called the standard error of $\bar{X}$. We denote it by $\sigma_{\bar{X}}$.

$$se = \sigma_{\bar{X}} = \sqrt{\text{var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \tag{1.10}$$

The standard error of X describes how much x tends to vary from sample to sample of size n.

1.4.5 Law of Large numbers and Central Limit Theorem

Because the standard error $\sigma = \frac{\sigma}{\sqrt{n}}$ decreases as n increases, $\bar{X}$ tends to be closer to μ as n increases. The result that $\bar{X}$ converges to μ as n grows indefinitely is called the law of large numbers.

Law of Large numbers

With n independent observations (such as in a randomized study) from a distribution with mean μ:

$$\text{for any } \epsilon > 0, P(|\bar{X} - \mu| \geq \epsilon) \to 0 \text{ as } n \to \infty \tag{1.11}$$

$$\text{for any } \epsilon > 0, P(|\bar{X} - \mu| < \epsilon) \to 1 \text{ as } n \to \infty \tag{1.12}$$

This type of convergence involves a sequence of random variables instead of fixed values as in ordinary sequences. To reflect the randomness, the convergence is called convergence in probability of $\bar{X}$ to μ, denoted by:

$$\bar{X} \xrightarrow{p} \mu \tag{1.13}$$

Normal, Binomial, and Poisson Sums of Random Variables Have the Same Distribution

For independent observations from some families of probability distributions, the sampling distribution of $\sum_i Y_i$ falls in the same family:

- **Normal Random Variables**
  If Y_i are independent normal random variables, with $Y_i \sim N(\mu_i, \sigma_i^2)$, then:

$$\sum_{i=1}^{n} Y_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \tag{1.14}$$

- **Binomial Random Variables**
  If Y_i are independent binomial random variables, with $Y_i \sim Binom(n_i, \pi)$, then:

$$\sum_{i=1}^{n} Y_i \sim \text{binom}\left(\sum_{i=1}^{n} n_i, \pi\right) \tag{1.15}$$

- **Poisson Random Variables**
  If Y_i are independent Poisson random variables, with $Y_i \sim Pois(\mu_i)$, then:

$$\sum_{i=1}^{n} Y_i \sim \text{Pois}\left(\sum_{i=1}^{n} \mu_i\right) \tag{1.16}$$

Definition: Central Limit Theorem (CLT)

For a large number n of independent observations from a population distribution having mean μ and finite standard deviation σ, the sampling distribution of the sample mean X is approximately a normal distribution.

Implications and interpretations of CLT

Here are some implications and interpretations of this remarkable result:

- The approximate normality of the sampling distribution applies no matter what the shape of the population distribution from which the samples are taken
- The sampling distribution of the sample mean takes more of a bell shape as the random sample size n increases.
- How large n must be before the sampling distribution is approximately normal depends on the skewness of the population distribution. If the population distribution is bell- shaped, then the sampling distribution is bell-shaped for all n. Unless the skewness is very severe, a sample size of about 30 is sufficient

- If the population distribution is approximately normal, then the sampling distribution is approximately normal for all sample sizes

CLT and Standardization

For n independent observations from a population with mean μ and finite standard deviation σ, the standardized version of the sample mean (i.e., subtracting its expected value and dividing by its standard error):

$$\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \tag{1.17}$$

has cdf that converges to the cdf of a standard normal distribution.
The convergence of a sequence of cdf 's to a limiting cdf is called convergence in distribution, in this case:

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1) \tag{1.18}$$

that is equivalent to stating that $\bar{X}$ has an approximate normal distribution with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$.

Chapter 2

Descriptive Statistics

2.1 Graphical Summaries of Data

Looking at a graph often gives you more of a feel for a variable and its distribution than looking at the raw data or a frequency table.

2.1.1 Graphs for Categorical Data

The two primary graphical displays for summarizing a categorical variable are the:

- **Pie Chart**
- **Bar Chart**

Pie Chart

- A pie chart is a circle having a slice of the pie for each category.
- The size of a slice corresponds to the percentage of observations in the category.
- It is easier to compare one category with the whole.
- It may not be suitable for too many categories.

Bar Chart

- bar graphs display a vertical bar for each category
- the height of the bar for each category is equal to the frequency (number of observations) or percentages (relative frequencies) in the category
- It is easier to compare categories

Pareto Chart

- Bar graphs are called Pareto Charts when the categories are ordered by their frequency, from the tallest bar to the shortest bar
- A bar graph with categories ordered by their frequency is called a Pareto chart, named after Italian economist Vilfredo Pareto (1848–1923), who advocated its use.
- The Pareto chart is often used in business applications to identify the most common outcomes, such as identifying products with the highest sales or identifying the most common types of complaints that a customer service center receives.

2.1.2 Graphs for Quantitative Data

The two primary graphical displays for summarizing a quantitative variable are the:

- **Dot Plots**
- **Histograms**

Dot Plot

- Shows a dot for each subject (observation) placed above its value on a number line.
- It is useful to show the relative positions of the data.
- To construct a dot plot:
  1. Draw a horizontal line and label it with the name of the variable
  2. Mark regular values of the variable on it.
  3. For each observation, place a dot above its value on the number line.

Histograms

- A histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

- Histograms are sometimes used for discrete quantitative variables drawing a separate bar for each possible value.

- Steps for constructing a histogram:
  1. Divide the range of the data into intervals of equal width. For a discrete variable with few values, use the actual possible values.
  2. Count the number of observations in each interval, creating a frequency table
  3. On the horizontal axis, label the values or the endpoints of the intervals.
  4. Draw a bar over each value or interval with height equal to its frequency (or proportion or percentage), values of which are marked on the vertical axis.
  5. Label and title appropriately

Choosing a Graph Type

We suggest always plotting the histogram to get an idea about the distribution. If the number of observations is small (e.g., less than 50), supplement the histogram with a dot plot to show the numerical values of the observations.

2.1.3 Shape of Distribution

A histogram (or dot plot) visualizes the distribution of a quantitative variable. Here are important questions to ask when describing this distribution:

- **Unimodality and Bimodality:** Does the distribution have a single mound or peak?
- **Symmetry and Skewness:** What is the shape of the distribution?
- **Outliers:** Do the data cluster together, or is there a gap such that one or more observations noticeably deviate from the rest, as in the histogram in the margin?

Unimodality and Bimodality

- **Unimodality:** A distribution of data with a single mound is called unimodal. The highest point is at the mode.
- **Bimodality:** A distribution with two distinct mounds is called bimodal.

Symmetry and Skewness

The shape of a unimodal distribution is often described as symmetric or skewed.

- **Symmetric:** A distribution is symmetric if the side of the distribution below a central value is a mirror image of the side above that central value.
- **Skewed:** A distribution is skewed if one side of the distribution stretches out longer than the other side.

Skewed Distribution

To skew means to stretch in one direction:

- **A left-skewed distribution stretches to the left**
  A distribution is skewed to the left if the left tail is longer than the right tail.
- **A right-skewed distribution stretches to the right**
  A distribution is skewed to the right if the right tail is longer than the left tail.

Outliers

An outlier is an observation that falls well above or well below the overall bulk of the data.

2.1.4 Time Plots: Displaying Data over Time

Time-Series

For some variables, observations are recorded over time. A data set collected over time is called a time series.

Time Plot

We can display time-series data graphically using a time plot. Plots each observation on the vertical scale against the time measured on the horizontal scale.

Trends

Points are usually connected. Common patterns in the data over time, known as trends, should be noted. So the trend indicates a tendency of the data to either rise or fall.

2.2 Measuring the Center of Quantitative Data

It's a good idea to look first at a variable with the appropriate graph to get a feel for the data. You can then consider computing numerical summaries (statistics) for the data to describe several other important features of a distribution. For instance, when describing a quantitative variable, we would like to know:

- **Center:** What is a typical value for the observations?
- **Variability (Spread):** Do the observations take similar values, or do they vary quite a bit?

There are several ways to compute the center of a distribution:

1. **Mean**
2. **Median**
3. **Mode**

2.2.1 Calculating the Mean

Definition: Mean

The mean is the sum of the observations divided by the number of observations. It is interpreted as the balance point of the distribution.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_n}{n} \tag{2.1}$$

The mean refers to averaging, that is, adding up the data points and dividing by how many there are.

Properties

There are some basic properties of the mean:

- **It is the balance point of data**
  If we were to place identical weights on a line representing where the observations occur, then the line would balance by placing a fulcrum at the mean.
- **Skewed Distribution and Mean**
  For a skewed distribution, the mean is pulled in the direction of the longer tail, relative to the median. The next example illustrates this idea.
- **Mean and Outlier**
  The mean can be highly influenced by an outlier, which is an unusually small or unusually large observation.
- **It is internal**
  The mean is always in between the minimum and the maximum
- **It is the fair value**
  Substituting the value of each observation with the mean we preserve the total amount.
- **Sum of deviations is zero**
  $(x_i - \bar{x}) = 0$
- **It is Linear**
  Let be consider:
  - $\bar{x}$ is the mean of $x_1, x_2, ..., x_n$
  - $\bar{y}$ is the mean of $y_1, y_2, ..., y_n$
  - linear equation: $y_i = a + bx_i$

  Then:

$$\bar{y} = a + b\bar{x} \tag{2.2}$$

2.2.2 Calculating the Median

Definition: Median

The median is the middle value of the observations when the observations are ordered from the smallest to the largest (or from the largest to the smallest).

How to Determine the Median

Steps to finding the median for a set of data:

1. Put the n observations in order of their size.
2. Find the location of median in the ordered data by:

$$\frac{(n+1)}{2} \tag{2.3}$$

3. Two ways:
   (a) **Number Observation is odd → Median is the middle observation**
   When the number of observations n is odd, the median is the middle observation in the ordered sample.
   (b) **Number Observation is even → median is the average of the two middle observations**
   When the number of observations n is even, two observations from the ordered sample fall in the middle, and the median is their average.

The median is the point that splits the data in two, half the data below it and half above it (much like the median on a highway splits a road into two equal parts).

Why is the median not affected by an outlier?

How far an outlier falls from the middle of the distribution does not influence the median. The median is determined solely by having an equal number of observations above it and below it.

2.2.3 Calculating the Mode

Definition: Mode

The mode is the value that occurs most frequently. It describes a typical observation in terms of the most common outcome.

Properties

- The concept of the mode is most often used to describe the category of a categorical variable that has the highest frequency (the modal category).
- With quantitative variables, the mode is most useful with discrete variables taking a small number of possible values.
- Not affected by extreme values
- There may be no mode or there may be several modes

2.2.4 Comparing the Mean and Median

Mean, median and mode are usually not equal

The shape of a distribution influences whether the mean is larger or smaller than the median. Generally, if the shape is:

- perfectly symmetric, the mean equals the median.
- skewed to the left, the mean is smaller than the median.
- skewed to the right, the mean is larger than the median.

So:

- **Symmetric Distribution**
  When a distribution is close to symmetric, the tails will be of similar length, and therefore the median and mean are similar.
- **Skewed Distribution**
  For skewed distributions, the mean lies toward the direction of skew (the longer tail) relative to the median. This is because extreme observations in a tail affect the balance point for the distribution, which is the mean.

Resistant Meausures

A numerical summary measure is resistant if extreme observations (outliers) have little, if any, influence on its value. So:

- **Mean is affected by extreme values. It is not resistant to outlier**
- **Median is resistant to outliers, in essence it is not affected by extreme values**

2.3 Measuring the Variability of Quantitative Data

To describe quantitative data more fully, we describe not only the center but also the variability about that center. There are several ways to compute the variability of a distribution:

1. **Range**
2. **Standard Deviation**

2.3.1 The Range

One way to measure the spread is to calculate the range.

Definition: Range

The range is the difference between the largest and smallest values in the data set:

$$\text{Range} = \text{Max} - \text{Min} \tag{2.4}$$

Remarks

- The range is simple to compute and easy to understand, but it uses only the extreme values and ignores the other values. Therefore, it's affected severely by outliers.
- The range is not a resistant statistic. It shares the worst property of the mean, not being resistant, and the worst property of the median, ignoring the numerical values of nearly all the data.

2.3.2 The Standard Deviation

More useful measures of variability use the distances of all the observations from the center, that is the deviation from the mean.

Definition: Deviation

The deviation of an observation x_i from the mean $\bar{x}$ is the difference between the observation and the sample mean.

$$x_i - \bar{x} \tag{2.5}$$

Properties of Deviation

- A deviation is positive if it falls above the mean and negative if it falls below the mean
- The sum of the deviations is always zero, in essence the mean is the center

Definition: Squared Deviations

The interpretation of the mean as the balance point implies that the positive deviations counterbalance the negative deviations. Because of this, the sum (and therefore the mean) of the deviations always equals zero, regardless of the actual data values. Hence, summary measures of variability from the mean use the squared deviations:

$$(x_i - \bar{x})^2 \tag{2.6}$$

Definition: Variance

A measure of variation can be obtained by summarizing the squared deviations of each observation from the mean and the average of the squared deviations is called the variance:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.7}$$

Definition: Standard Deviation

Because the variance uses the square of the units of measurement for the original data, its square root is easier to interpret: this is called Standard Deviation. The original unit of measurement can be recovered by computing:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.8}$$

Sample Variance and Sample Standard Deviation

When the dataset is a sample, the two previous indices are adjusted dividing the sum of the squared deviations by $n-1$ instead of n. In formulas:

- **Sample Variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.9}$$

- **Sample Standard Deviation**

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{\text{sum of squared deviations}}{\text{sample size } -1}} \tag{2.10}$$

Remark

σ^2 and σ are also called population variance and population standard deviation, respectively.

Properties of the Standard Deviation (σ, s)

- The Standard Deviation measures the spread (variability) of the data.
- $\sigma[s] = 0$ only when all observations have the same value, otherwise $\sigma[s] > 0$
- $\sigma[s]$ has the same units of measurement as the original observations. The variance $\sigma^2[s^2]$ has units that are squared.
- $\sigma[s]$ is not resistant (can be influenced by outliers). A few extreme values can greatly increase its value. Also, outliers have large deviations, so they tend to have extremely large squared deviations.

2.4 Using Measures of Position to Describe Variability

We'll now learn about some other ways of describing a distribution using measures of position.

- **Percentage**
  One type of measure of position tells us the point where a certain percentage of the data fall above or fall below that point.
- **Z-Score**
  Another type of measure of position, called a z-score, tells us how far an observation falls from a particular point, such as the number of standard deviations an observation falls from the mean.

2.4.1 Measures of Position: The Quartiles and Percentiles

Definition: Percentiles

The pth percentile ($p\%$) is a value such that p percent of the observations fall at or below that value.

Median as Percentile

The median is a special case of a more general set of measures of position called percentiles. The 50th percentile is usually referred to as the median.

Definition: Quartile

The quartiles split the distribution into four parts, each containing one quarter (25%) of the observations

1. **First Quartile** $p = 25$
   25% is below the first quartile (Q1)
2. **Second Quartile** $p = 50$
   25% is between the first quartile and the second quartile (the median, Q2)
3. **Third Quartile** $p = 75$
   25% is between the second quartile and the third quartile (Q3)

Finding Quartiles

1. Arrange the data in order
2. Consider the median. This is the second quartile, Q2
3. Consider the lower half of the observations (excluding the median itself if n is odd)
4. The median of these observations is the first quartile, Q1
5. Consider the upper half of the observations (excluding the median itself if n is odd)
6. Their median is the third quartile, Q3

2.4.2 Measuring Variability: The Interquartile Range

The quartiles are also used to define a measure of variability that is more resistant than the range and the standard deviation. This measure summarizes the range for the middle half of the data.

Definition: Interquartile Range (IQR)

The interquartile range is the distance between the third and first quartiles:

$$IQR = Q_3 - Q_1 = \text{upper quartile} - \text{lower quartile} \tag{2.11}$$

$$= 75\text{th percentile} - 25\text{th percentile} \tag{2.12}$$

Remarks

- the IQR is not affected by any observations below the first quartile or above the third quartile. In other words, it is not affected by outliers.
- the range depends solely on the minimum and the maximum values, the most extreme values, so the range changes as either extreme value changes.

2.4.3 Detecting Potential Outliers

Examining the data for unusual observations, such as outliers, is important in any statistical analysis. One way is to use the interquartile range.

The 1.5 X IQR Criterion for Identifying Potential Outliers

An observation is a potential outlier if it falls a distance of more than $1.5 * IQR$ below the first quartile or a distance of more than $1.5 * IQR$ above the third quartile.

$$\text{Potential Outlier} \implies 1.5 * IQR < Q_1 \vee 1.5 * IQR > Q_3 \tag{2.13}$$

Why do we identify an observation as a potential outlier rather than calling it a definite outlier?

- When a distribution has a long tail, some observations may be more than 1.5 * IQR below the first quartile or above the third quartile even if they are not outliers, in the sense that they are not separated far from the bulk of the data.
- For instance, for a distribution with a long right tail, the largest obser- vation may not be that far away from the remaining observations, and there isn't enough of a gap to call it an outlier.

2.4.4 The Box Plot: Graphing a Five-Number Summary of Positions

The quartiles and the maximum and minimum values are five numbers often used as a set to summarize positions that help describe center and variability of a distribution.

The Five-Number Summary of Positions

The five-number summary of a data set consists of the minimum value, first quartile Q1, median, third quartile Q3, and the maximum value.

The Box Plot

The five-number summary is the basis of a graphical display called the box plot and consists of:

1. **Minimum value**
2. **First Quartile**
3. **Median**
4. **Third Quartile**
5. **Maximum value**

The box of a box plot indicates:

- the central 50% of the distribution, from the first quartile to the third quartile
- A line inside the box marks the median.
- The lines extending from each side of the box are called whiskers. These extend to show the stretch of the rest of the data, except for potential outliers

Constructing a Box Plot

1. A box goes from the lower quartile Q1 to the upper quartile Q3.
2. A line is drawn inside the box at the median.
3. Line goes from the lower end of the box to the smallest observation that is not a potential outlier. A separate line goes from the upper end of the box to the largest observation that is not a potential outlier. These lines are called whiskers.
4. The potential outliers (more than 1.5 IQR below the first quartile or above the third quartile) are shown separately with special symbols (such as a dot or a star).

The Box Plot Compared with the Histogram

- A box plot does not portray certain features of a distribution, such as distinct mounds and possible gaps, as clearly as a histogram.
- A box plot does indicate skew from the relative lengths of the whiskers and the two parts of the box. The side with the larger part of the box and the longer whisker usually has skew in that direction.
- the box plot will not show us whether there is a large gap in the distribution contributing to the skew.
- box plots are useful for identifying potential outliers.

Comparing Distribution

- Quantitative data can be broken down by levels of a categorical variable.
- The side-by-side box-plot produces an excellent visual comparison for shape, outliers, variability, etc.
- The box-plot is less accurate than the histogram but very useful in comparing several distributions.

2.4.5 The Z-Score

Z-Score for an Observation

Z-value or Z-score or simply Z, represents the number of standard deviations that it falls from the mean of all observations. The z-score for a particular observation is calculated as:

$$z_i = \frac{x_i - \bar{x}}{s} = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} \tag{2.14}$$

- Positive z-score indicates the observation is above the mean
- Negative z-score indicates the observation is below the mean

Properties of Z-Score

- The z-score allows us to tell quickly how surprising or extreme an observation is.
- The z-score converts an observation (regardless of the observation's unit of measurement) to a common scale of measurement, which allows comparisons.
- The z-score identifies position and potential outliers.

2.5 Concept of Association/Relationship

2.5.1 Response and Explanatory Variables

Response Variable (Dependent Variable)

The response variable is the outcome variable on which comparisons are made for different values of the explanatory variable.

Explanatory Variable (Indipendent Variable)

- When the explanatory variable is categorical, it defines the groups to be compared with respect to the response variable.
- When the explanatory variable is quantitative, we examine how different values of the explanatory variable relate to changes in the response variable

Three Cases

In practice, when we investigate the association between two variables, there are three types of cases:

1. **The variables could be categorical**
   The variables could be categorical as food type and pesticide status are. In this case, as we have already seen, the data are displayed in a contingency table, and we can explore the association by comparing conditional proportions.
2. **One variable could be quantitative and the other could be categorical**
   One variable could be quantitative and the other could be categorical such as analyzing height and gender or income and race. We can compare the categories (such as females and males) using summaries of center and variability for the quantitative variable (such as the mean and standard deviation of height) and graphs such as side-by-side box plots.
3. **Both variables could be quantitative**
   Both variables could be quantitative. In this case, we analyze how the outcome on the response variable tends to change as the value of the explanatory variable changes.

Important Remark

We use the term association to study how two categorical variables relate, and the term relationship to study how two quantitative variables relate.

2.5.2 Association Between Two Variables

The main purpose of a data analysis with two variables is to investigate whether there is an association and to describe the nature of that association.

Definition: Association

An association (dependence) exists between two variables if a particular value/category for one variable is more likely to occur with certain values/categories of the other variable

Remark

When there's an association, the likelihood of particular values for one variable depends on the values of the other variable.

2.6 The Association Between Two Categorical Variables

2.6.1 Contingency Tables

Definition

A contingency table is a display for two categorical variables.

Characteristic

- Its rows list the categories of one variable
- Its columns list the categories of the other variable
- Each entry in the table, called a cell, gives the count for a particular pair of categories (frequencies)
- The response variable is typically used as the column variable, and the explanatory variable is used to form the rows of the contingency table.

Example: bivariate distribution of Gender by Occupational status

Raw data

Code	Gen	OS
1	M	YES
2	F	YES
3	M	NO
4	F	NO
5	M	YES
6	M	YES
7	M	YES
8	F	NO
9	M	YES
10	F	NO

Contingency table

	OS		
Gen	**Yes**	**No**	**Tot.**
M	5	1	6
F	1	3	4
Tot.	6	4	10

- It displays two categorical variables
- The rows list the categories of one variable
- The columns list the categories of the other variable
- Entries in the table are frequencies

2.6.2 Calculate Conditional Proportions or Percentages

- **Association and Conditional Proportion**
  If OS is the response then we study how its distribution changes when Gen varies. In order to compare different distributions of the same variable, we compute the relative frequencies (or percentages).

	OS		
Gen	**Yes**	**No**	**Tot.**
M	500	100	600
	0.83	0.17	1
F	10	30	40
	0.25	0.75	1
Tot.	510	130	640
	0.80	0.20	1

- 0.83 = proportion of YES under the condition Gen=M (conditional prop.)
- 0.25 = proportion of YES under the condition Gen=F (conditional prop.)
- 0.80 = proportion of YES (marginal proportion)

- **No Association and Conditional Proportion**
  If there is no association between OS and Gen, then the conditional proportions for the response variable categories (OS) would be the same for each gender. In this case the two variables are said to be independent.

	OS		
Gen	**Yes**	**No**	**Tot.**
M	500	100	600
	0.83	0.17	1
F	50	10	60
	0.83	0.17	1
Tot.	550	110	660
	0.83	0.17	1

- In general, two variables are **independent** if the conditional percentage distributions of the response (given the categories of the explanatory) are **equal**.

2.7 The Relationship Between Two Quantitative Variables

2.7.1 Looking for a Trend: The Scatterplot

Definition

A scatterplot is a graphical display for the relationship between two quantitative variables.

Constructing Scatterplot

- **Horizontal Axis: Explanatory variable, x:** using the horizontal (x) axis for the explanatory variable x
- **Vertical Axis: Response variable, y:** using the vertical (y) axis for the response variable y.

Observation

Each point (x, y) in the scatterplot represents an observation with value x for the explanatory and value y for the response variable. With a total of n observations, the scatterplot has n points.

2.7.2 How to Examine a Scatterplot: Positive, Negative and Linear Relationship

We examine a scatterplot to study the relationship between two quantitative variables. How do values on the response variable change as values of the explanatory variable change?

Positive Relationship and Negative Relationship

Two quantitative variables x and y are:

- **Positvely Associated: As x goes up, y tends to go up.**
  Two quantitative variables x and y are said to have a positive relationship when:
  - high values of x tend to occur with high values of y
  - low values of x tend to occur with low values of y.

- **Negatively Associated: As x goes up, y tends to go down.**
  Two quantitative variables x and y are said to have a positive relationship when high values of one variable tend to pair with low values of the other variable:
  - High values of x tend to occur with low values of y
  - Low values of x tend to occur with high values of y

Linear Relationship

A relationship between two quantitative variables is called linear if the scatterplot reveals a trend that is well approximated by a straight line.

2.7.3 The Correlation Coefficient

When the data points follow a roughly straight-line trend, the variables are said to have an approximately linear relationship. A summary measure called the correlation coefficient describes the strength of the linear relationship.

Concept of Correlation

- Correlation is a way to measure the positive or negative association between two quantitative variables.
- It Measures the strength and direction of the linear association between x and y.

Definition

The correlation coefficient summarizes the strength and direction of the linear (straight-line) relationship between two quantitative variables. Denoted by r, it takes values between -1 and +1:

$$-1 \leq r \leq +1 \tag{2.15}$$

where:

- A positive value for r indicates a positive relationship.
- A negative value for r indicates a negative relationship.
- The closer r is to ± 1 the closer the data points fall to a straight line, and the stronger the linear association is.
- The closer r is to 0, the weaker the linear association is.

Formula for the Correlation Coefficient

The Correlation Coefficient is given by the following expession:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2.16}$$

Another Formula with Z-score

Let z_x, z_y denote the z-score for, respectively, an observation x, y on the explanatory variable, that are:

$$z_x = \frac{\text{observed value } - \text{ mean of all } x}{\text{standard deviation of all } x} = \frac{(x - \bar{x})}{s_x} \tag{2.17}$$

$$z_y = \frac{\text{observed value } - \text{ mean of all } y}{\text{standard deviation of all } y} = \frac{(y - \bar{y})}{s_y} \tag{2.18}$$

The Correlation Coefficient is given by the following expession:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} z_x z_y = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right) \tag{2.19}$$

Notes that:

- The product of the z-scores for any point in the upper-right quadrant is positive. The product is also positive for each point in the lower-left quadrant. Such points make a positive contribution to the correlation coefficient.
- The product of the z-scores for any point in the upper-left and lower-right quadrants is negative. Such points make a negative contribution to the correlation coefficient.

Properties

- The correlation coefficient r always falls between -1 and $+1$. The closer the value to 1 in absolute value (see the margin comments), the stronger the linear (straight-line) relationship as the data points fall nearer to a straight line.
- A positive correlation coefficient indicates a positive relationship, and a negative correlation coefficient indicates a negative relationship.
- The correlation coefficient does not depend on which variable is treated as the response and which as the explanatory variable.
- The correlation coefficient does not depend on the variables' units. The correlation coefficient is not affected by a linear transformation of the response or explanatory variable.
- the correlation coefficient is only appropriate to calculate for quantitative variables which show an approximate linear relationship.
- Correlation only measures strength of linear relationship

2.8 Cautions in Analyzing Associations

2.8.1 Association Does Not Imply Causation

Just observing an association between two variables is not enough to imply a causal connection. There may be some alternative explanation for the association.

2.8.2 Extrapolation could be Dangerous

Extrapolation refers to using a regression line to predict y values for x values outside the observed range of data. This is riskier as we move farther from that range. There is no guarantee that the relationship given by the regression equation holds outside the range of observed x-values.

Chapter 3

Inferential Statistics

3.1 Point Estimates and Confidence Intervals

Population parameters have two types of estimates, a point estimate and an interval estimate.

3.1.1 Overview: Point Estimate and Confidence Intervals

Point Estimate

- A point estimate is a single number that is our best guess for the parameter.
- A point estimate does not tell us how close the estimate is likely to be to the parameter.
- A point estimate is the value of a statistic that predicts the parameter value.

Confidence Interval

- An interval estimate is an interval of numbers that is quite likely to contain the actual value of the parameter.
- An interval estimate is more useful. It incorporates a margin of error which helps us to gauge the accuracy of the point estimate.
- An interval estimate is an interval of numbers around the point estimate, called a confidence interval, within which the parameter value is believed to fall with a specified degree of confidence.

3.1.2 Properties of Point Estimates (or/and Estimator)

For any particular parameter, there are several possible point estimates. For a normal distribution, for instance, the center is the mean and the median since that distribution is symmetric.

So, two possible estimates of that center value are the sample mean and the sample median. A good estimator of a parameter has two desirable properties:

1. **Estimator is Unbiased**
2. **Small Standard Error**

Terminology: Estimates vs Estimator

- **Estimator is a Random Variable**
  The term estimator refers to a particular statistic used to construct a point estimate. For example, the sample proportion is an estimator of a population proportion.
- **Estimate is a Particular Number**
  The term estimate is short for its value (the point estimate) for a particular sample. For example, the value 0.81 is an estimate for the population proportion believing in life after death in the U.S.

We denote θ an arbitrary parameter and $\hat{\theta}$ an estimator of θ

Property 1: Estimator is Unbiased

A good estimator has a sampling distribution that is centered at the parameter, in essence the mean of the distribution equals the true value of the parameter. An estimator with this property is unbiased. Formally:

$$E(\hat{\theta}) = \theta \tag{3.1}$$

So:

- The sample mean is an unbiased estimator of the population mean
- The sample proportion is an unbiased estimator of the population proportion

Property 2: Small Standard Error

A good estimate has a small standard deviation compared to other estimates. This tells us the estimate tends to fall closer than other estimates to the parameter. Formally, an efficient estimator tends to fall closer to θ, on the average, than other estimators. A common way to summarize the closeness is by the mean squared error (MSE):

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{3.2}$$

An efficient estimator by this criterion has minimum mean squared error, uniformly for the possible values of θ. The mean squared error satisfies:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 = E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2[E(\hat{\theta}) - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 = \text{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2\end{aligned} \tag{3.3}$$

The bias of $\hat{\theta}$ is defined as $E(\hat{\theta}) - \theta$, so:

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + (\text{ bias })^2 \tag{3.4}$$

Of the unbiased estimators, the one with smallest MSE is a minimum variance unbiased estimator:

$$E(\hat{\theta}) = \theta, \quad \text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) \tag{3.5}$$

Remark: Standard Error

Beginning here, we'll use the term standard error to refer to the estimated standard deviation of the sampling distribution for a statistic. The standard error is the estimated standard deviation of a sampling distribution.

- Error of estimation: the difference between the statistic (estimate) and the true value of the parameter.
- Standard error (se):the standard absolute error of estimation of the estimator (square root of squared errors mean).
- It is important to note that when the mean of the estimator sampling distribution is equal to the value of the parameter, then the standard error coincides with the standard deviation. This is so because the deviations from the mean coincide with the errors of estimation.

3.1.3 Interval Estimator: Confidence Intervals for Parameters

Definition: Confidence Intervals

A confidence interval is an interval containing the most plausible values for a parameter. (so, for a parameter θ is an interval of numbers within which θ is predicted to fall).

The probability that this method produces an interval that contains the parameter θ is called the confidence level, which is a number chosen to be close to 100%, most commonly 95%.

It is constructed by a method that combines a point estimate with a margin of error.

Margin of Error

The margin of error estimates the precision of a point estimate for a given confidence level. It measures how close the point estimate is likely to fall to the population parameter.

For a sampling distribution that is approximately normal, the margin of error is a multiple of the standard error (se), such as ± 1.96(se) when the confidence level is 95%.

Logic of Confidence Intervals

To construct a confidence interval for a population proportion, start with the sampling distribution of a sample proportion:

1. gives the possible values for the sample proportion and their probabilities
2. is approximately a normal distribution for large random samples by the CLT
3. has mean equal to the population proportion

Approximately 95% of a normal distribution falls within 1.96 standard deviations of the mean:

- With probability 0.95, the sample proportion falls within about 1.96 standard errors of the population proportion
- The distance of 1.96 standard errors is the margin of error in calculating a 95% confidence interval for the population proportion

Summary

1. A confidence interval is constructed by taking a point estimate and adding and subtracting a margin of error.

$$\text{point estimate } \pm \text{ margin of error.} \tag{3.6}$$

2. The margin of error is based on the standard deviation of the sampling distribution of that estimator.
3. When the sampling distribution is approximately normal, a 95% confidence interval has a margin of error equal to 1.96 standard deviations.

$$\text{point estimate } \pm 1.96(\text{se}) = \hat{p} \pm 1.96(\text{ standard deviation }) \tag{3.7}$$

3.2 Constructing Confidence Intervals

3.2.1 Concept of Confidence Interval

Definition: Confidence Interval

In a study employing randomization, let $T_L(\mathbf{Y})$ and $T_U(\mathbf{Y})$ (respectively called lower and upper confidence limits) be two functions of observations $\mathbf{Y} = (Y_1, ..., Y_n)$ such that:

$$P\left[T_L(\mathbf{Y}) \leq \theta \leq T_U(\mathbf{Y})\right] = 1 - \alpha, \text{ for all possible values of } \theta. \tag{3.8}$$

Once the data $\mathbf{Y} = \mathbf{y}$ are observed, the interval:

$$[T_L(\mathbf{y}), T_U(\mathbf{y})] \tag{3.9}$$

is called a $100(1-\alpha)\%$ confidence interval for θ.

Confidence Level and Error Probability

- **Confidence Level:** The probability $1 - \alpha$, called the confidence level, is usually chosen close to 1, such as 0.95 or 0.99
- **Error Probability:** The corresponding probability α is an error probability.

Standard Normal Quantile

Let z_a denote the standard normal quantile for right-tail probability equal to a, such as: $z_{0.025} = 1.96$. and $z_{0.005} = 2.58$. Then, for confidence level $1 - \alpha$, the confidence interval uses quantiles with tail probabilities equal to $\frac{\alpha}{2}$ and has the form:

$$\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}} = \hat{\theta} \pm z_{\alpha/2}(\text{se}) \tag{3.10}$$

As the confidence level grows, α decreases and $z_{\frac{\alpha}{2}}$ increases, so that the confidence interval gets wider:

- As the confidence level increases, the confidence interval gets wider.
- As n grows, the standard error decreases, so the confidence interval gets narrower.

3.2.2 Confidence Interval to Estimate a Population Proportion

Assumptions

- We symbolize a population proportion by π
- The point estimate of the population proportion is the sample proportion
- We symbolize the sample proportion by $\hat{\pi}$
- A 95% confidence interval uses a margin of error = 1.96(standard errors)

Confidence Interval

The Confidence Interval is given by:

$$CI = [\text{ point estimate } \pm \text{ margin of error}] = \tag{3.11}$$

$$= \hat{\pi} \pm z_{\alpha/2}(\text{se}) = \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \tag{3.12}$$

and finding the 95% Confidence Interval for a Population Proportion ($\alpha = 5\% \quad z_{\alpha/2} = 1.96$):

$$\hat{\pi} \pm 1.96\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \tag{3.13}$$

Sample Size Needed for Large-Sample Confidence Interval for a Proportion

The sampling distribution of the sample proportion is well approximated by a normal if both np and $n(1-p)$ are greater than or equal to 15. In this case we do not know the true value of p but only its point estimate. It follows that the rule becomes:

For the 95% confidence interval for a proportion p to be valid, you should have at least 15 successes and 15 failures:

$$n\hat{p} \geq 15 \text{ and } n(1-\hat{p}) \geq 15 \tag{3.14}$$

Effects of Confidence Level and Sample Size

The margin of error for a confidence interval:

- Increases as the confidence level increases.
- Decreases as the sample size increases.

Interpretation of the Confidence Level

- If we used the 95% confidence interval method to estimate many times the same population proportion, then in the long run about 95% of those intervals would give correct results, containing the population proportion
- If we used the 95% confidence interval method to estimate many times the same population proportion, then in the long run about 5% of those intervals would give incorrect results, do not containing the population proportion

3.2.3 Confidence Interval to Estimate a Population Mean

Assumptions

- The sample mean is the point estimate of the population mean

$$\text{Point Estimator: } \hat{\mu} = \bar{X} \Rightarrow \text{ Point Estimate: } \bar{x} = \mu \tag{3.15}$$

- The exact standard error of the sample mean is:

$$se = \frac{\sigma}{\sqrt{n}} \tag{3.16}$$

  but in practice, we do not know the population standard deviation σ, we estimate it by the sample standard deviation, s:

$$se = \frac{s}{\sqrt{n}} \tag{3.17}$$

Now, there are two ways:

- The number of population is small
- The number of population is large

Recall: T-Distribution

- The t distribution is bell shaped and symmetric about 0 (mean).
- Its standard deviation is a bit larger than 1, the precise value depending on what is called the degrees of freedom, denoted by df. The probabilities depend on the degrees of freedom:

$$df = n - 1 \tag{3.18}$$

- The t distribution has thicker tails and therefore allows for more variability (spread out) than the standard normal distribution. The larger the df value, however, the closer it gets to the standard normal.
- A t-score obtained from the t-distribution multiplied by the standard error gives the margin of error for a confidence interval for the mean μ

C.I. for a Population Mean when n is small and the population is normal

For small n from any population and also for every n from an underlying population that is normal, then the standardization of population mean is a T-distribution:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim T_{n-1} \tag{3.19}$$

then the Confidence Interval for the population mean is given by:

$$CI = [\text{ point estimate } \pm \text{ margin of error}] = \tag{3.20}$$

$$= \bar{x} \pm z_{\alpha/2}(\text{se}) = \bar{x} \pm z_{df,\alpha/2}(\frac{s}{\sqrt{n}}) \tag{3.21}$$

and finding the 95% Confidence Interval for a Population Mean ($\alpha = 5\%$):

$$\bar{x} \pm z_{df,0.025}(\frac{s}{\sqrt{n}}) \tag{3.22}$$

C.I. for a Population Mean when n is large

For large n from any population and also for every n from an underlying population that is normal, then the population mean is a Normal Distribution by the Central Limit Theorem:

$$\bar{X} \sim N\left(\mu, s^2/n\right) \tag{3.23}$$

then the Confidence Interval for the population mean is given by:

$$CI = [\text{ point estimate } \pm \text{ margin of error}] = \tag{3.24}$$

$$= \bar{x} \pm z_{\alpha/2}(\text{se}) = \bar{x} \pm z_{\alpha/2}(\frac{s}{\sqrt{n}}) \tag{3.25}$$

and finding the 95% Confidence Interval for a Population Mean ($\alpha = 5\% \quad z_{\alpha/2} = 1.96$):

$$\bar{x} \pm 1.96(\frac{s}{\sqrt{n}}) \tag{3.26}$$

3.3 Significance Test

3.3.1 What is Significance Test?

Definition: Hypothesis

In statistics, a hypothesis is a statement about a population, usually claiming that a population parameter takes a particular numerical value or falls in a certain range of values.

Definition: Significance Test

A significance test is a method of using data to summarize the evidence about a hypothesis. In other terms, it is a method to check whether or not the data support certain statements or predictions. These statements are hypotheses about population. A significance test about a hypothesis has five steps:

1. **Assumptions**
2. **Hypotheses**
3. **Test Statistic**
4. **P-Value**
5. **Conclusion**

3.3.2 The Steps of a Significance Test

Assumptions

Like other statistical methods, each significance test makes certain assumptions or has certain conditions to be valid. These pertain to:

- **Type of Data:** Each significance test applies for either quantitative or categorical data.
- **Randomization:** Each significance test assumes that the data gathering employed randomization, such as a simple random sample or a randomized experiment.
- **Population Distribution:** ome significance tests assume that the variable has population distribution in a particular family of probability distributions, such as the normal or binomial.

- **Sample Size:** Some significance tests, such as those based on the Central Limit Theorem, are valid only when the sample size is sufficiently large.

Each significance test makes certain assumptions about some characteristics of the distribution of the population and the sampling distribution of a statistic.
It implies that the sampling distribution of sample mean is normal as well (in real applications, it would be more realistic to assume the normality of the population on the log scale).

Hypothesis

Significance tests have two hypotheses about population distributions, usually expressed in terms of parameters of those distributions:

- **Null Hypothesis**
  The null hypothesis, denoted by H_0, is a statement that a parameter takes a particular value. The value in the null hypothesis usually represents no effect. It has a single parameter value.
- **Alternative Hypothesis**
  The alternative hypothesis, denoted by H_a or H_1, states that a parameter takes value in some alternative range. The value in the alternative hypothesis then represents an effect in the positive, in the negative, or in both the positive and negative direction. It specifies how the null can be false.

In a significance test, the null hypothesis is presumed to be true unless the data give strong evidence against it. Formally, we have to decide, on the basis of the information contained in the sample, if H_0 have to be rejected or not.

Test Statistic

In order to decide if the null should be rejected or not, we summarize the evidence against the null by computing a (test) statistic.

The population parameter to which the null and alternative hypotheses refer has a point estimate. A test statistic describes how far that point estimate falls from the parameter value given in the null hypothesis. Usually this distance is measured by the number of standard errors between the point estimate and the parameter.

- The test statistic is zero when the estimate is equal to the value specified in the null.
- If the test statistic falls far from zero in the direction specified by the alternative hypothesis, it is evidence against the null hypothesis and in favor of the alternative hypothesis.

P-Value

To interpret a test statistic value, we use a probability summary of the evidence against the null hypothesis, H_0.

1. First, we presume that H_0 is true
2. Next, we consider the sampling distribution of the test statistic

The P-value is the probability that the test statistic equals the observed value or a value even more extreme. It is calculated by presuming that the null hypothesis H_0 is true.

In essence, we summarize how far out in the tail of this sampling distribution the test statistic falls by computing, when H_0 is true, the probability of that value and values even more extreme.

The smaller the p-value, the stronger the evidence the data provide against the null hypothesis. That is, a small p-value indicates a small likelihood of observing the sampled results if the null hypothesis were true.

Conclusion

Report and interpret the p-value in the context of the study. Based on the p-value, make a decision about H0 (either reject or do not reject H_0 if the p-value is small or not). Before seeing the data, we decide how small the p-value would need to be to reject H0. This cutoff point is called the significance level.

In practice, the most common significance level is 0.05 (or 0.01 if the sample size is very large). When we reject H0 we say the results are statistically significant.

3.3.3 Significance Test About Proportions

Assumptions

- The variable is categorical in 0,1 coding
- The data are obtained using randomization (such as a random sample or a randomized experiment).
- The sample size is sufficiently large that the sampling distribution of the sample proportion $\hat{\pi}$ is approximately normal. The approximate normality happens when the expected numbers of successes and failures are both at least 15, using the null hypothesis value for π.

Hypothesis

- **Null Hypothesis**
  The null hypothesis of a test about a population proportion has form:
  $$H_0 : \pi = \pi_0 \tag{3.27}$$
- **Alternative Hypothesis**
  The alternative hypothesis has the form:
  - $H_1 : \pi < \pi_0$ (one-sided, left tail, lower tailed)
  - $H_1 : \pi > \pi_0$ (one-sided, right tail, upper tailed)
  - $H_1 : \pi \neq \pi_0$ (two-sided, two tail, two tailed)

Test Statistic

The test statitic is:

$$Z_{obs} = \frac{\hat{\pi} - \pi_0}{se_0}, \quad \text{where} \quad se_0 = \sqrt{\frac{\pi_0 (1 - \pi_0)}{n}}. \tag{3.28}$$

It has a standard normal distribution when the null is true

P-Value

The p-value summarizes the evidence. It describes how unusual the observed data would be if H0 were true. There are three ways to compute the p-value:

- One-sided,left tail:
  $$H_1 : \pi < \pi_0 \Longrightarrow P(Z < z_{obs}) \tag{3.29}$$
- One-Side, two tails
  $$H_1 : \pi > \pi_0 \Longrightarrow P(Z > z_{obs}) \tag{3.30}$$
- Two-Sided, Two tails:
  $$H_1 : \pi \neq \pi_0 \Longrightarrow P(|Z| > |z_{obs}|) = \mathrm{P}\left(Z < -|z_{obs}|\right) + \mathrm{P}\left(Z > |z_{obs}|\right) = \tag{3.31}$$
  $$= 2 \times \mathrm{P}\left(Z > |z_{obs}|\right) \tag{3.32}$$

Conclusion

Compare the p-value with the pre-specified significance level (usually 0.05).

3.3.4 Significance Test About Means

Assumptions

- The variable is quantitative
- The data are obtained using randomization (such as a random sample or a randomized experiment)
- The population distribution is approximately normal.
- The sample size is sufficiently large that the sampling distribution of the sample mean is approximately normal (population variance is unknown); the exact distribution is a T-distribution.

Hypothesis

- **Null Hypothesis**
  The null hypothesis of a test about a population proportion has form:

$$H_0 : \mu = \mu_0 \tag{3.33}$$

- **Alternative Hypothesis**
  The alternative hypothesis has the form:
  - $H_1 : \mu < \mu_0$ (one-sided, left tail, lower tailed)
  - $H_1 : \mu > \mu_0$ (one-sided, right tail, upper tailed)
  - $H_1 : \mu \neq \mu_0$ (two-sided, two tail, two tailed)

Test Statistic

The test statistic measures how far the sample mean falls from the null hypothesis value μ_0, as measured by the number of standard errors between them. The test statistic is:

$$T_{obs} = \frac{\bar{Y} - \mu_0}{SE} \quad \text{where} \quad SE = \frac{S}{\sqrt{n}}. \tag{3.34}$$

P-Value

The p-value summarizes the evidence. It describes how unusual the observed data would be if H0 were true. There are three ways to compute the p-value:

- One-sided,left tail:

$$H_1 : \pi < \pi_0 \Longrightarrow P(Z < t_{obs}) \tag{3.35}$$

- One-Side, two tails

$$H_1 : \pi > \pi_0 \Longrightarrow P(Z > t_{obs}) \tag{3.36}$$

- Two-Sided, Two tails:

$$H_1 : \pi \neq \pi_0 \Longrightarrow P(|Z| > |t_{obs}|) = \mathrm{P}\left(Z < -|t_{obs}|\right) + \mathrm{P}\left(Z > |t_{obs}|\right) = \tag{3.37}$$

$$= 2 \times \mathrm{P}\left(Z > |t_{obs}|\right) \tag{3.38}$$

Conclusion

Compare the p-value with the pre-specified significance level (usually 0.05).

3.3.5 Decisions and Types of Errors in Significance Tests

Significance Level

The α-level, also called the significance level, is a number α between 0 and 1 such that we reject H_0 if the $P - value \leq \alpha$. In practice, common α-levels are 0.05 and 0.01.

Two Potential Types of Errors in Test Decisions

Because of sampling variability, decisions in significance tests always have some uncertainty. A decision can be in error. Two are the possible errors:

1. When H_0 is true, a Type I error occurs when H_0 is rejected.
2. When H_0 is false, a Type II error occurs when H_0 is not rejected.

	Decision	
P-Value	H_0	H_a
P-value ≤ 0.05	Reject	Accept
P-value > 0.05	Do not reject	Do not accept

(3.39)

The Probability of a Type I Error

Suppose H_0 is true. The probability of rejecting H_0, thereby making a Type I error, equals the chosen significance level α for the test.

3.3.6 Optimal Test

The two error probabilities are closely (inversely) related. It follows that it is not possible to minimize both simultaneously. How do we choose the best test?

- the best test minimizes a weighted mean of the two error probabilities
- (Neyman-Pearson approach) given an upper bound for the probability of a type I error, the best test minimizes the probability of a type II error for each possible value of the parameter in the region of the alternative hypothesis.

3.4 Comparing Two Population Means or Proportions

3.4.1 Sampling Distribution of the Difference Between Two Sample Proportions

There are two cases that we will consider:

1. when the populations are independent and different proportion
2. when the populations are independent and common proportion

3.4.2 ... with different Proportion

For two independent random samples of size n and m from two groups:

$$X_1, X_2, \ldots, X_n \sim Be\,(p_1) \tag{3.40}$$

$$Y_1, Y_2, \ldots, Y_m \sim Be\,(\,p_2) \tag{3.41}$$

with population proportion p_1 and p_2, the sampling distribution of the difference in the sample proportions has:

$$\text{mean: } \mu_{X-Y} = p_1 - p_2 \tag{3.42}$$

$$\text{standard deviation: } \sigma_{X-Y} = \sqrt{\frac{p_1\,(1-p_1)}{n} + \frac{p_2\,(1-p_2)}{m}} \tag{3.43}$$

In practice, we'll have to estimate the standard deviation because it depends on the unknown population proportions (with standard error). The standard error is obtained by replacing the unknown population proportion p_1 and p_2 in the formula for the standard deviation by the corrisponding sample proportiond $\hat{p_1}$ and $\hat{p_2}$:

$$se = \sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{m}} \tag{3.44}$$

3.4.3 Confidence intervals for the difference in two population proportions (different Proportion)

For large random samples, an (approximate) $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$, the difference in two population proportions, is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}} \tag{3.45}$$

3.4.4 ... with common Proportion

For two independent random samples of size n and m from two groups:

$$X_1, X_2, \ldots, X_n \sim Be\,(p) \tag{3.46}$$
$$Y_1, Y_2, \ldots, Y_m \sim Be\,(\ p) \tag{3.47}$$

with population proportion $p_1 = p_2 = p$, the sampling distribution of the difference in the sample proportions has:

$$\text{mean: } \mu_{X-Y} = p_1 - p_2 = p - p = 0 \tag{3.48}$$

$$\text{standard deviation: } \sigma_{X-Y} = \sqrt{p(1-p)\left(\frac{1}{n} + \frac{1}{m}\right)} \tag{3.49}$$

In practice, we'll have to estimate the standard deviation because it depends on the unknown population proportions (with standard error). The standard error is obtained by replacing the unknown population proportion p_1 and p_2 in the formula for the standard deviation by the corrisponding sample proportiond $\hat{p_1}$ and $\hat{p_2}$:

$$se = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)} \tag{3.50}$$

where $\hat{p}$ is the pooled estimate of the common proportion.

3.4.5 Sampling Distribution for the Difference of Two Sample Means

There are two cases that we will consider:

1. when the populations are independent and normally distributed with unequal variances
2. when the populations are independent and normally distributed with a common variance

3.4.6 ... with unequal Variances

For two independent random samples of size n and m from two groups:

$$X_1, X_2, \ldots, X_n \sim N\left(\mu_X, \sigma_X^2\right) \tag{3.51}$$
$$Y_1, Y_2, \ldots, Y_m \sim N\left(\mu_Y, \sigma_Y^2\right) \tag{3.52}$$

with population means μ_X and μ_Y and population standard deviations σ_X^2 and σ_Y^2 , the sampling distribution of the difference in the sample means has:

$$\text{mean: } \mu = \mu_X - \mu_Y \tag{3.53}$$

$$\text{standard deviation: } \sigma_{X-Y} = \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \tag{3.54}$$

In practice, we will have to estimate the standard deviation because it depends on the unknown population standard deviations. Replacing S_X^2 and S_Y^2 by the sample standard deviations σ_X^2 and σ_Y^2, we get the standard error of the difference:

$$se = \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \tag{3.55}$$

3.4.7 Confidence Interval for the Difference Between Two Population Means (with unequal Variances)

A $(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, the difference in the population means is:

$$(\bar{X} - \bar{Y}) \pm \left(t_{\alpha/2, n+m-2}\right) \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} \tag{3.56}$$

3.4.8 ... with common Variances

For two independent random samples of size n and m from two groups:

$$X_1, X_2, \ldots, X_n \sim N\left(\mu_X, \sigma^2\right) \tag{3.57}$$

$$Y_1, Y_2, \ldots, Y_m \sim N\left(\mu_Y, \sigma^2\right) \tag{3.58}$$

with population means μ_X and μ_Y and population standard deviations $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, the sampling distribution of the difference in the sample means has:

$$\text{mean: } \mu = \mu_X - \mu_Y \tag{3.59}$$

$$\text{standard deviation: } \sigma_{X-Y} = \sigma^2 \sqrt{\frac{1}{n} + \frac{1}{m}} \tag{3.60}$$

In practice, we will have to estimate the standard deviation because it depends on the unknown population standard deviations. Replacing S_p that is a pooled sample variance with σ and we get the standard error of the difference:

$$se = S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \tag{3.61}$$

where the pooled sample variance is given by:

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \tag{3.62}$$

3.4.9 Confidence Interval for the Difference Between Two Population Means (with common Variances)

A $(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$, the difference in the population means is:

$$(\bar{X} - \bar{Y}) \pm \left(t_{\alpha/2, n+m-2}\right) S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \tag{3.63}$$

Chapter 4

Statistical Models

4.1 Predicting the Outcome of a Variable

4.1.1 Definition of a Regression Line

The regression line predicts the value for the response variable y as a straight-line function of the value x of the explanatory variable. Let $\hat{y}$ denote the predicted value of y. The equation for the regression line has the form:

$$\hat{y} = a + bx \tag{4.1}$$

In essence, a regression line is a straight line that describes how the response variable (y) changes as the explanatory variable (x) changes. We denote:

- a: y-intercept
- b: the slope

4.1.2 Interpreting the y-Intercept and Slope

y-Intercept

- The predicted value for y when $x = 0$
- May not have any interpretative value if 0 is not in the range of x values
- Represents omitted variables that does not vary in the data

Slope

- measures the change in the predicted variable for a 1 unit increase in the explanatory variable (x)
- When the slope is negative, the predicted value $\hat{y}$ decreases as x increases. The straight line then goes downward, and the relationship is negative.
- When the slope = 0, the regression line is horizontal (parallel to the x-axis). The predicted value $\hat{y}$ of y stays constant at the y-intercept for any value of x. Then the predicted value yn does not change as x changes, and the variables do not exhibit an association.

4.1.3 Concept of Error

The following equation predicts a single value of the response variable:

$$\hat{y} = a + bx \tag{4.2}$$

But we should not expect all subjects at that value of x to have the same value of y. Variability occurs in the y variable given x.

Error

The true equation is:

$$\hat{y} = a + bx + \epsilon \tag{4.3}$$

where ϵ is an error term representing:

- the variability of y given x
- omitted variables
- errors in the equation (it could be not perfectly linear)

4.1.4 Concept of Residual

In a scatterplot, the vertical distance between the point and the regression line is the absolute value of the residual.

Residual

The residual (or called observation errors) is denoted as the difference:

$$e_i = y - \hat{y} \tag{4.4}$$

between the actual value and the predicted value of the response variable.

4.1.5 Residual Sum of Squares (RSS)

The actual summary measure used to evaluate regression lines is called the Residual Sum of Squares:

$$RSS = \sum(\text{ residual })^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{4.5}$$

This formula squares each vertical distance between a point and the line and then adds up these squared values.

- The better the line, the smaller the residuals tend to be, and the smaller the residual sum of squares tends to be.
- For each potential line, we have a set of predicted values, a set of residuals, and a residual sum of squares

4.1.6 The Method of Least Squares Yields the Regression Line

Least Squares Method

Among the many possible lines that could be drawn through data points in a scatterplot, the least squares method gives what we call the regression line. This method produces the line that has the smallest value for the residual sum of squares using $\hat{y} = a + bx$ to predict y:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{4.6}$$

So, the least squares regression line is the line that minimizes the vertical distance between the points and their predictions, in essence, it minimizes the residual sum of squares

Properties

Besides making the errors as small as possible, this regression line:

- Has some positive residuals and some negative residuals, and the sum (and mean) of the residuals equals 0; so the property tells us that the too-low predictions are balanced by the too-high predictions.
- Passes through the point $(\bar{x}, \bar{y})$; so the property tells us that the line passes through the center (as measured by the mean) of the data.

Regression Formulas for y-Intercept and Slope

- **Slope**

$$b = r\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4.7}$$

- **Intercept**

$$a = \bar{y} - b\bar{x} \tag{4.8}$$

4.1.7 r-Squared (r^2)

It is important to understand if the line fit well the data, i.e. if the residuals are small.

r-Squared

r^2 is interpreted as the percent of the variability in the response variable that can be explained by the linear relationship between x and y. The goodness of fit can be measured by using the r2 (the square of the correlation coefficient).

$$r^2 = \frac{\text{variance of } \hat{y}}{\text{variance of y}} \tag{4.9}$$

Properties

- lies between 0 and 1
- is 1 if and only if the sum of squared residuals is 0 (perfect fit)
- is 0 if and only if the slope is 0 ($\hat{y}$ does not have variability)

4.2 Concept of Regression

4.2.1 Definition: Regression

Regression is a statistical methodology that use the statistical relation between predictor variable(s)(input, independent variable(s), etc.) and a response variable (output, dependent variable, etc.), so that a response can be predicted from the others.

4.2.2 Two Distinct Goals

1. Construct and Tests about statistical relation between predictor variables and response variables

2. Prediction

4.2.3 Sample vs Population

Not only Regression, Statistical Methods are usually used to make generalization about population based on information of sample:

- **Sample**
  A sample is the collection of units (people, animals, cities, fields, whatever you study) that is actually measured or surveyed in a study. The sample, a subset of the population, is used to estimate characteristics of the population.

- **Population**
  The population is the large group of units we are interested in, from which the sample was selected.

4.2.4 Notations

Notations for Sample and Population

Different notations are used for sample and population characteristics:

- The mean of a sample is often denoted as $\bar{y}$

- The mean of a population is often denoted as μ. An alternative notation for population mean is:

$$\mu = E(Y) \tag{4.10}$$

In practice, we do not know the exact value of μ but only the value of $\bar{y}$. Therefore, we often use the sample characteristic to estimate the feature of the population.

Regression Notation

A component of the simple regression model is that the mean value of the Y-variable is a straight line function of an X-variable. The two coefficients of a straight line are the intercept and the slope.

$$\text{Population: } E(Y) = \beta_0 + \beta_1 X \tag{4.11}$$

$$\text{Sample (Regression Analysis): } \hat{y} = b_0 + b_1 x \tag{4.12}$$

- **Intercept**

$$\begin{aligned} \text{population} &: \beta_0 \\ \text{sample} &: b_0 \text{ or } \hat{\beta}_0 \end{aligned} \tag{4.13}$$

- **Slope**

$$\begin{aligned} \text{population} &: \beta_1 \\ \text{sample} &: b_1 \text{ or } \hat{\beta}_1 \end{aligned} \tag{4.14}$$

4.3 Model for Simple Linear Regression

Regression is a method for studying the relationship between a response variable Y and a covariate X. The covariate is also called a predictor variable or a feature. One way to summarize the relationship between X and Y is through the regression function:

$$r(x) = \mathbb{E}(Y \mid X = x) = E(Y) = \int y f(y \mid x) dy \tag{4.15}$$

4.3.1 The Linear Model Describes a Conditional Expectation

A basic statistical model with a single explanatory variable has equation:

$$E(Y) = \beta_0 + \beta_1 x \tag{4.16}$$

and an alternative formulation for the model expresses Y_i, rather than $E(Y_i)$, in terms of x_i. The alternative model formula is:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.17}$$

where:

- y : response/dependent variable
- x : predictor/independent variable
- ϵ : random error of Y

This model is called the the simple linear regression model.

Assumptions of Errors

- All the errors ϵ are independent with mean 0

$$E(\epsilon) = 0 \tag{4.18}$$

- All the errors ϵ have the same degree of variation from the regression line for all x

$$Var(\epsilon) = \sigma^2 \tag{4.19}$$

- For the purpose of statistical inference, we assume that the errors have a normal distribution

$$\epsilon \sim N\left(0, \sigma^2\right) \tag{4.20}$$

4.3.2 Sample Estimates of Model

Assume simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n \tag{4.21}$$

Fitted model:

$$\hat{y}_i = b_0 + b_1 x_i, i = 1, \ldots, n \tag{4.22}$$

Our goal is to estimate $\beta_0, \beta_1 \text{by} b_0, b_1$ based on a sample of observations (x_i, y_i) of size n.

4.3.3 How to decide the best estimates based on the sample?

Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 \tag{4.23}$$

Based on the rule of least sums of squared errors, the estimated β coefficients in the simple linear regression model have the following expressions:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ b_0 &= \bar{y} - b_1\bar{x} \end{aligned} \tag{4.24}$$

b_0 and b_1 are called the least square estimates of β_0 and β_1.

Mean Squared Error

$$MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{4.25}$$

MSE is the sample variance of the errors and estimates σ^2.

Important Note: The divisor n - 2 only applies to simple regression. The general rule is that the divisor is n - p, where p =number of parameters in the regression equation.

Standard Deviation of Errors

$$s = \sqrt{MSE} \tag{4.26}$$

which is the sample standard deviation (residuals) of the errors from the regression line.

Sums of Squares Total (Total Sum of Squares)

$$SST(\text{or } SSTO) = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{4.27}$$

R-square

$$R^2 = \frac{SST - SSE}{SST} \tag{4.28}$$

where:

$$\text{correlation: } \sqrt{R^2} = r \tag{4.29}$$

Interpretation: R^2 is interpreted as the fraction of variation in y that is explained by the fitted regression equation. It is often converted to a percentage.

4.4 Sampling Distribution of b_0 and b_1 in Linear Regression

4.4.1 Remark Normal Distribution

- If $W \sim N\left(\mu, \sigma^2\right)$ then:
$$W + a \sim N\left(\mu + a, \sigma^2\right) \tag{4.30}$$
where a is a constant.
- If $W \sim N\left(\mu, \sigma^2\right)$ then a standard normal distribution or Z-distribution is:
$$\frac{W - \mu}{\sigma} \sim N(0, 1) \tag{4.31}$$
- If $W \sim N\left(\mu, \sigma^2\right)$ then:
$$a \times W \sim N\left(a\mu, a^2\sigma^2\right) \tag{4.32}$$
- If $W_1 \sim N\left(\mu_1, \sigma_1^2\right)$, $W_2 \sim N\left(\mu_2, \sigma_2^2\right)$ and W_1 and W_2 are indipendent then:
$$W_1 + W_2 \sim N\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right) \tag{4.33}$$

4.4.2 Sampling Distribution of b_0 and b_1

Assumption

For simple linear regression:
$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4.34}$$
We assume:

- ϵ follows normal distribution with mean 0 and variance σ^2 (usually unknown):
$$\epsilon \sim N(0, \sigma^2) \tag{4.35}$$
- β_0, β_1 are population parameters(fixed constant).
- X are known constant.

By the properties of Normal Distribution, Y follows normal distribution with:

- Mean: $\beta_0 + \beta_1 X$
- Variance: σ^2
$$Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right) \tag{4.36}$$

Estimation

We will use sample estimates $b_0(\hat{\beta}_0)$ and $b_1(\hat{\beta}_1)$ to estimate and make inference. Recall the Sum Square Error:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{4.37}$$

We minimize the equation:

$$min_{\beta_0,\beta_1} SSE = min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{4.38}$$

We obtain:

$$\begin{aligned} \frac{\partial SSE}{\partial \beta_0} &= -\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0 \\ \frac{\partial SSE}{\partial \beta_1} &= -\sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0 \end{aligned} \tag{4.39}$$

We rewrite these equations as:

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i, \quad \sum_{i=1}^{n} x_i y_i = \beta_0 \left(\sum_{i=1}^{n} x_i \right) + \beta_1 \sum_{i=1}^{n} x_i^2. \tag{4.40}$$

Simultaneous solution of the two equations yields the least squares estimates, which are unique. They are:

$$b_1 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \tag{4.41}$$

$$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4.42}$$

where:

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}} \text{ and } s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}. \tag{4.43}$$

we can write s_x and s_{xy} in other expressions:

$$s_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i (x_i - \bar{x}) \tag{4.44}$$

$$s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x}) y_i \tag{4.45}$$

Finding the Distribution

We know that:

$$\begin{cases} Y \sim N\left(\beta_0 + \beta_1 X, \sigma^2\right) \\ b_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_i (X_i - \bar{X}) Y_i}{s_{xx}} \end{cases} \tag{4.46}$$

So we find that b_1 is a Normal Distribution:

$$b_1 = \frac{\sum_i \left(X_i - \bar{X}\right)}{s_{xx}} Y_i \sim N(\frac{\sum_i \left(X_i - \bar{X}\right)}{s_{xx}} \mathbb{E}(Y_i), (\frac{\sum_i \left(X_i - \bar{X}\right)}{s_{xx}})^2, Var(Y_i)) \tag{4.47}$$

Of course, same for b_0.

Distribution of b_0 and b_1

b1 and b0 are normally distributed and:

$$\begin{aligned} b_1 &\sim N\left(\beta_1, \operatorname{Var}\left(b_0\right)\right) \\ b_0 &\sim N\left(\beta_0, \operatorname{Var}\left(b_1\right)\right) \end{aligned} \tag{4.48}$$

where:

$$\operatorname{Var}\left(b_0\right) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right], \operatorname{Var}\left(b_1\right) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{4.49}$$

4.4.3 Standardized b_0, b_1

Since b1 and b0 are normally distributed, we know the standardized statistic:

$$\frac{b_1 - \beta_1}{\sqrt{\operatorname{Var}\left(b_1\right)}} \sim N(0,1) \tag{4.50}$$

$$\frac{b_0 - \beta_0}{\sqrt{\operatorname{Var}\left(b_0\right)}} \sim N(0,1) \tag{4.51}$$

However if we replace Var by estimate s^2 (usually reported in R), we have:

$$\begin{aligned} \frac{b_1 - \beta_1}{s\left(b_1\right)} &\sim t(n-2) \\ \frac{b_0 - \beta_0}{s\left(b_0\right)} &\sim t(n-2) \end{aligned} \tag{4.52}$$

4.5 Statistical Inference Tools

Basically, there are two main statistical inference tools:

- Hypothesis testing
- Confidence interval

They are equivalent in some sense.

4.5.1 Hypothesis Testing for slope β_1

For instance, we usually want to know whether the slope of the simple regression model equals 0 or not, since the slope directly tells us about the link between mean y and x.

- When the true population slope β_1 does not equal 0, the variables y and x are linearly related.
- When the slope is 0, there is no linear relationship because mean y does not change when the value of x is changed.

Step 1: the null and alternative hypotheses

$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0 \tag{4.53}$$

Step 2: Construct a test statistic

- A test statistic should not contain the unknown parameter.
- The test statistic is a random variable.
- The distribution of the test statistic should be known.

Start with the sampling distributions of b1:

$$\frac{b_1 - \beta_1}{s\left(b_1\right)} \sim t(n-2) \tag{4.54}$$

Hence, T below can serve as a test statistic:

$$T = \frac{b_1}{s\left(b_1\right)} \sim t(n-2) \text{ under } H_0 : \beta_1 = 0 \tag{4.55}$$

Step 3: Make conclusion

Under H_0, the density curve of the constructed T variable is bell-shaped and symmetric at 0. The area under the t-curve gives us the probability of the variable taking values in a certain interval.

Critical value approach

- If H0 is true, that is, the random variable:

  $$T \sim t(n-2) \tag{4.56}$$

  then the sample value of T, denoted by T0 (T0 is the value of T given the specific sample), should have a large chance to fall in the middle area, that is, T0 should be close to 0.

- if $\mid T_0 \mid$ is "too large", H0 is likely to be wrong. In practice, we use:

  $$\mid T_0 \mid > t_{\frac{\alpha}{2}}(n-2) \tag{4.57}$$

  to indicate "too large" and that we should reject H0, e.g $\alpha = 0.05$.

p-value approach

- If H0 is true, the sample value T0 has a large chance to fall in the middle area, i.e. $\mid T_0 \mid$ is small. Then the probability that random variable T is more extreme than T0 is large:

  $$\text{p-value } = P\left(|T| > |T_0|\right) \tag{4.58}$$

  should be large.

- if p-value is "very small", H0 is likely to be wrong; in practice, we use:

  $$\text{p-value } < \alpha \tag{4.59}$$

  to indicate "very small", e.g. $\alpha = 0.05$.

4.5.2 Hypothesis Testing for intercept β_0

Step 1: the null and alternative hypotheses

$$H_0 : \beta_0 = 0, H_1 : \beta_0 \neq 0 \tag{4.60}$$

Step 2: Construct a test statistic

- A test statistic should not contain the unknown parameter.
- The test statistic is a random variable.
- The distribution of the test statistic should be known.

Start with the sampling distributions of b1:

$$\frac{b_0 - \beta_0}{s\left(b_0\right)} \sim t(n-2) \tag{4.61}$$

Hence, T below can serve as a test statistic:

$$T = \frac{b_0}{s\left(b_0\right)} \sim t(n-2) \text{ under } H_0 : \beta_0 = 0 \tag{4.62}$$

Step 3: Make conclusion

Same as for the slope.

4.5.3 Confidence Interval for slope β_1

A $1-\alpha$ confidence interval for the unknown slope β_1 can be computed as:

$$\begin{aligned} &(\text{point estimate } \pm \text{ Multiplier } t^\star \times \text{ sd of the point estimate}) \\ &\left(b_1 \quad \pm t_{\alpha/2}(n-2) \times s\left(b_1\right)\right) \\ &b_1 \pm t_{\alpha/2}(n-2) \times \frac{\sqrt{MSE}}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)}} \end{aligned} \tag{4.63}$$

4.5.4 Hypothesis Testing v.s. Confidence Interval

They are equivalent in some sense. Consider the hypothesis testing:

$$H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0 \tag{4.64}$$

Reject the null hypothesis if 0 is not included in the $1-\alpha$ confidence interval for the slope. So two cases:

1. If CI for β_1 contains 0, we conclude that there is no evidence of a linear relationship between the predictor and the response in the population
2. If CI for β_1 does not contain 0, we conclude that there is evidence of a linear relationship between the predictor and the response in the population.

4.6 Confidence Interval and Prediction Inteval

4.6.1 Confidence Interval for $E(Y)$

A $1-\alpha$ confidence interval for $E(Y)$ is an interval estimate for the mean value of y or, $E(Y)$ in the population level given an x.

$$\begin{gathered} (\text{fitted value} \pm \text{Multiplier } t^\star \times \text{ standard error of the fitted value}) \\ \left(\widehat{y} \pm t_{(1-0.95)/2}(n-2) \times s\{\widehat{y}\}\right) \end{gathered} \tag{4.65}$$

where:

$$\widehat{y} = b_0 + b_1 x \tag{4.66}$$

$$s\{\widehat{y}\} = \sqrt{MSE \cdot \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right)} \tag{4.67}$$

Notes that:

- a confidence interval for E(Y) estimates the location of the line at a specific x value.
- It counts the variation of different samples.
- C.I with higher confidence level will be wider. e.g. 99% C.I is wider than 95% C.I.

4.6.2 Prediction Interval for y

A prediction interval is an interval estimate for a new observation y corresponding to a given level x_h. A $(1-\alpha)$ prediction interval for y give x_h is:

$$\hat{y} \pm t_{\alpha/2}(n-2)\sqrt{s^2\{\hat{y}\} + MSE} \tag{4.68}$$

where:

$$\hat{y} = b_0 + b_1 * x_h \tag{4.69}$$

Notes that:

- It has to count for the variation in the mean (as in confidence intervals), but also count for random error of observations
- Wider than a confidence interval.
- Interpretation similar as a confidence interval.

4.6.3 CI for $E(Y)$ and PI for y : Comparison

Object and Formula

	object	formula
C.I.	mean $E(Y)$	$\hat{y} \pm t_{\alpha/2}(n-2) \times s\{\hat{y}\}$
P.I.	observation y	$\hat{y} \pm t_{\alpha/2}(n-2) \times \sqrt{MSE + s^2\{\hat{y}\}}$

Intepretation

	Intepretation
C.I.	with 95% confidence, we can estimate the mean of the response for a given x is in the C.I.
P.I.	with 95% confidence, we can predict the response value for a given x is in the P.I.

4.6.4 Important Notes

1. C.I. and P.I. can only be used when x is a value within the range of the x values in the data set.
2. Extrapolation: when x is "out the scope of the model":
   - Don't know whether it still follows the same linear regression model.
   - Does not make sense.
3. But x does not have to be one of the actual x values in the data set.

4.7 Multiple Regression Model

The linear regression model with a single explanatory variable extends to include multiple explanatory variables. For subject i in the sample, let x_{ij} denote the value of explanatory variable j. The linear model with p explanatory variables, called the multiple regression model, is:

$$E\left(Y_i\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{4.70}$$

Where:

- $(x_{i1}, \ldots, x_{ip})$: explanatory variables
- $\mathbb{E}(Y_i)$: represents the conditional expectation of the response variable

The standard linear model assumes that Y has constant variance σ^2 at each combination of values of the explanatory variables.

4.7.1 Linear Model

the relationship between a response variable y and more than more predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon \tag{4.71}$$

$$Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon_i \tag{4.72}$$

4.7.2 Notation for Multiple Regression Model

Model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \end{aligned} \tag{4.73}$$

Assumption

The linear model above is reasonable, and ϵ_i are independent and follows $N(0, \sigma^2)$ distribution.

Note

The subscript i refers to the ith individual or unit. In the notation for the x-variables, the subscript following i indicates which x-variable it is. (example: x_{i2} represents the 2nd x-variable for subject i)

4.7.3 Matrix Notation for Multiple Regression Model

Remarks

Consider the multiple regression models for all the individuals in the sample (size n), with intercept and p-1 predictor variables:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots + \beta_{p-1} x_{1,p-1} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots + \beta_{p-1} x_{2,p-1} + \epsilon_2 \\ &\ldots \\ y_i &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i \\ &\ldots \\ y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_{p-1} x_{n,p-1} + \epsilon_n \end{aligned} \tag{4.74}$$

And the corresponding regression equations:

$$\begin{aligned}
E(y_1) &= \beta_0 + \beta_1 x_{1,1} + \beta_2 x_{1,2} + \ldots + \beta_{p-1} x_{1,p-1} \\
E(y_2) &= \beta_0 + \beta_1 x_{2,1} + \beta_2 x_{2,2} + \ldots + \beta_{p-1} x_{2,p-1} \\
&\ldots \\
E(y_i) &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} \\
&\ldots \\
E(y_n) &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_{p-1} x_{n,p-1}
\end{aligned} \tag{4.75}$$

Then, we have:

$$\begin{aligned}
y_1 &= E(y_1) + \epsilon_1 \\
y_2 &= E(y_2) + \epsilon_2 \\
&\ldots \\
y_i &= E(y_i) + \epsilon_i \\
&\ldots \\
y_n &= E(y_n) + \epsilon_n
\end{aligned} \tag{4.76}$$

Now we write E(Y) as a design matrix (or say X matrix):

$$\begin{aligned}
E(y_i) &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} \\
&= 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \ldots + x_{i,p-1} \cdot \beta_{p-1} \\
&= (1 x_{i,1} x_{i,2} \ldots x_{i,p-1}) \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_{p-1} \end{pmatrix}.
\end{aligned} \tag{4.77}$$

Then:

$$E(y_i) = 1 \cdot \beta_0 + x_{i,1} \cdot \beta_1 + x_{i,2} \cdot \beta_2 + \ldots + x_{i,p-1} \cdot \beta_{p-1}; i = 1, 2, 3, \ldots, n$$

$$E(Y) = \begin{pmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \\ \ldots \\ E(y_n) \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \ldots & x_{3,p-1} \\ & \cdot & \cdot & & \ldots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_{p-1} \end{pmatrix} \tag{4.78}$$

General Equations

Regression equation: $E(Y) = X\boldsymbol{\beta}$; Regression model: Y = E(Y) + $\boldsymbol{\epsilon}$ or Y = X$\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where:

$$\begin{aligned}
\mathrm{Y} &= \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \cdot \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \ldots & x_{3,p-1} \\ 1 & \cdot & \cdot & & \ldots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p-1} \end{pmatrix} \\
\boldsymbol{\beta} &= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \ldots \\ \beta_{p-1} \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_n \end{pmatrix}
\end{aligned} \tag{4.79}$$

4.7.4 Estimates (Least Square Criterion)

The estimates of the β coefficients minimize the sum of squared distances from the observation points to the regression hyperplane.

$$\min \sum_i (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2 \tag{4.80}$$

The letter b is used to represent a sample estimate of a β coefficient ($\hat{\beta}$):

$$(b_0, b_1, \ldots, b_{p-1}) = \text{argmin}_{\beta_0,\ldots,\beta_{p-1}} \sum (y_i - \beta_0 - \ldots - \beta_{p-1} x_{i,p-1})^2 \tag{4.81}$$

4.7.5 Two Notation of Estimation

Vector Notation

$$\mathrm{b} = \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ \cdots \\ b_{p-1} \end{pmatrix} \tag{4.82}$$

where:

- b_0: The estimated mean y when all x variables are 0.
- b_j: The estimated mean change in y when x_j increases one unit when other x-variables remain the same.

Matrix Form

Notes that:

$$\textbf{b} \text{ minimize } \sum_i (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_{p-1} x_{i,p-1})^2 \Leftrightarrow \text{ minimize } (Y - X\boldsymbol{\beta})^\top (Y - X\boldsymbol{\beta}) \tag{4.83}$$

Then:

$$\mathrm{b} = \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ \cdots \\ b_{p-1} \end{pmatrix} = \left(\mathrm{X}^\top \mathrm{X}\right)^{-1} \mathrm{X}^\top \mathrm{Y} \tag{4.84}$$

4.8 Sampling Distribution of b in Multiple Regression

4.8.1 Sampling Distribution of b

Under normality assumption about random errors (same operation of Simple Linear), we have:

$$\mathrm{b} = \left(X^\top X\right)^{-1} X^\top Y \sim \text{MVN}\left(\boldsymbol{\beta}, \sigma^2 \left(X^\top X\right)^{-1}\right) \tag{4.85}$$

where MVN is a Multivariate Normal Distribution. Each b_i follows a normal distribution with:

$$\text{Mean: } \beta_i \tag{4.86}$$

$$\text{Variance: the ith diagonal elements of } (X^T X)^{-1} \tag{4.87}$$

4.8.2 Multicollinearity

The problem of Multicollinearity is the following:

Whenever some of the columns in design matrix X are highly correlated (i.e. some of the predictor variables are linear related), there is no unique solution for inverse of $X^T X$. That is, in such a case, $(X^T X)^{-1}$ does not exist uniquely.

Consequences

Consequences of high multicollinearity:

- Often confusing and misleading results: interpretation.
- Unstable estimates

4.8.3 Model fitting

Preticted/Fitted Value

- **Linear Notation**

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_{p-1} x_{i,p-1}, i = 1, \ldots, n \tag{4.88}$$

- **Matrix Notation**

$$\hat{Y} = Xb \tag{4.89}$$

Residual

- **Linear Notation**

$$e_i = y_i - \hat{y}_i, i = 1, \ldots, n \tag{4.90}$$

- **Matrix Notation**

$$e = Y - \hat{Y}, \text{ where } e = (e_1, \ldots, e_n)^\top \tag{4.91}$$

SSE

- **Linear Notation**

$$SSE = \sum_i (y_i - b_0 - b_1 x_{i,1} - \ldots - b_{p-1} x_{i,p-1})^2 \tag{4.92}$$

- **Matrix Notation**

$$SSE = (Y - Xb)^\top (Y - Xb) = e^\top e \tag{4.93}$$

Estimate σ and σ^2

- **Mean Square Error (MSE)**

$$MSE = \frac{SSE}{n-p} \text{ estimates } \sigma^2 \tag{4.94}$$

- **Standard Devation of Error (S)**

$$S = \sqrt{MSE} = \text{ estimate } \sigma \tag{4.95}$$

Specific expression of the standard error is the following:

$$\sqrt{\frac{\text{Residual SS}}{df}} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - (\text{ number of parameters in regression equation })}} \tag{4.96}$$

4.8.4 Validate Model Assumptions

Residual plot

Still a 2-D plot ($e_i$versus$\hat{y}_i$)! Ideally should resemble a horizontal random band.

Normal Probability Plot

Expect the normal probability plot to be a straight line and R-J test to have p-value larger than 0.05.

4.8.5 Inference: Testing Significance of Each β Coefficient

We may want to assess whether a particular x-variable is making a useful contribution to the model. That is, given the presence of the other x-variables in the model.
Consider, for example, the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{4.97}$$

Null and Alternative Test

To determine whether variable x1 is a useful predictor variable in this model, we could test:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{4.98}$$

If the null hypothesis above were the case, y and x1 are not significantly related, or x1 is not important when x2 and x3 are in the model.

Test Statistic

Carry out the test:

$$T = \frac{\text{sample coefficient}}{\text{standard error of the coefficient}} \sim t(n-p) \text{ under } H_0. \tag{4.99}$$

Critival Value Approach vs p-Value Approach

- **Critical Value Approach**

$$\text{if } |T_0| > t_{\alpha/2}(n-p), \text{ reject } H_0 \tag{4.100}$$

- **p-Value Approach**

$$\text{if } p\text{-value } = P\left(|T| > |T_0|\right) < \alpha, \text{ reject } H_0 \tag{4.101}$$

4.9 Analysis of Variance (ANOVA)

4.9.1 ANOVA Table

Two uses of ANOVA table:

1. A decomposition of variance in y
2. A significance test of whether x and y are really linearly related in the population.

4.9.2 A decomposition of variance in y

ANOVA table displays quantities that measure how much of the variability in the y-variable is explained and how much is not explained by the y-variables relationship with the x-variables. The main quantities are:

- **Sums of Squares for Total (SST)**
- **Sums of Squared Errors (SSE)**
- **Sum of Squares due to Regression (SSR)**

Sums of Squares for Total (SST)

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.102}$$

- it quantifies how much the observed responses vary if you don't take into account their predictor values. SST is a measure of the overall variation in the y-variables.
- Total degrees of freedom:

$$df(SST) = n - 1 \tag{4.103}$$

Sums of Squared Errors (SSE)

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4.104}$$

- it quantifies how much the data points vary around the estimated regression line.
- Errors degrees of freedom:

$$df(SSE) = n - p \tag{4.105}$$

Sum of Squares due to Regression (SSR)

$$SSR = SST - SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{4.106}$$

- it quantifies how far the estimated regression line is from the no relationship line.
- Regression degrees of freedom:

$$df(SSR) = df(SST) - df(SSE) = (n-1) - (n-p) = p - 1 \tag{4.107}$$

Summary Table

Souce	DF	SS	MS	F
Regression	$p-1$	SST-SSE	MSR	MSR/MSE
Error	$n-p$	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	MSE	
Total	$n-1$	$\sum_{i=1}^{n} (y_i - \bar{y})^2$		

(4.108)

4.9.3 A Significance Test: F Test

Mean Square for the Regression (MSR)

$$MSR = \frac{SSR}{df(SSR)} = \frac{SSR}{p-1} \tag{4.109}$$

Mean Squared Error(MSE):estimates of σ^2

$$MSE = \frac{SSE}{df(SSE)} = \frac{SSE}{n-p} \tag{4.110}$$

F statistics: used to test the significance in the linear relation

$$F = \frac{MSR}{MSE} \tag{4.111}$$

Hypothesis Testing

The F statistic in the analysis of variance can be used to test whether the y-variable is related to at least one x-variables in the model.

1. **The Null and Alternative Hypotheses**
   F statistic to test whether the y-variable and x-variable are related:

$$H_0 \quad : \quad \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \tag{4.112}$$
$$H_a \quad : \quad \text{at least one of the } \beta_i \neq 0 \text{ for } i = 1, \ldots, p-1 \tag{4.113}$$

   - The null hypothesis means that the y-variable is not related to any of the x-variables in the model.
   - The alternative hypothesis means that the y-variable is related to one or more of the x-variables in the model.

2. **Construct a test statistic**

$$F = MSR/MSE \sim F(p-1, n-p) \text{ under } H_0 \tag{4.114}$$

3. **Make Conclusion**

   - **Critical value approach**
     Reject H0 if the calculated statistic F0:

$$\text{If } F_0 > F_\alpha(p-1, n-p), \text{ reject } H_0 \tag{4.115}$$

   - **p-value approach**
     Usually:

$$p- \text{ value } \leq 0.05 \tag{4.116}$$

     reject the null hypothesis, and conclude that y is related to at least one of the x-variables in the model.

Remark: t-test vs F-test

- F can only be used for two-sided tests, t can also be used for one-sided tests.
- In multiple linear regression, the T-test and F-test are different:
  - T-test tests linear relation between y and a certain xi while all other x-variables are in the model
  - F-test test the linear relation between y and all x-variables together

Hypothesis Testing in Simple Linear Regression

F statistic to test whether the y-variable and x-variable are related:

1. **The Null and Alternative Hypotheses**
   F statistic to test whether the y-variable and x-variable are related:

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0 \tag{4.117}$$

   And in simple linear regression p = 2:

2. **Construct a test statistic**

$$F = MSR/MSE \sim F(p-1, n-p) \text{ under } H_0 : \beta_1 = 0 \tag{4.118}$$

3. **Make Conclusion**

- **Critical value approach**
  Reject H0 if the calculated statistic F0:

$$\text{If } F_0 > F_\alpha(2, n-1), \text{ reject } H_0 \tag{4.119}$$

  where $F_\alpha(2, n-1)$ is the $1-\alpha$ percentile of $F(1, n-1)$ distribution, α usually takes as 0.05

- **p-value approach**
  Usually:

$$p- \text{ value } \leq 0.05 \tag{4.120}$$

  reject the null hypothesis, and conclude that y is related to at least one of the x-variables in the model.

Other Uses of ANOVA Table: MSE and R^2

- MSE is the estimate of the error variance. Thus:

$$S = \sqrt{MSE} \tag{4.121}$$

- The Total and Error lines give the SS values used in the calculation of:

$$R^2 = \frac{SST - SSE}{SST} = \frac{SSR}{SSE} \tag{4.122}$$

```
Residuals:
    Min      1Q  Median      3Q     Max 
-78.231 -41.710   7.646  33.552 108.831 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 576.6819    23.4709  24.570  < 2e-16 ***
Age          -3.0068     0.4243  -7.086 1.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.76 on 28 degrees of freedom
Multiple R-squared:  0.642,	Adjusted R-squared:  0.6292 
F-statistic: 50.21 on 1 and 28 DF,  p-value: 1.041e-07

> anova(out2)
Analysis of Variance Table

Response: Distance
          Df Sum Sq Mean Sq F value    Pr(>F)    
Age        1 124333  124333  50.211 1.041e-07 ***
Residuals 28  69334    2476                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```