Metode Kuantitatif : Overview

Metode Kuantitatif

Tim pengajar

MK METODE KUANTITATIF

- SKS : 3 (2 – 2)
- Pengajar
  - Prof. Dr. Ir. Agus Buono, M.Si., M.Kom
  - Dr. Sony Hartono Wijaya, S.Kom., M.Kom→Koordinator
  - Dr. Toto Haryanto, S.Kom, MSi
  - Lailan Sahrina Hasibuan S.Kom MKom
  - Endang Purnama Giri, S.Kom, MKom
  - Aziz Kustiyo, SSi, M.Kom
- Jadwal Kuliah
  - Senin 08.00 dan Senin 10.00
- Jadwal Praktikum
  - Jumat 07.00 dan Jumat 09.00

Manfaat, Deskripsi dan Tujuan

- **Manfaat Matakuliah**
  - Matakuliah ini akan memberi manfaat bagi mahasiswa dalam memahami konsep-konsep statistika dalam merancang percobaan, pembangkitan data, reduksi dimensi, dan pemodelan data untuk menyelesaikan permasalahan sederhana dengan menggunakan *tools* yang relevan.
- **Deskripsi Matakuliah**
  - Mata kuliah ini membahas mengenai dasar dan teknik analisis dalam perancangan percobaan, pengumpulan data, metode survei, pemodelan linear, eksplorasi peubah ganda, reduksi dimensi, konsep jarak, *clustering*, klasifikasi dan pengenalan analisis regresi.
- **Tujuan**
  - Setelah menyelesaikan mata kuliah ini, mahasiswa diharapkan dapat memahami dan mampu menerapkan metode-metode kuantitatif dalam permasalahan bidang komputer.

Strategi perkuliahan

- Perkuliahan dilakukan sebanyak 14 kali pertemuan kuliah dan 14 kali praktikum secara daring dengan minimal 1 kali pertemuan kuliah setelah UTS yang dilaksanakan secara *hybrid*.
- Persyaratan mahasiswa yang dapat mengikuti kuliah luring di masa pandemi sesuai ketentuan Departemen Ilmu Komputer FMIPA IPB.
- Metode perkuliahan adalah *synchronous* dan *asynchronous*.
- Materi kuliah dan praktikum dapat diakses di newLMS **KOM1221 Metode Kuantitatif**.
- Silahkan bergabung sesuai kelas paralel penyelenggaraan kuliah (K1 - **genap2122_k1** dengan *enrollment key* **metkuan2022** dan K2 - **genap2122_k2** dengan *enrollment key* **2022metkuan**).

Penilaian

- Nilai akhir (NA) adalah nilai kumulatif dengan bobot nilai sebagai berikut:
- **UTS dan UAS** dilakukan melalui ujian tertulis dengan bobot total **50%**, dengan masing-masing memiliki bobot 25%.
- Aktivitas **partisipatif dan proyek akhir** dengan bobot **50%**, dengan rincian:
  - Tugas akhir (aktivitas partisipatif 5% dan laporan proyek akhir 20%)
  - Praktikum (aktivitas partisipatif diambil dari lembar kerja praktikum/nilai praktikum 20%) dan pengetahuan (5% diambil dari kuis praktikum)
- Selang nilai untuk menetapkan huruf mutu A, AB, B, BC, C, D, atau E ditentukan berdasarkan nilai rataan dan standar deviasi dengan menggunakan sebaran normal.

Tata Tertib Perkuliahan dan Praktikum

- Sesuai dengan ketentuan yang terdapat pada Buku Panduan Sarjana IPB:
  - Hadir dalam media pembelajaran daring/luring paling lambat 5 menit sebelum kegiatan dimulai.
  - Berpenampilan dan berbusana sopan serta rapi sesuai dengan tata terbit yang berlaku di IPB.
  - Tidak ada ujian perbaikan.
  - Mahasiswa yang melakukan kecurangan dalam pelaksanaan ujian (UTS, UAS, dan plagiat tugas) akan diberikan sanksi sesuai dengan ketentuan yang berlaku.

Presensi

- Mahasiswa diwajibkan menghadiri perkuliahan setidaknya 11 dari seluruh pertemuan, atau mendapat sanksi cekal UAS.
- Setiap mahasiswa diwajibkan mengikuti seluruh kegiatan praktikum (kehadiran praktikum: 100%). Jika mahasiswa tidak hadir dalam praktikum maka komponen tugas pada pertemuan tersebut adalah 0 dan tidak ada tugas pengganti. Hal ini dikecualikan bagi mahasiswa yang sakit atau mendaftar mata kuliah Metode Kuantitatif dalam KRS B.
- Surat keterangan sakit yang sah diserahkan kepada asisten praktikum selambat-lambatnya satu minggu sejak tanggal mahasiswa tersebut tidak hadir dalam praktikum.
- Pengumuman nama-nama mahasiswa yang tidak dapat mengikuti ujian akhir semester akan diberikan selambat-lambatnya 3 (tiga) hari sebelum ujian akhir semester dilaksanakan.

Materi Perkuliahan

Perancangan percobaan, pengumpulan data, metode survei, pemodelan linear, eksplorasi peubah ganda, reduksi dimensi, konsep jarak, *clustering*, klasifikasi dan pengenalan analisis regresi

Pertemuan	Kompetensi Dasar	Materi	Dosen
1	Mahasiswa dapat menjelaskan contoh penerapan metode kuantitatif dalam permasalahan bidang komputer dan dapat menyebutkan hal-hal terkait survei: kenapa, terminologi, tahapan-tahapan dalam survei dan memilih jenis survei yang sesuai	Overview materi perkuliahan, elemen-elemen survei, perencanaan survei	LSH
2	Mahasiswa dapat menjelaskan dan melakukan kegiatan survei dengan Simple Random Sampling (SiRS)	Metode survei Simple Random Sampling: definisi dan kenapa SiRS, bagaimana mengambil sampel, pendugaan rataan, total, dan proporsi, menentukan ukuran sampel.	LSH
3	Mahasiswa dapat menjelaskan dan melakukan kegiatan survei dengan Stratified Random Sampling (StRS)	Metode survei Stratified Random Sampling: definisi dan kenapa StRS, bagaimana mengambil sampel, pendugaan rataan, total, dan proporsi, menentukan ukuran sampel.	LSH
4-5	Mahasiswa dapat melakukan analisis regresi untuk permasalahan sederhana	Analisis regresi linear	LSH
		Analisis regresi logistik	LSH
6-7	Mahasiswa dapat menerapkan konsep perancangan percobaan pada penelitian bidang komputer	Konsep perancangan percobaan faktorial dalam Rancangan Acak Lengkap (RAL)	AGB
		Contoh kasus percobaan faktorial dalam RAL	AGB
UTS			

Pertemuan	Kompetensi Dasar	Materi	Dosen
8-9-10	Mahasiswa dapat menjelaskan dan menerapkan konsep eksplorasi peubah ganda untuk permasalahan dalam ilmu komputer	Analisis Komponen Utama (matriks covariance, nilai eigen, vektor eigen)	AGB
		Analisis Biplot	AGB
		Analisis Multidimensional scalling	AGB
11-12	Mahasiswa dapat menjelaskan dan menerapkan konsep jarak pada analisis klaster	Konsep jarak dan partitional clustering	SHW
		Hierarchical clustering	SHW
13-14	Mahasiswa dapat menjelaskan dan menerapkan metode klasifikasi	Metode klasifikasi KNN	TTH
		Metode klasifikasi dengan fungsi diskriminan	TTH
UAS			

Pendahuluan

- Sebutkan beberapa survey yang kalian ketahui
  - penyelenggara
  - Tujuan
  - Responden
  - Hasil survey
- Sebutkan percobaan yang kalian ketahui
  - pelaksana
  - Tujuan
  - hasil

Pendahuluan 1

- **Survey Kepuasan Pelanggan**
- Setiap tahun PT PLN (Persero) selalu mengadakan kegiatan Survey Kepuasan Pelanggan (SKP). Tujuan dilakukan survey ini adalah untuk mengukur kinerja pelayanan PT PLN (Persero) melalui indeks Kepuasan dan Ketidakpuasan Pelanggan yang meliputi seluruh Unit PLN Wilayah/Distribusi, tak terkecuali PT PLN (Persero) Wilayah Aceh.
- Pelanggan yang dilakukan survey terdiri dari semua golongan tarif, baik tarif Bisnis (B), Industri (I), Rumah Tangga (R), Sosial (S), dan Pemerintah (P). Selain terdiri dari semua tarif, baik pelanggan pascabayar maupun prabayar juga dijadikan target survey.
- Tahun ini area yang dilakukan SKP adalah Area Banda Aceh dan Area Lhokseumawe. Hasilnya cukup menggembirakan, nilai kepuasan pelanggan PLN Aceh adalah 88,40% dengan kategori Puas, sedangkan untuk nilai ketidakpuasan pelanggan PLN Aceh adalah 1,03%. Untuk Regional Sumatera, PLN Aceh berada di peringkat 3 untuk kepuasan pelanggan, sedangkan untuk ketidakpuasan pelanggan berada di peringkat 1.

- Sumber : http://www.pln.co.id/aceh/?p=805

Pedahuluan 2

- **Pengaruh Media Dasar dan Konsentrasi BAP terhadap Pertumbuhan Stek Buku Tunggal In vitro Tanaman Zaitun (Olea europea L.)**

  oleh Kusumaningsih, Nita Ayu

Metode Penelitian

- **Pertumbuhan Tunas Lateral**
- Eksplan yang steril dari metode S4 di tanam ke dalam media percobaan. - Rancangan percobaan yang digunakan dalam percobaan ini adalah Rancangan Acak Lengkap (RAL) faktorial 2 faktor yang terdiri atas faktor perbedaan jenis media dasar dan konsentrasi BAP. Jenis media yang digunakan adalah WPM dan DKW. Konsentrasi BAP yang digunakan adalah 0, 1,2, dan 4 ppm. Masing-masing perlakuan dibuat 10 ulangan dengan 1 eksplan per botol. Kultur dipelihara dalam ruang kultur dengan pencahayaan 800-1000 lux selama 16 jam per hari, dan suhu ruangan 25 ± 2 °C di ruang kultur. Pengamatan dilakukan selama 8 minggu dengan waktu pengamatan setiap seminggu sekali. Parameter yang diamati meliputi waktu awal pertumbuhan tunas, jumlah tunas pada setiap eksplan, panjang tunas dan jumlah daun serta kondisi kultur.

- Metode pembangkitan data
  - Survey
  - Perancangan perobaan

http://pages.uoregon.edu/aarong/teaching/G4074_Outline/node16.html

Types of data collection

- **Anecdotal evidence**: "My uncle once ...." - Haphazardly selected small number of observations - completely unreliable, not considered valid for scientific inquiry
- **Observational study** - We observe individuals (or process or units, etc) and (potentially) measure variables as they are in the population. A particularly important kind of observational study is one drawn from a **sample**, because if the sample is well chosen, then we can better generalize from the data. If we have a **census**, then we have entire population, and generalizability is not an issue.
- **Experiment** - We actively impose treatment on some individuals in order to observe a response. If we choose individuals for treatment well, then we will be able to make a causal connection between treatment and response.

Comparing Statistical Surveys and Statistical Experiments

By Deborah J. Rumsey from Statistics For Dummies, 2nd Edition

- There are two major types of statistical studies: surveys and experiments.
- After a question has been formed, researchers must design an effective study to collect data that will help answer that question.
- This means that they must decide whether to use a survey or experiment to get the data they need.

Statistical surveys

- An *observational study* is one in which data is collected on individuals in a way that doesn't affect them.
- The most common observational study is the survey.
- *Surveys* are questionnaires that are presented to individuals who have been selected from a population of interest.
- Surveys take many different forms: paper surveys sent through the mail, questionnaires on Web sites, call-in polls conducted by TV networks, phone surveys, and so on.

Statistical surveys

- If conducted properly, surveys can be very useful tools for getting information.
- However, if not conducted properly, surveys can result in bogus information.
- Some problems include improper wording of questions, which can be misleading, lack of response by people who were selected to participate, or failure to include an entire group of the population.
- These potential problems mean a survey has to be well thought out before it’s given.

Statistical experiments

- An *experiment* imposes one or more treatments on the participants in such a way that clear comparisons can be made. After the treatments are applied, the responses are recorded.
- For example, to study the effect of drug dosage on blood pressure, one group may take 10 mg of the drug, and another group may take 20 mg.
- When designed correctly, an experiment can help a researcher establish a cause-and-effect relationship if the difference in responses between the treatment group and the control group is statistically significant (unlikely to have occurred just by chance).

Statistical experiments

- Experiments are credited with helping to create and test drugs, determining best practices for making and preparing foods, and evaluating whether a new treatment can cure a disease, or at least reduce its impact.
- Our quality of life has certainly been improved through the use of well-designed experiments.
- However, not all experiments are well-designed, and your ability to determine which results are credible and which results are incredible (pun intended) is critical, especially when the findings are very important to you.

Types of Sampling

- Simple Random Sample
- Stratified Random Sample
- Cluster sampling
- Systematic
- Convenience

Math Alliance Project

Simple Random Sample

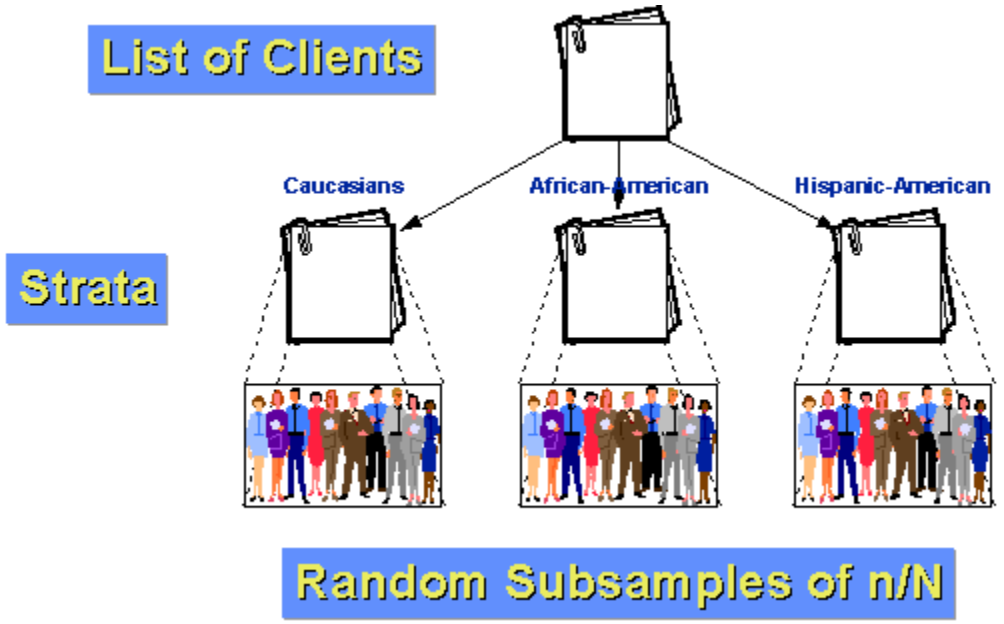
- Every subset of a specified size n from the population has an equal chance of being selected

A subset of the population.

Math Alliance Project

Stratified Random Sample

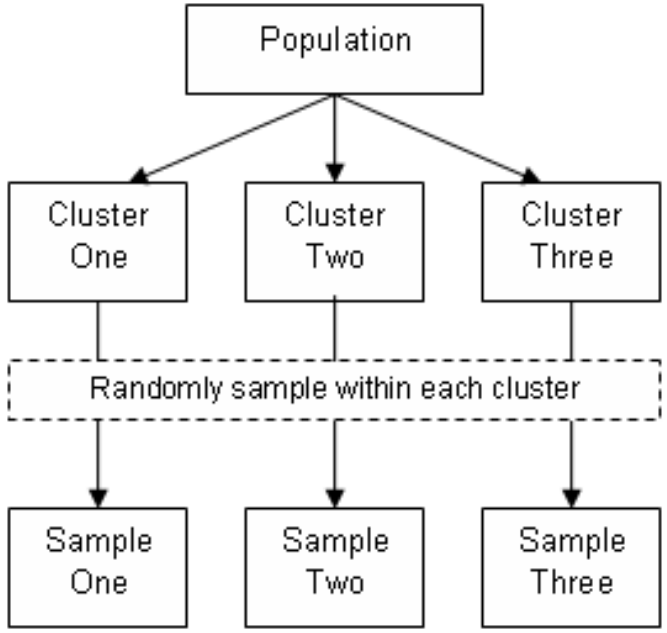
- The population is divided into two or more groups called strata, according to some criterion, such as geographic location, grade level, age, or income, and subsamples are randomly selected from each strata.



Math Alliance Project

Cluster Sample

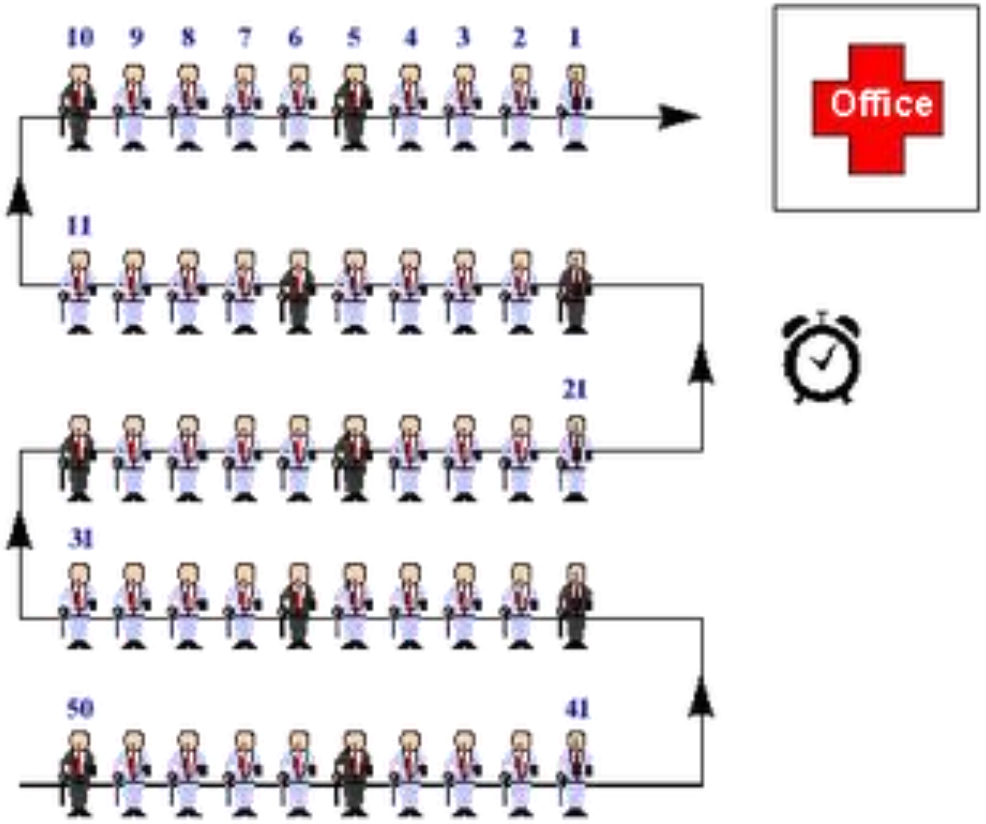
- The population is divided into subgroups (clusters) like families. A simple random sample is taken of the subgroups and then all members of the cluster selected are surveyed.



Math Alliance Project

Systematic Sample

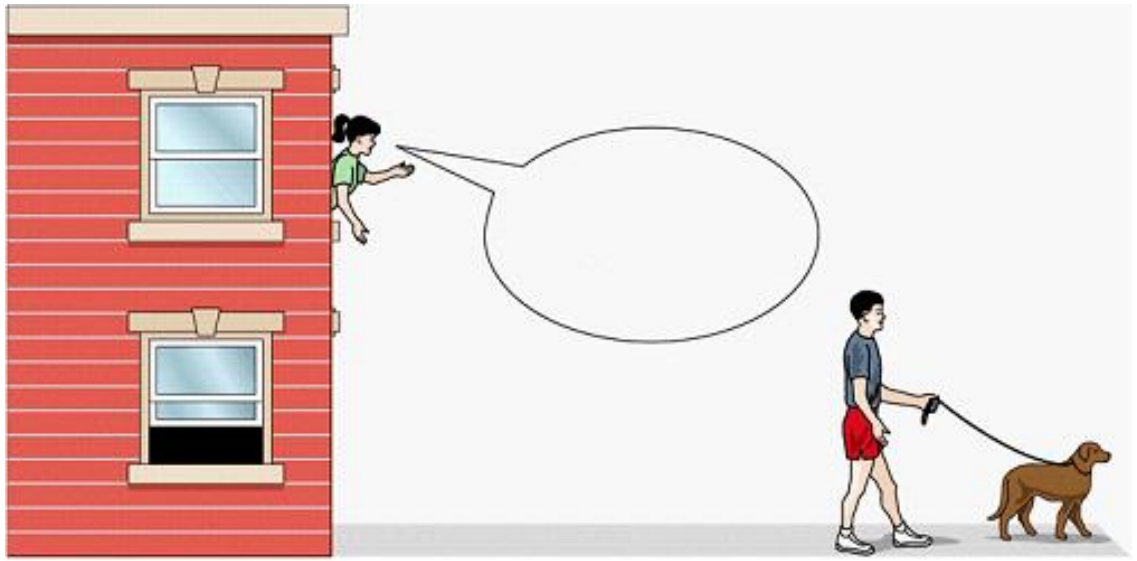
- Every kth member (for example: every 10th person) is selected from a list of all population members.



Math Alliance Project

Convenience Sample

- Selection of whichever individuals are easiest to reach
- It is done at the “convenience” of the researcher



Math Alliance Project

Analisis regresi

S E V E N T H E D I T I O N, Elementary Statistics, A Step by Step Approach
Allan G. Bluman, Professor Emeritus
Community College of Allegheny County

Analisis regresi

Do Dust Storms Affect Respiratory Health?

- SoutheastWashington state has a long history of seasonal dust storms. Several researchers decided to see what effect, if any, these storms had on the respiratory health of the people living in the area. They undertook (among other things) to see if there was a relationship between the amount of dust and sand particles in the air when the storms occur and the number of hospital emergency room visits for respiratory disorders at three community hospitals in southeast Washington. Using methods of correlation and regression, which are explained in this chapter, they were able to determine the effect of these dust storms on local residents. See Statistics Today—Revisited at the end of the chapter.

Source: B. Hefflin, B. Jalaludin, N. Cobb, C. Johnson, L. Jecha, and R. Etzel, "Surveillance for Dust Storms and Respiratory Diseases in Washington State, 1991," *Archives of Environmental Health 49, no. 3 (May–June 1994), pp. 170–74. Reprinted with* permission of the Helen Dwight Reid Education Foundation. Published by Heldref Publications, 1319 18th St. N.W., Washington, D.C. 20036-1802. Copyright 1994

Analisis regresi

Contoh lain:

- Hubungan antara biaya iklan dengan keuntungan perusahaan
- Hubungan antara uang saku per bulan dengan besarnya belanja pulsa
- Hubungan antara lamanya belajar dengan Indeks prestasi
- Hubungan antara lamanya mengakses internet dengan indeks prestasi

Clustering

- https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf

Clustering for Understanding

- Classes, or conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world.
- Indeed, human beings are skilled at dividing objects into groups (clustering) and assigning particular objects to these groups (classification).
- For example, even relatively young children can quickly label the objects in a photograph as buildings, vehicles, people, animals, plants, etc.
- In the context of understanding data, clusters are potential classes and cluster analysis is the study of techniques for automatically finding classes

Clustering

Biology.

- Biologists have spent many years creating a taxonomy (hierarchical classification) of all living things: kingdom, phylum, class, order, family, genus, and species

Business.

- Businesses collect large amounts of information on current and potential customers.
- Clustering can be used to segment customers

  into a small number of groups for additional analysis and marketing activities

Clustering

Information Retrieval.

- The World Wide Web consists of billions of Web pages, and the results of a query to a search engine can return thousands of pages.
- Clustering can be used to group these search results into a small number of clusters, each of which captures a particular aspect of the query.
- For instance, a query of “movie” might return Web pages grouped into categories such as reviews, trailers, stars, and theaters.
- Each category (cluster) can be broken into subcategories (sub-clusters), producing a hierarchical structure that further assists a user’s exploration of the query results

What Is Cluster Analysis?

- Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships.
- The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering.

Classification

- http://charuaggarwal.net/classbook.pdf
- The problem of data classification has numerous applications in a wide variety of mining applications. This is because the problem attempts to learn the relationship between a set of feature variables and a target variable of interest.
- Since many practical problems can be expressed as associations between feature and target variables, this provides a broad range of applicability of this model.
- The problem of classification may be stated as follows: Given a set of training data points along with associated training labels, determine the class label for an unlabeled test instance.

Classification

- Classification algorithms typically contain two phases:
  - Training Phase: In this phase, a model is constructed from the training instances.
  - Testing Phase: In this phase, the model is used to assign a label to an unlabeled test instance

- In some cases, such as lazy learning, the training phase is omitted entirely, and the classification is performed directly from the relationship of the training instances to the test instance.
- Instance-based methods such as the nearest neighbor classifiers are examples of such a scenario

Classification

Customer Target Marketing:

- Since the classification problem relates feature variables to target classes, this method is extremely popular for the problem of customer target marketing.
- In such cases, feature variables describing the customer may be used to predict their buying interests on the basis of previous training examples.
- The target variable may encode the buying interest of the customer.

Medical Disease Diagnosis:

- In recent years, the use of data mining methods in medical technology has gained increasing traction.
- The features may be extracted from the medical records, and the class labels correspond to whether or not a patient may pick up a disease in the future.
- In these cases, it is desirable to make disease predictions with the use of such information

Principal Component Analysis

- https://onlinecourses.science.psu.edu/stat505/node/49

- Sometimes data are collected on a large number of variables from a single population.
- With a large number of variables, the dispersion matrix may be too large to study and interpret properly. There would be too many pairwise correlations between the variables to consider. Graphical display of data may also not be of particular help incase the data set is very large. With 12 variables, for example, there will be more than 200 three-dimensional scatterplots to be studied!
- To interpret the data in a more meaningful form, it is therefore necessary to reduce the number of variables to a few, interpretable linear combinations of the data. Each linear combination will correspond to a principal component.

Elements of The Sampling Problem

Metode kuantitatif

Tim pengajar

Elements

- Technical term
- Sources of error in surveys
- Reducing errors in surveys
- Designing questionnaire
- Planning survey

Technical term

- An element is an object which a measurement is taken
- A population is a collection of elements about which we wish to make an inference
- Sampling units are nonoverlapping collections of elements from the population that cover the entire population
- A frame is a list of sampling units
- A sample is a collection of sampling units drawn from a frame or frames

Sources of error in surveys

- Survey errors can be divided into two major groups:
  - ***Error of nonobservation,*** where the sampled elements make up only part of the target population, and
  - ***errors*** of ***observation,*** where recorded data deviate from the truth.
- Errors of nonobservation can be attributed to sampling, coverage, or nonresponse.
- Errors of observation can be attributed to the interviewer (data collector), respondent, instrument, or method of data collection.

Reducing Errors in Surveys

- Both errors of nonobservation and errors of observation can seriously affect the accuracy of a survey.
- Errors can not be eliminated from a survey, but their effects can be reduced by careful adherence to a good sampling plan.

- **Callbacks**
- Nonresponse can be reduced by having a carefully prepared plan for callbacks on sampled elements. A fixed number of callback should be required for each sampled element, and these callbacks should be on different days of the week and at different hours of the day.

Reducing Errors in Surveys

Rewards and Incentives

- Sometimes, an appropriate tactic for encouraging responses is to offer a reward for responding.
- This reward may be a cash payment to a person who agrees to participate in a study.
- In studies of consumer products, a participant may be given a supply of the product.
- The rewards should be offered to potential participants in a study only after they have been selected for the sample by some objective procedure.

Reducing Errors in Surveys

Trained Interviewers

- The skill of the interviewer is directly related to the quantity and quality of data resulting from a survey, whether the interview is in person or over the telephone.
- Good interviewers can ask questions in such a way as to encourage honest responses and can tell the difference between those who really don't know the answer and those who are simply reluctant to answer.

Reducing Errors in Surveys

Data Checks

- Completed questionnaires should be scrutinized carefully by someone other than the interviewer to see that the form has been tilled out correctly.
- At this stage, and again later if data have been entered into a computer, a predesigned system of data checks should be made to spot obvious errors in information.

Questionnaire Construction

- After simple selection, the most important component of a well-run, informative, and accurate sample survey is a properly designed questionnaire.

Designing a Questionnaire

- As stated earlier, one objective of any survey design is to minimize the nonsampling errors that may occur.
- If a survey is to obtain information from people. then many potential nonsampling errors should be considered and, it is hoped, controlled by the careful design of the questionnaire.

Designing a Questionnaire

Question Ordering

- Respondents to questionnaires generally try to be consistent in their responses to questions.
- Respondent consistency may cause the ordering of the questions to affect the responses, somtimes in ways that seem unpredictable to the inexperienced investigator.

Designing a Questionnaire

An experiment was conducted with the following two questions:

A. Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?

B. Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?

Designing a Questionnaire

- For surveys in which the questions appeared in the order (A, B), 54.7% of the respondents answered yes to A and 63.7% answered yes to B.
- For surveys in which the questions appeared in the order (B, A), 74.6% answered yes to A and 81.9% answered yes to B.
- So the evidence suggests that asking question B first puts the respondents in a more lenient frame of mind toward allowing Communist reporters into the United States.

Designing a Questionnaire

Open versus Closed Questions

- Because questionnaires today are often designed to be electronically scored after completion. with the data in a form for computer handling, most questions are closed questions. That is, each question has either a single numerical answer (such as age of the respondent) or fixed number of predetermined choices, one of which is to be selected by the respondent.
- Eventhough closed questions allow for easy data coding and analysis, some thought should be given to open question, in which the respondent is allowed to freely state an unstructured answer.
- The open question allows the respondent to express some depth and shades of meaning in the answer.

Designing a Questionnaire

Response Options

- On almost any question that can be posted, someone being interviewed will want to say that he doesn't know or has no opinion.
- Because such responses give no useful information about the question and essentially reduce the sample size, typical survey practice is to avoid using these options.
- The respondent is forced to make a choice from among the listed informative answers, unless the interviewer decides that such a choice simply cannot be made.

Designing a Questionnaire

Wording of Questions

- Even for questions in which the number of response options is clearly determined, the designer should be concerned about the phrasing of the main body of the question.
  - Do you favor the use of capital punishment?
  - Do you favor or oppose the use of capital punishment?
- The second question is a more balanced form

Planning a Survey

1. ***Statement of objective***

   State the objectives of the survey clearly and concisely and refer to these objectives regularly as the design and the implementation of the survey progress. Keep the objectives simple enough to be understood by those working on the survey and to be met successfully when the survey is completed.

2. ***Target population***

   Carefully define the population to be sampled. If adults are to be sampled, then define what is meant by *adult* (all those over the age of **13,** for example) and state which group of adults are included (all permanent residents of a city, for example). Keep in mind that a sample must be selected from this population and define the population so that sample selection is possible.

Planning a Survey

3. ***The frame***

   Select the frame (or frames) so that the list of sampling units and the target population show close agreement. Keep in mind that multiple frames may make the sampling more efficient. For example, residents of a city can be sampled from a list of city blocks coupled with a list of residents within blocks.

4. ***Sample design***

   Choose the design of the sample, including the number of sample elemenls, so that the sample provides sufficient information for the objectives of the survey. Many surveys have produced little or no useful information because they were nor properly designed.

Planning a Survey

5. ***Method of measurement***

   Decide on the method of measurement, usually one or more of the following methods: personal interviews, telephone interviews, mailed questionnaires, or direct observations.

6. ***Measurement***

   In conjunction with step 5, carefully specify how and what measurements are to be obtained. If a questionnaire is to be used, plan the questions so that they minimize nonresponse and incorrect response bias.

Planning a Survey

7. ***Selection and training of field-workers***

   Carefully select and train fieldworkers. After the sampling plan has been clearly and completely set up, someone must collect the data. Those collecting data, the fieldworkers, must be carefully taught what rneasurements to make and how to make them. Training is especially important if interviews, either personal or telephone, are used because the rate of response and the accuracy of responses are affected by the interviewer's personal style and tone of voice.

8. ***The pretest***

   Select a small sample for a pretest. The pretest is crucial because it allows you to field-test the questionnaire or other measurement device, to screen interviewers, and to check on the management of field operations. The results of the pretest usually suggest that some modifications must be made before a fullscale sampling is undertaken.

Planning a Survey

9. **Organization *of field work***

   Plan the fieldwork in detail. Any large-scale survey involves numerous people working as interviewers, coordinators, or data managers. The various jobs should be carefully organized and lines of authority clearly established before the survey is begun.

10. ***Organization of data management***

    Outline how each datum is to be handled for all stages of the survey. Large surveys generate huge amounts of data. Hence, a well-prepared data management plan is of the utmost importance. This plan should include the steps for processing data from the time a measurement is taken in the field until the final analysis is completed. A quality control scheme should also be included in the plan in order to check for agreement between processed data and data gathered in the field.

Planning a Survey

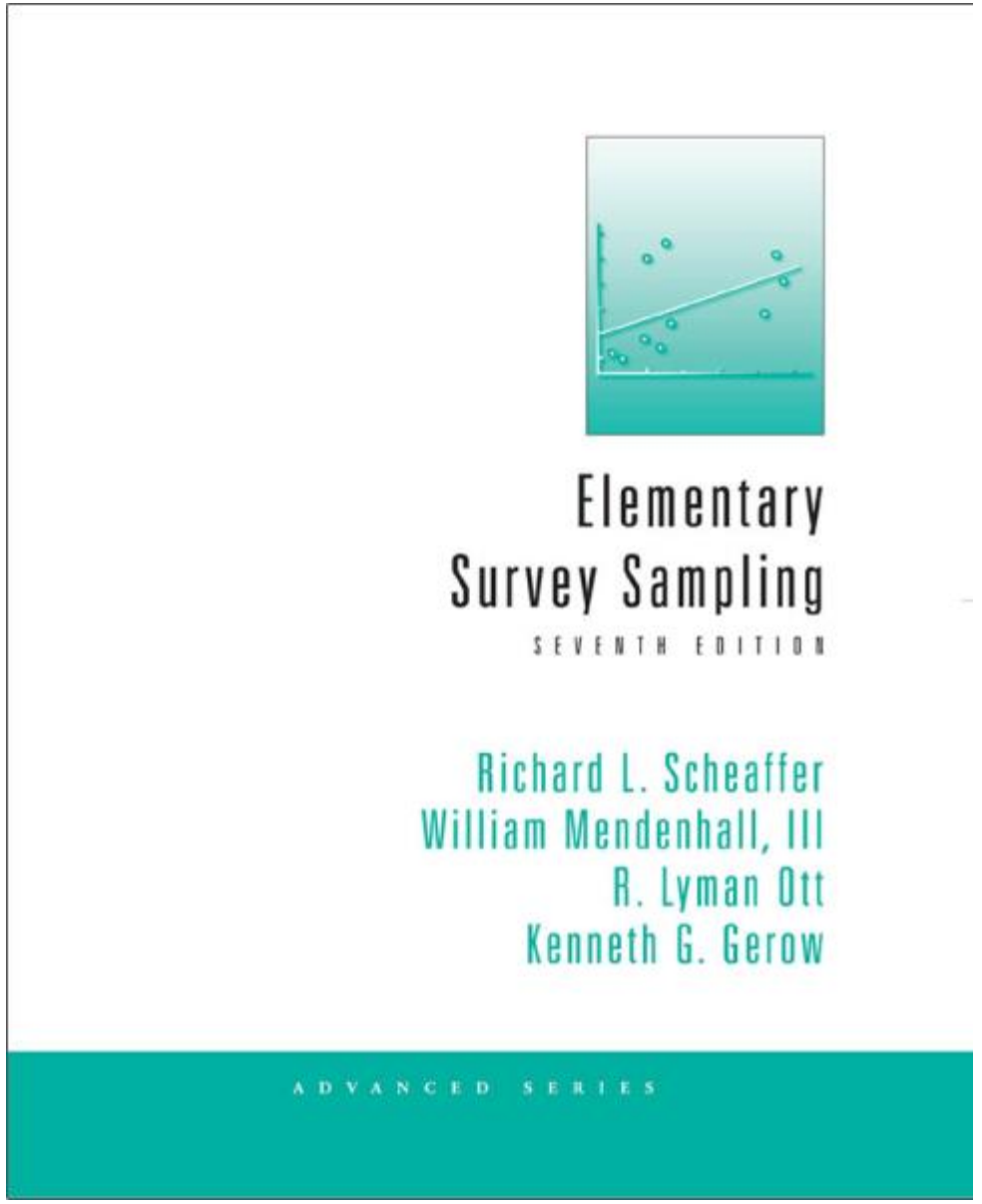
11. Data analysis

Outline the analyses that are to be completed. Closely related to step 10, this step involves the detailed specification of which analyses are to be performed. It may also list the topics to be included in the final report. If you think about the final report before a survey is run, you may be more careful in selecting items to be measured in the survey.

If these steps are followed diligently, the survey will be off to a good start and should provide useful information for the investigator.

Reference

- Scheaffer RL, William M, Lyman O. 2006. Elementary Survey Sampling sixth ed. PWS-KENT Publishing Company.



Simple random sampling

Metode kuantitatif

Tim pengajar

2022

Simple random sampling

- Introduction
- How to draw a simple random sampling
- Estimation of population mean and total
- Selecting the sample size for estimating population means and totals

Introduction

- The objective of a sample survey is to make an inference about population parameters from information contained in a sample.
- Two factors affect the quantity of information contained in the sample and hence the precision of our inference-making procedure.
  - The size of the sample selected from the population.
  - The amount of variation in the data; variation can frequently be controlled by the method of selecting the sample.

Introduction

- The procedure for selecting the sample is called the *sample survey design.*
- For a fixed sample size n, we will consider various designs, or sampling procedures, for obtaining the n observations in the sample.
- Because observations cost money, a design that provides a precise estimator of the parameter for a fixed sample size yields a savings in cost to the experimenter.
- The basic design, or sampling technique, is called simple random sampling

Introduction

- Definition 4.1
  - If a sample of size n is drawn from a population of size N in such a way that every possible sample of size *n* has the same chance of being selected, the sampling procedure is called *simple random sampling.* The sample thus obtained is called a *simple random sample.*
- All individual elements in a population have the same chance of being selected *and* that the selection of individual elements is mutually independent.

Introduction

- A federal auditor is to examine the accounts for a city hospital
- The hospital records obtained from a computer data file show a particular accounts receivable total, and the auditor must verify this total.
- If there are 28,000 open accounts in the hospital, the auditor cannot afford the time to examine every patient record to obtain a total accounts receivable figure.

Introduction

- Suppose that all N = 28,000 patient records are recorded in a computer file and a sample size n= 100 is to be drawn.
- The sample is called a simple random sample if every possible sample of *n* = 100 records has the same chance of being selected.

Introduction

Simple random sampling, which forms the bases of most sampling designs discussed in this book, is the foundation of most scientific surveys.

Introduction

- Marketing research often involves a simple random sample of potential users of a product. The researcher may want to estimate the proportion of potential buyers who prefer a certain color of car or flavor of food
- A forester may estimate the volume of timber or proportion of diseased trees in a forest by selecting geographic points in the area covered by the forest and then attaching a plot of fixed size and shape (such as a circle of 10-meter radius) to that point. All the trees within the sample plots may be studied, but, again, the basic design is a simple random sample.

How to draw simple random sample

A table of random numbers is shown in Appendix A, Table 2

Line/ Col.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
1	10480	15011	01536	02011	81647	91646	69179	14194	62590	36207	20969	99570	91291	90700
2	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
3	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
4	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
6	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
7	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
8	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
9	89579	14342	63661	10281	17453	18103	57740	84378	25331	12565	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
11	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
12	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
13	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
14	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
16	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
17	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
18	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
19	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
21	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
22	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
23	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
24	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
26	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
27	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
28	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
29	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
31	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
32	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
33	69011	65795	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
34	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
36	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
37	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	82517	36103	42791
38	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
39	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
41	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
42	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
43	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
44	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583

How to draw simple random sample

- Simple random samples can be selected by using tables of random numbers.
- A random number table is a set of integers generated so that in the long run the table will contain all ten integers *(0.* 1.. . . . 9) in approximetely equal proportions, with no trends in the pattern in which the digits were generated.
- Thus, if one number is selected from a random point in the table, it is equally likely to be any of the digits 0 through 9.

How to draw simple random sample

- Choosing numbers from the table is analogous to drawing numbers out of a hat containing those numbers on thoroughly mixed pieces of paper.
- Suppose we want a simple random sample of three people to be selected from seven.
- We could number the people from 1 to 7, put slips of paper containing these numbers (one number to a slip) into a hat, mix them, and draw out three, *without replacing* the drawn numbers.

Example 4.1

- For simplicity, assume there are N = 1000 patient records from which a simple random sample of n = 20 is to be drawn.
- The digits in Appendix A, Table 2 is generated to satisfy the conditions of simple random sampling.
- **Determine which records are to be included in a sample of size n = 20!**

Solution

TABLE 4.1

Patient records, to be included in the sample

104	**779**	**289**	**510**
223	**995**	**635**	**023**
241	**963**	**094**	**010**
421	**895**	**103**	**521**
375	**854**	**071**	**070**

Estimation of a Population Mean and Total

- We stated previously that the objective of survey sampling is to draw inferences about a population from information contained in a sample.
- One way to make inferences is to estimate certain population parameters by using the sample information.
- The objective of a sample survey is often to estimate a population mean, denoted μ or a population total, denoted by τ

Estimation of a Population Mean and Total

- Estimator of the population mean:

$$\hat{\mu} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- Estimator variance of $\bar{y}$:

$$\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) \text{ where } s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} {y_i}^2 - n\bar{y}}{n-1}$$

- Bound on the error of estimation

$$2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$$

Example 4.2

- Refer to the hospital audit in Example 4.1 and suppose that a random sample of n = 200 accounts is selected from the total of N = 1000.
- The sample mean of the accounts is found to be $\bar{y}$ = \$94.22, and the sample variance is s^2 = 415.21.
- Estimate μ the average due for all 1000 hospital accounts, and place a bound on the error of estimation!

Solution

- We use $\bar{y}$ = \$94.22 to estimate μ.
- A bound on the error of estimation can be found by using
- $2\sqrt{\hat{V}(\bar{y})} = 2\sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{415.21}{200}\left(\frac{1000-200}{1000}\right)} = \$2.58$
- Thus, we estimate the mean value per account, μ, to be $\bar{y}$ =\$94.22.
- Because *n* is large, the sample mean should possess approximately a normal distribution, so that \$94.22 $\pm$ \$2.58 is approximately a 95% confidence interval for the population mean

Estimation of population total

- Estimator of the population total:

$$\hat{\tau} = N\bar{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$$

- Estimator variance of $\hat{\tau}$:

$$\hat{V}(\hat{\tau}) = \hat{V}(N\bar{y}) = N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right) \text{ where } s^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$$

- Bound on the error of estimation

$$2\sqrt{\hat{V}(N\bar{y})} = 2\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)}$$

Example 4.4

- An industrial firm is concerned about the time spent each week by scientists on certain trivial tasks.
- The time-log sheets of a simple random sample of n = 50 employees show the average amount of time spent on these tasks is 10.31 hours, with a sample variances s^2 = 2.25. The company employs N = 750 scientists.
- Estimate the total number of worker-hours lost each week on trivial tasks and place a bound on the error of estimation.

Solution

- ***N*** = 750 employees with n = 50 and $\bar{y}$ = 10.31 hours/week

  $\hat{\tau} = N\bar{y} =$ 750(10.31) = 7732.5 hours

- To place a bound on the error of estimation, we apply

  - $2\sqrt{\hat{V}(N\bar{y})} = 2\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)}$ = $2\sqrt{750^2\left(\frac{2.25}{50}\right)\left(\frac{750-50}{750}\right)}$ = 307.4 hours

- Thus, the estimate of total time lost is $\hat{\tau} = 7732.5$ hours. We are reasonably confident that the error of estimation is less than 307.4 hours.

Selecting the Sample Size for Estimating Population Means and Totals

- At some point in the design of the survey, someone must make a decision about the size of the sample to be selected from the population.
- So far, we have discussed a sampling procedure (simple random sampling) but have said nothing about the number of observations to be included in the sample.
- The implications of such a decision are obvious. Observations cost money. Hence if the sample is too large, time and talent are wasted. Conversely, if the number of observations included in the sample is too small, we have bought inadequate information for the time and effort expended and have again been wasteful.

Selecting the Sample Size for Estimating Population Means and Totals

- Sample size required to estimate μ with a bound on the error of estimation ***B:***

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2} \quad \textit{where} \quad D = \frac{B^2}{4}$$

- Sample size required to estimate τ with a bound on the error of estimation ***B:***

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2} \quad \textit{where} \quad D = \frac{B^2}{4N^2}$$

Example 4.6

- An investigator is interested in estimating the total weight gain in 0 to 4 weeks for N = 1000 chicks fed on a new ration.
- Obviously, to weigh each bird would be time consuming and tedious. Therefore, determine the number of chicks to be sampled in this study in order to estimate τ with a bound on the error of estimation equal to 1000 grams.
- Many similar studies on chick nutrition have been run in the past. Using data from these studies. the investigator found that $\boldsymbol{\sigma^2}$, the population variance, was approximately equal to 36.00 $(\text{grams})^2$.
- Determine the required sample size.

Solution

- We can obtain an approximate sample size with $\boldsymbol{\sigma^2}$ equal to 36.00 and
  - $\boldsymbol{D = \frac{B^2}{4N^2}} = \frac{\mathbf{1000^2}}{\mathbf{4(1000)^2}} = 0.25$
  - $\boldsymbol{n} = \frac{\mathbf{1000(36.00)}}{\mathbf{(1000-1)(0.25)+36.00}} = 125.98$
- The investigator therefore needs to weigh n = 126 chicks to gain $\hat{\tau}$, the total weight for N = 1000 chickens in 0 to 1 weeks, with a bound on the error of estimation equal to 1000 grams.

Estimation of population proportion

- The investigator conducting a sample survey is frequently interested in estimating the proportion of the population that possesses a specified characteristic.
- A congressional leader investigating the merits of an 18-year-old voting age may want to estimate the proportion of the potential voters in the district between the ages of 18 and 21.
- A marketing research group may be interested in the proportion of the total sales market in diet preparations that is attributable to a particular product. That is, what percentage of sales is accounted for by a particular product?
- A forest manager may he interested in the proportion of trees with a diameter of 12 inches or more.
- Television ratings are often determined by estimating the proportion of the viewing public that hatches a particular program.

- We denote the population proportion and its estimator by the symbols ***p*** and $\hat{p}$, respectively. The properties of $\hat{p}$ for simple random sampling parallel those of the sample mean $\bar{y}$ if the response measurements are defined as follows
- Let y_i= *0* if the ith element sampled does not possess the specified characteristic and y_i = 1 if it does. Then the total number of elements in a sample of size *n* possessing a specified characteristic is

$$\sum_{i=1}^{n} y$$

Estimator of the population proportion p:

$$\hat{p} = \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{4.14}$$

Estimated variance of $\hat{p}$:

$$\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right) \tag{4.15}$$

where

$$\hat{q} = 1 - \hat{p}$$

Bound on the error of estimation:

$$2\sqrt{\hat{V}(\hat{p})} = 2\sqrt{\frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)} \tag{4.16}$$

- A simple random sample of n = 100 college seniors was selected to estimate
  - (1) the fraction of N = 300 seniors going on to graduate school and
  - (2) the fraction of students that have held palt-time jobs during college.
- Let y_i and x_i (i = 1, 2. . . . , 100) denote the responses of the ith student sampled.
  - y_i = 0 if the ith student does not plan to attend graduate school
  - y_i = 1 if he or she does.
- Similarly, let
  - x_i = 0 if he or she has not held a part-time job sometime during college
  - x_i = 1 if he or she has
- Estimate p_1, the proportion of seniors planning to attend graduate school, and p_2 the proportion of seniors who have had a part-time job sometime during their college careers (summers included).

Student	y	x
1	1	0
2	0	1
3	0	1
4	1	1
5	0	0
6	0	0
7	0	1
.	.	.
.	.	.
.	.	.
96	0	1
97	1	0
98	0	1
99	0	1
100	1	1
	$\sum_{i=1}^{100} y_i = 15$	$\sum_{i=1}^{100} x_i = 65$

The sample proportions from Eq. (4.14) are given by

$$\hat{p}_1 = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{15}{100} = 0.15$$

and

$$\hat{p}_2 = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{65}{100} = 0.65$$

The bounds on the errors of estimation of p_1 and p_2 are, respectively,

$$2\sqrt{\hat{V}(\hat{p}_1)} = 2\sqrt{\frac{\hat{p}_1\hat{q}_1}{n-1}\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{(0.15)(0.85)}{99}\left(\frac{300-100}{300}\right)}$$
$$= 2(0.0293) = 0.059$$

and

$$2\sqrt{\hat{V}(\hat{p}_2)} = 2\sqrt{\frac{\hat{p}_2\hat{q}_2}{n-1}\left(\frac{N-n}{N}\right)}$$

$$= 2\sqrt{\frac{(0.65)(0.35)}{99}\left(\frac{300-100}{300}\right)} = 2(0.0391) = 0.078$$

Thus, we estimate that 0.15 (15%) of the seniors plan to attend graduate school, with a bound on the error of estimation equal to .059 (5.9%). We estimate that 0.65 (65%) of the seniors have held a part-time job during college, with a bound on the error of estimation equal to .078 (7.8%). ■

Selecting the Sample Size for Estimating Proportion

- In a practical situation, we do not know *p*. An approximate sample size can be found by replacing ***p*** with an estimated value. Frequently, such an estimate can be obtained from similar past surveys. However, if no such prior information is available, we can substitute $p = 0.5$ into Eq. (4.18) to obtain a conservative sample size (one that is likely to be larger than required).

Sample size required to estimate p with a bound on the error of estimation B:

$$n = \frac{Npq}{(N-1)D + pq} \qquad (4.18)$$

where

$$q = 1 - p \qquad \text{and} \qquad D = \frac{B^2}{4}$$

Example 4.8

- Student government leaders at a college want to conduct a survey to determine the proportion of students who favor a proposed honor code.
- Interviewing N = 2000 student in a reasonable length of time is almost impossible.
- Determine the sample size (number of students to be interviewed) needed to estimate p
- Bound on the error of estimation of magnitude B = 0.05. Assume that no prior information is available to estimate $\boldsymbol{p}$.

Solution

We can approximate the required sample sizes when no prior information is available by setting $p = 0.5$ in Eq. (4.18). We have

$$D = \frac{B^2}{4} = \frac{(0.05)^2}{4} = 0.000625$$

Hence,

$$\begin{aligned} n &= \frac{Npq}{(N-1)D + pq} \\ &= \frac{(2000)(0.5)(0.5)}{(1999)(0.000625) + (0.5)(0.5)} = \frac{500}{1.499} \\ &= 333.56 \end{aligned}$$

That is, 334 students must be interviewed to estimate the proportion of students who favor the proposed honor code with a bound on the error of estimation of $B = 0.05$. ■

Stratified Random Sampling

Metode kuantitatif

Tim pengajar

2022

Stratified Random Sampling

- Introduction
- How to draw a stratified random sampling
- Estimation of population mean and total
- Selecting the sample size for estimating population means and totals

Introduction

- *A stratified random sample* is one obtained by separating the population elements into **non-overlapping groups**, called strata, and then selecting a simple random sample from each stratum.

Introduction

- Suppose a public opinion poll designed to estimate the **proportion** of voters who favor spending more **tax revenue** on an improved **ambulance service** is to be conducted in a certain county.
- The county contains **two cities** and a rural area. The population elements of interest for the poll are all men and women of voting age who reside in the county.
- A stratified random sample of adults residing in the county can be obtained by selecting a **simple random sample of adults from each city** and another simple random sample of adults from the rural area.

Introduction

- Why should we choose a stratified random sample rather than a simple random sample?
  1. Our goal in designing surveys is to maximize the information obtained (or to minimize the bound on the error of estimation) for a fixed expenditure. The samples display small variability among the measurements will produce small bounds on the errors of estimation.
  2. The cost of obtaining observations varies with the design of the survey. The cost of selecting the adults to be sampled, the cost of interviewer time and travel, and the cost of administering the overall sampling procedure may all be minimized by a carefully planned stratified random sample in compact, well-defined geographic areas.

Introduction

3. Estimates of a population parameter may be desired for certain subsets of the population. In the country poll, each city commission may want to see an estimate of the proportion of voters favoring an expanded anmbulances service for its own city.

- Stratified random sampling allows for separate estimates of population parameters within each stratum.

Introduction

In summary, the principal reasons for using stratified random sampling rather than simple random sampling are as follows:

1. Stratification may produce a smaller bound on the error of estimation than would be produced by a simple random sample of the same size. This result is particularly true if measurements within strata are homogeneous.
2. The cost per observation in the survey may be reduced by stratification of the population elements into convenient groupings.
3. Estimates of population parameters may be desired for subgroups of the population. These subgroups should then be identifiable strata

Introduction

- The consumer price index (CPI) is a measure of the average change in prices for a fixed collection of goods and services for urban consumers.
- The CPI is actually calculated from at least four different types of surveys: surveys of cities, surveys of urban families, surveys of outlets providing goods and services, and surveys of specific goods and services.
- In the design of most CPI surveys, sampling units (counties or groups of contiguous counties) are identified in the population and then grouped into strata.
- Strata are chosen on the basis of geography, population size, rate of population increase, major industry, percentage nonwhite, and percentage urban.
- The sampling units within a stratum are chosen to be as much alike as possible with regard to these characteristics.

How to draw a stratified random sample

- The first step in the selection of a stratified random sample is to clearly specify the strata; then each sampling unit of the population is placed into its appropriate stratum.
- After the sampling units have been divided into strata, we select a simple random sample from each stratum.
- Some additional notation is required for stratified random sampling. Let
  - L = number of strata
  - Ni = number of sampling units in stratum i
  - N = Number of sampling units in the population ($N=N1 + N2 +... + N_L$)

Estimation of a population mean and total

Estimator of the population mean μ:

$$\bar{y}_{st} = \frac{1}{N}[N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L] = \frac{1}{N}\sum_{i=1}^{L} N_i\bar{y}_i \qquad \textbf{(5.1)}$$

Estimated variance of $\bar{y}_{st}$:

$$\begin{aligned}\hat{V}(\bar{y}_{st}) &= \frac{1}{N^2}\left[N_1^2\hat{V}(\bar{y}_1) + N_2^2\hat{V}(\bar{y}_2) + \cdots + N_L^2\hat{V}(\bar{y}_L)\right] \\ &= \frac{1}{N^2}\left[N_1^2\left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{s_1^2}{n_1}\right) + \cdots \right. \\ &\quad \left. + N_L^2\left(\frac{N_L - n_L}{N_L}\right)\left(\frac{s_L^2}{n_L}\right)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right) \qquad \textbf{(5.2)}\end{aligned}$$

Example 5.1

- An advertising firm, interested in determining how much to emphasize television advertising in a certain county, decides to conduct a sample survey to estimate the average number of hours each week that households within the county watch television. The county contains two towns, A and B and a rural area.
- Town A is built around a factory, and most households contain factory workers with school-age children.
- Town B is an exclusive suburb of a city in a neighboring county and contains older residents with few children at home.
- There are 155 households in town A, 62 in town B, and 93 in the rural area.

Example 5.2

- Suppose the survey is carried out. The advertising firm has enough time and money to interview **n= 40** households and decides to select random samples of size ***n1* = 20 from town A, n2 = 8 from town B, and n3 = 12 from the rural area**. The simple random samples are selected and the interviews conducted. The results, with measurements of television-viewing time in hours per week, are shown in Table 5.1. **Estimate the average television-viewing time, in hours per week**, for (a) all households in the county and (b) all households in town B. In both cases, place a **bound on the error of estimation**.

Example 5.2

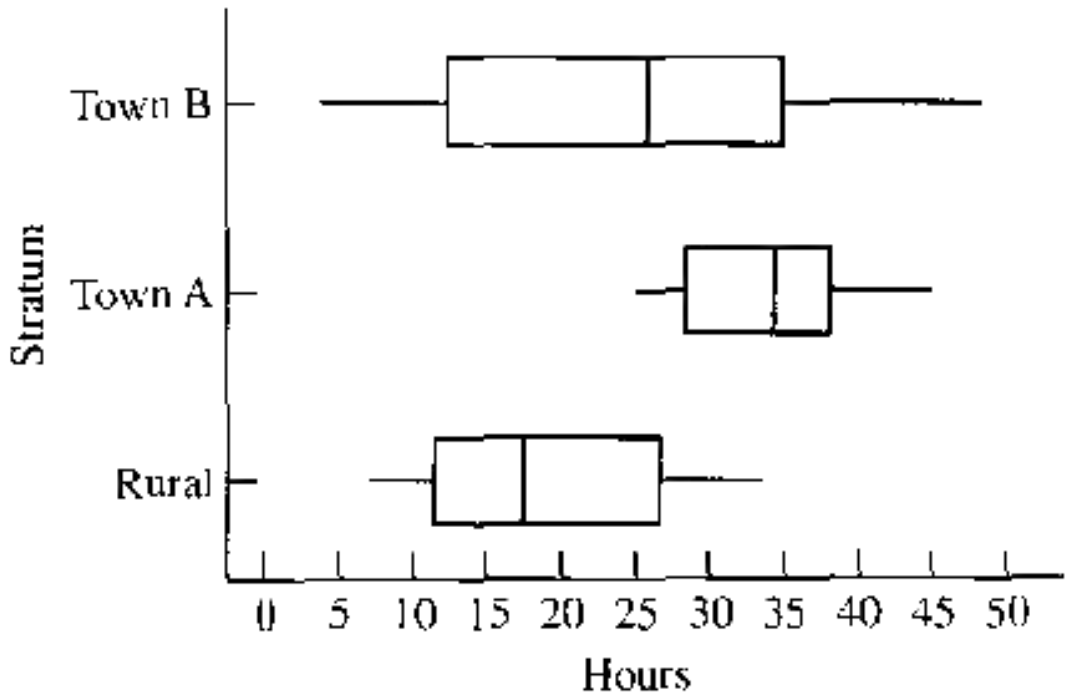
- Television viewing time, in hours per week

Stratum 1	Stratum 2	Stratum 3
Town A	Town B	Rural area
35 28 26 41	27 4 49 10	8 15 21 7
43 29 32 37	15 41 25 30	14 30 20 11
36 25 29 31		12 32 34 24
39 38 40 45		
28 27 35 34		

Example 5.2

FIGURE **5.1**

Box plots of television-viewing time



Example 5.2

TABLE **5.2**
Summary of the data from Table 5.1

	n	Mean	Median	SD
Town A	20	33.90	34.50	5.95
Town B	8	25.12	26.00	15.25
Rural	12	19.00	17.50	9.36

$$\bar{y}_{st} = \frac{1}{N}[N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L]$$
$$= \frac{1}{310}[(155)(33.90) + (62)(25.12) + (93)(19.00)]$$
$$= 27.7$$

$$\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$$
$$= \frac{1}{(310)^2}\left[\frac{(155)^2(0.871)(5.95)^2}{20} + \frac{(62)^2(0.871)(15.25)^2}{8} + \frac{(93)^2(0.871)(9.36)^2}{12}\right]$$
$$= 1.97$$

$$\bar{y}_{st} \pm 2\sqrt{\hat{V}(\bar{y}_{st})} \quad \text{or} \quad 27.675 \pm 2\sqrt{1.97} \quad \text{or} \quad 27.7 \pm 2.8$$

$$\bar{y}_2 \pm 2\sqrt{\left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{s_2^2}{n_2}\right)} \quad \text{or} \quad 25.1 \pm 2\sqrt{\left(\frac{62-8}{62}\right)\left(\frac{232.56}{8}\right)}$$
$$\text{or} \quad 25.1 \pm 10.0$$

Population total estimation

Estimator of the population total τ:

$$N\bar{y}_{st} = N_1\bar{y}_1 + N_2\bar{y}_2 + \cdots + N_L\bar{y}_L = \sum_{i=1}^{L} N_i\bar{y}_i \qquad (5.3)$$

Estimated variance of $N\bar{y}_{st}$:

$$\hat{V}(N\bar{y}_{st}) = N^2\hat{V}(\bar{y}_{st}) = \sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right) \qquad (5.4)$$

Example 5.3

Refer to Example 5.2 and estimate the total number of hours each week that households in the county view television. Place a bound on the error of estimation.

For the data in Table 5.1,

$$N\bar{y}_{st} = 310(27.7) = 8587 \text{ hours}$$

The estimated variance of $N\bar{y}_{st}$ is given by

$$\hat{V}(N\bar{y}_{st}) = N^2\hat{V}(\bar{y}_{st}) = (310)^2(1.97) = 189,278.560$$

The estimate of the population total with a bound on the error of estimation is given by

$$N\bar{y}_{st} \pm 2\sqrt{\hat{V}(N\bar{y}_{st})} \quad \text{or} \quad 8587 \pm 2\sqrt{189,278.560}$$

$$\text{or} \quad 8587 \pm 870$$

Thus, we estimate the total weekly viewing time for households in the county to be 8587 hours. The error of estimation should be less than 870 hours. ■

Selecting the sample size for estimating population means and totals

- To be

Approximate sample size required to estimate μ or τ with a bound B on the error of estimation:

$$n = \frac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \qquad (5.6)$$

where a_i is the fraction of observations allocated to stratum i, σ_i^2 is the population variance for stratum i, and

$$D = \frac{B^2}{4} \qquad \text{when estimating } \mu$$

$$D = \frac{B^2}{4N^2} \qquad \text{when estimating } \tau$$

- We must obtain approximations of the population variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_L^2$ before we can use Eq. *(5.6).* One method of obtaining these approximations is to use the sample variances $s_1^2, s_2^2, \ldots, s_L^2$ from a previous experiment to estimate $\sigma_1^2, \sigma_2^2, \ldots, \sigma_L^2$.

Example 5.1

- An advertising firm, interested in determining how much to emphasize television advertising in a certain county, decides to conduct a sample survey to estimate the average number of hours each week that households within the county watch television. The county contains two towns, A and B and a rural area.
- Town A is built around a factory, and most households contain factory workers with school-age children.
- Town B is an exclusive suburb of a city in a neighboring county and contains older residents with few children at home.
- There are 155 households in town A, 62 in town B, and 93 in the rural area.

Example 5.5

- A prior survey suggests that the stratum variances for Example 5.1 are approximately $\sigma_1^2 \approx 25, \sigma_2^2 \approx 225, \sigma_3^2 \approx 100$. We wish to estimate the population mean by using $\bar{y}_{st}$. Choose the sample size to obtain a bound on the error of estimation equal to 2 hours if the allocation fractions are given by using $a_1 = {}^1/_3, a_2 = {}^1/_3, a_3 = {}^1/_3$. In other words, you are to take an equal number of observations from each stratum.

Solution

- A bound on the error of 2 hours means that
- $2\sqrt{V(\bar{y}_{st})} = 2$ or $V(\bar{y}_{st}) = 1$
- Therefore, D = 1

In Example 5.1, $N_1 = 155$, $N_2 = 62$, and $N_3 = 93$. Therefore,

$$\sum_{i=1}^{3} \frac{N_i^2 \sigma_i^2}{a_i} = \frac{N_1^2 \sigma_1^2}{a_1} + \frac{N_2^2 \sigma_2^2}{a_2} + \frac{N_3^2 \sigma_3^2}{a_3}$$

$$= \frac{(155)^2(25)}{(1/3)} + \frac{(62)^2(225)}{(1/3)} + \frac{(93)^2(100)}{(1/3)}$$

$$= (24{,}025)(75) + (3844)(675) + (8649)(300)$$

$$= 6{,}991{,}275$$

$$\sum_{i=1}^{3} N_i \sigma_i^2 = N_1 \sigma_1^2 + N_2 \sigma_2^2 + N_3 \sigma_3^2$$

$$= (155)(25) + (62)(225) + (93)(100) = 27{,}125$$

$$N^2 D = (310)^2(1) = 96{,}100$$

From Eq. (5.6), we then have

$$n = \frac{\sum_{i=1}^{3} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2} = \frac{6{,}991{,}275}{96{,}100 + 27{,}125} = \frac{6{,}991{,}275}{123{,}225} = 56.7$$

Thus, the experimenter should take $n = 57$ observations with

$$n_1 = n(a_1) = 57\left(\frac{1}{3}\right) = 19$$

$$n_2 = 19$$

$$n_3 = 19$$

■

Example 5.6

- As in Example 5.5, suppose the variances of Example 5.1 are approximated by $\sigma_1^2 \approx 25, \sigma_2^2 \approx 225, \sigma_3^2 \approx 100$. We wish to estimate the population total τ with a bound of 400 hours on the error of estimation. Choose the appropriate sample size if an equal number of observations is to be taken from each stratum.

Solution

- The bound on the error of estimation is to be 400 hours and, therefore,

- $D = \frac{B^2}{4N^2} = \frac{400^2}{4N^2} = \frac{40000}{N^2}$

To calculate n from Eq. (5.6), we need the following quantities:

$$\sum_{i=1}^{3} \frac{N_i^2 \sigma_i^2}{a_i} = 6{,}991{,}275 \qquad \text{(from Example 5.5)}$$

$$\sum_{i=1}^{3} N_i \sigma_i^2 = 27{,}125 \qquad \text{(from Example 5.5)}$$

$$N^2 D = N^2 \left(\frac{40{,}000}{N^2} \right) = 40{,}000$$

Using Eq. (5.6) yields

$$\frac{\sum_{i=1}^{3} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2} = \frac{6{,}991{,}275}{40{,}000 + 27{,}125} = 104.2 \approx 105$$

Then $n_1 = n_2 = n_3 = 35$. ■

Allocation of sample

- Recall that **the objective of a sample survey** design is to provide estimators with **small variances at the lowest possible cost**.
- After the sample size ***n*** is chosen, there are many ways to divide ***n*** into the individual stratum sample sizes $n_1, n_2, \ldots, n_L$
- Each division may result in a different variance for the sample mean.
- Hence, our **objective** is to use an allocation that gives a specified amount of information at **minimum cost**.

Allocation of sample

In terms of our objective, the best allocation scheme is affected by three factors:

1. The total number of elements in each stratum.
2. The variability of observations within each stratum.
3. The cost of obtaining an observation from each stratum.

Allocation of sample

- The number of elements in each stratum affects the quantity of information in the sample.
- A sample size 20 from a population of 200 elements should contain more information than a sample of 20 from 20.000 elements.
- Thus, large-sample sizes should be assigned to strata containing large numbers of elements.
- Variability must be considered because a larger sample is needed to obtain a good estimate of a population parameter when the observations are less homogeneous.
- If the cost of obtaining an observation varies from stratum to stratum, we take small samples from strata with high costs. We do so because our objective is to keep the cost of sampling at a minimum.

Allocation of sample

Approximate allocation that minimizes cost for a fixed value of $V(\bar{y}_{st})$ or minimizes $V(\bar{y}_{st})$ for a fixed cost:

$$n_i = n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{N_1\sigma_1/\sqrt{c_1} + N_2\sigma_2/\sqrt{c_2} + \cdots + N_L\sigma_L/\sqrt{c_L}}\right) \qquad \textbf{(5.7)}$$

$$= n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right)$$

where N_i denotes the size of the ith stratum, σ_i^2 denotes the population variance for the ith stratum, and c_i denotes the cost of obtaining a single observation from the ith stratum. Note that n_i is directly proportional to N_i and σ_i and inversely proportional to $\sqrt{c_i}$.

Allocation of sample

- We must approximate the variance of each stratum before sampling in order to use the allocation formula Eq. (5.7).
- The approximations can be obtained from earlier surveys or from knowledge of the range of the measurements within each stratum.

Allocation of sample

- Substituting the n_i/n given by formula (5.7) for a_i in Eq. (5.6) gives

$$n = \frac{\left(\sum_{k=1}^{L} N_k \sigma_k / \sqrt{c_k}\right)\left(\sum_{i=1}^{L} N_i \sigma_i \sqrt{c_i}\right)}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \qquad \textbf{(5.8)}$$

- for optimal allocation with the variance of $\bar{y}_{st}$ fixed at *D*

Example 5.7

- The advertising firm in Example 5.1 finds that obtaining an observation from a rural household costs more than obtaining a response in town A or B. The increase is due to the costs of traveling from one rural household to another.
- The cost per observation in each town is estimated to be \$9 (that is $c_1 = c_2 = \$9$, and the costs per observation in the rural area to be \$16 (that is $c_3 = \$16$).
- The stratum standard deviations (approximated by the strata sample variances from a prior survey) are $\sigma_1 \approx 5, \sigma_2 \approx 15, \sigma_3 \approx 10$.
- Find the overall sample size *n* and the stratum sample sizes n_1, n_2, n_3 that allow the firm to estimate, at minimum cost, the average television-viewing time with a bound on the error of estimation equal to 2 hours.

Solution

We have

$$\sum_{k=1}^{3} \frac{N_k \sigma_k}{\sqrt{c_k}} = \frac{N_1 \sigma_1}{\sqrt{c_1}} + \frac{N_2 \sigma_2}{\sqrt{c_2}} + \frac{N_3 \sigma_3}{\sqrt{c_3}}$$

$$= \frac{155(5)}{\sqrt{9}} + \frac{62(15)}{\sqrt{9}} + \frac{93(10)}{\sqrt{16}} = 800.83$$

and

$$\sum_{i=1}^{3} N_i \sigma_i \sqrt{c_i} = N_1 \sigma_1 \sqrt{c_1} + N_2 \sigma_2 \sqrt{c_2} + N_3 \sigma_3 \sqrt{c_3}$$

$$= 155(5)\sqrt{9} + 62(15)\sqrt{9} + 93(10)\sqrt{16} = 8835$$

Thus

$$n = \frac{\left(\sum_{k=1}^{3} N_k \sigma_k / \sqrt{c_k}\right)\left(\sum_{i=1}^{3} N_i \sigma_i \sqrt{c_i}\right)}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2}$$

$$= \frac{(800.83)(8835)}{(310)^2(1) + 27,125} = 57.42 \approx 58$$

Then

$$n_1 = n\left(\frac{N_1\sigma_1/\sqrt{c_1}}{\sum_{k=1}^{3} N_k\sigma_k/\sqrt{c_k}}\right) = n\left[\frac{155(5)/3}{800.83}\right] = 0.32n = 18.5 \approx 18$$

Similarly,

$$n_2 = n\left[\frac{62(15)/3}{800.83}\right] = 0.39n = 22.6 \approx 23$$

$$n_3 = n\left[\frac{93(10)/4}{800.83}\right] = 0.29n = 16.8 \approx 17$$

Hence, the experimenter should select 18 households at random from town A, 23 from town B, and 17 from the rural area. He or she can then estimate the average number of hours spent watching television at minimum cost with a bound of 2 hours on the error of estimation. ■

Allocation of sample

- In some stratified sampling problems, the cost of obtaining an observation is the same for all strata. If the costs are unknown, we may be willing to assume that the costs per observation are equal. If $c_1 = c_2 = \cdots = c_L$, then the cost terms cancel in Eq. (5.7) and

$$n_i = n\left(\frac{N_i\sigma_i}{\sum_{k=1}^{L} N_k\sigma_k}\right) \qquad (5.9)$$

Allocation of sample

- This method of selecting $n_1, n_2, \ldots, n_L$ is called *Neymann allocation*. Under *Neymann allocation*, Eq. (5.8) for the total sample size n becomes

$$n = \frac{\left(\sum_{k=1}^{L} N_k \sigma_k\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2} \qquad \textbf{(5.10)}$$

Example 5.8

- The advertising firm in Example 5.1 decides to use telephone interviews rather than personal interviews because all households in the country have telephones, and this method reduces costs. The cost of obtaining an observation is then the same in all three strata.
- The stratum standard deviations are again approximated $\sigma_1 \approx 5, \sigma_2 \approx 15, \sigma_3 \approx 10$. The firm desires to estimate the population mean μ with a bound on the error of estimation equal to 2 hours.
- Find the appropriate sample size n and stratum sample size n_1, n_2, n_3.

Solution

- We now use Eqs. (5.9) and (5.10) because the costs are the same in all strata. Therefore to find the allocation fractions, a_1, a_2, a_3,we use Eq. (5.9). Then

$$\sum_{i=1}^{3} N_i\sigma_i = N_1\sigma_1 + N_2\sigma_2 + N_3\sigma_3$$

$$= (155)(5) + (62)(15) + (93)(10) = 2635$$

Allocation of sample

and from Eq. (5.9)

$$n_1 = n\left(\frac{N_1\sigma_1}{\sum_{i=1}^{3} N_i\sigma_i}\right) = n\left[\frac{(155)(5)}{2635}\right] = n(0.30)$$

Similarly,

$$n_2 = n\left[\frac{62(15)}{2635}\right] = n(0.35)$$

$$n_3 = n\left[\frac{93(10)}{2635}\right] = n(0.35)$$

Thus, $a_1 = 0.30$, $a_2 = 0.35$, and $a_3 = 0.35$.

Now let us use Eq. (5.10) to find n. A bound of 2 hours on the error of estimation means that

$$2\sqrt{V(\bar{y}_{st})} = 2 \quad \text{or} \quad V(\bar{y}_{st}) = 1$$

Therefore,

$$D = \frac{B^2}{4} = 1 \quad \text{and} \quad N^2 D = (310)^2(1) = 96{,}100$$

Also, from Example 5.6,

$$\sum_{i=1}^{3} N_i \sigma_i^2 = 27{,}125$$

and Eq. (5.10) gives

$$n = \frac{\left(\sum_{i=1}^{3} N_i \sigma_i\right)^2}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2}$$

$$= \frac{(2635)^2}{96{,}100 + 27{,}125} = 56.34 \approx 57$$

Allocation of sample

and Eq. (5.10) gives

$$n = \frac{\left(\sum_{i=1}^{3} N_i \sigma_i\right)^2}{N^2 D + \sum_{i=1}^{3} N_i \sigma_i^2}$$

$$= \frac{(2635)^2}{96,100 + 27,125} = 56.34 \approx 57$$

Then

$$n_1 = na_1 = (57)(0.30) = 17$$
$$n_2 = na_2 = (57)(0.35) = 20$$
$$n_3 = na_3 = (57)(0.35) = 20 \quad \blacksquare$$

- The sample size *n* in Example 5.8 is nearly the same as in Example 5.7, but the allocation has changed. More observations are taken from the rural area because these observations no longer have a higher cost.

Allocation of sample

- In the television-viewing example, suppose the costs are as specified in Example 5.7. That is, $c_1 = c_2 = \$9, c_3 = \$16$. Let the stratum standard deviations be approximated by $\sigma_1 \approx 5, \sigma_2 \approx 15, \sigma_3 \approx 10$. Given that the advertising firm has only \$500 to spend on sampling, choose the sample size and the allocation that minimize $V(\bar{y}_{st})$.

The allocation scheme is still given by Eq. (5.7). In Example 5.7, we find $a_1 = 0.32$, $a_2 = 0.39$, and $a_3 = 0.29$.

Because the total cost must equal \$500, we have

$$c_1 n_1 + c_2 n_2 + c_3 n_3 = 500$$
$$9n_1 + 9n_2 + 16n_3 = 500$$

Because $n_i = na_i$, we can substitute as follows:

$$9na_1 + 9na_2 + 16na_3 = 500$$
$$9n(0.32) + 9n(0.39) + 16n(0.29) = 500$$

Solving for n, we obtain

$$11.03n = 500$$
$$n = \frac{500}{11.03} = 45.33$$

Therefore, we must take $n = 45$ to ensure that the cost remains below \$500. The corresponding allocation is given by

$$n_1 = na_1 = (45)(0.32) = 14$$
$$n_2 = na_2 = (45)(0.39) = 18$$
$$n_3 = na_3 = (45)(0.29) = 13 \quad \blacksquare$$

Estimation of a Population Proportion

- In our numerical examples, we have been interested in estimating the average or the total number of hours per week spent watching television. In contrast, suppose that the advertising firm wants to estimate the proportion (fraction) of households that watches a particular show.
- The population is divided into strata, just as before and a simple random sample is taken from each stratum. Interviews are then conducted to determine the proportion $\hat{p}_i$, of households in stratum i that view the show.
- This $\hat{p}_i$, is an unbiased estimator of p_i, the population proportion in stratum i (as described in Chapter 4).

Estimation of a Population Proportion

- Reasoning as we did in Section 5.3, we conclude that $N_i\hat{p}_i$is an unbiased estimator of the total number of households in stratum i that view this particular show.
- Hence, $N_1\hat{p}_1 + N_2\hat{p}_2 + \cdots + N_L\hat{p}_L$ is a good estimator of the total number of viewing households in the population.
- Dividing this quantity by N we obtain an unbiased estimator of the population proportion p of households viewing the show.

Estimation of a Population Proportion

Estimator of the population proportion p:

$$\hat{p}_{\text{st}} = \frac{1}{N^2}(N_1\hat{p}_1 + N_2\hat{p}_2 + \cdots + N_L\hat{p}_L) = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{p}_i \qquad \textbf{(5.13)}$$

Estimated variance of $\hat{p}_{\text{st}}$:

$$\begin{aligned}\hat{V}(\hat{p}_{\text{st}}) &= \frac{1}{N}\left[N_1^2\hat{V}(\hat{p}_1) + N_2^2\hat{V}(\hat{p}_2) + \cdots + N_L^2\hat{V}(\hat{p}_L)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\hat{V}(\hat{p}_i) \\ &= \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{\hat{p}_i\hat{q}_i}{n_i - 1}\right) \qquad \textbf{(5.14)}\end{aligned}$$

Example 5.12

TABLE **5.3**
Data for Example 5.12

Stratum	Sample size	Number of households viewing show X	$\hat{p}_i$
1	$n_1 = 20$	16	0.80
2	$n_2 = 8$	2	0.25
3	$n_3 = 12$	6	0.50

Estimation of a Population Proportion

- The advertising firm wants to estimate the proportion of households in the country of Example 5.1 that view show X. The country is divided into three strata, town A, town B and the rural area.
- The strata contain $N_1 = 155, N_2 = 62, N_3 = 93$ households. A stratified random sample of *n* = 40 households is chosen with proportional allocation. In other words, a simple random sample is taken from each stratum; the sizes of the samples are $n_1 = 20$, $n_2 = 8$, $n_3 = 12$.
- Interviews are conducted in the 40 sampled households; results are shown in Table 5.3.
- Estimate the proportion of households viewing show X, and place a bound on the error of estimation.

Estimation of a Population Proportion

The estimate of the proportion of households viewing show X is given by $\hat{p}_{st}$. Using Eq. (5.13), we calculate

$$\hat{p}_{st} = \frac{1}{310}[(155)(0.80) + 62(0.25) + 93(0.50)] = 0.60$$

The variance of $\hat{p}_{st}$ can be estimated by using Eq. (5.14). First, let us calculate the $\hat{V}(\hat{p}_i)$ terms. We have

$$\hat{V}(\hat{p}_1) = \left(\frac{N_1 - n_1}{N_1}\right)\left(\frac{\hat{p}_1\hat{q}_1}{n_1 - 1}\right) = \left(\frac{155 - 20}{155}\right)\left[\frac{(0.8)(0.2)}{19}\right]$$
$$= (0.871)(0.008) = 0.007$$
$$\hat{V}(\hat{p}_2) = \left(\frac{N_2 - n_2}{N_2}\right)\left(\frac{\hat{p}_2\hat{q}_2}{n_2 - 1}\right) = \left(\frac{62 - 8}{62}\right)\left[\frac{(0.25)(0.75)}{7}\right]$$
$$= (0.871)(0.027) = 0.024$$
$$\hat{V}(\hat{p}_3) = \left(\frac{N_3 - n_3}{N_3}\right)\left(\frac{\hat{p}_3\hat{q}_3}{n_3 - 1}\right) = \left(\frac{93 - 12}{93}\right)\left[\frac{(0.5)(0.5)}{11}\right]$$
$$= (0.871)(0.023) = 0.020$$

From Eq. (5.14)

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{3} N_i^2 \hat{V}(\hat{p}_i)$$

$$= \frac{1}{(310)^2}[(155)^2(0.007) + (62)^2(0.024) + (93)^2(0.020)]$$

$$= 0.0045$$

Then the estimate of proportion of households in the county that view show X, with a bound on the error of estimation, is given by

$$\hat{p}_{st} \pm 2\sqrt{\hat{V}(\hat{p}_{st})} \quad \text{or} \quad 0.60 \pm 2\sqrt{0.0045}$$

$$\text{or} \quad 0.60 \pm 2(0.07) \quad \text{or} \quad 0.60 \pm 0.14 \quad \blacksquare$$

The bound on the error in Example 5.12 is quite large. We could reduce this bound and make the estimator more precise by increasing the sample size. The problem of choosing a sample size is considered in the next section.

Selecting the Sample Size and Allocating the Sample to Estimate Proportions

Approximate sample size required to estimate p with a bound B on the error of estimation:

$$n = \frac{\sum_{i=1}^{L} N_i^2 p_i q_i / a_i}{N^2 D + \sum_{i=1}^{L} N_i p_i q_i} \qquad \textbf{(5.15)}$$

where a_i is the fraction of observations allocated to stratum i, p_i is the population proportion for stratum i, and $D = B^2/4$.

Selecting the Sample Size and Allocating the Sample to Estimate Proportions

Approximate allocation that minimizes cost for a fixed value of $V(\hat{p}_{st})$ or minimizes $V(\hat{p}_{st})$ for a fixed cost:

$$n_1 = n\left(\frac{N_i\sqrt{p_iq_i/c_i}}{N_1\sqrt{p_1q_1/c_1} + N_2\sqrt{p_2q_2/c_2} + \cdots + N_L\sqrt{p_Lq_L/c_L}}\right)$$

$$= n\left(\frac{N_i\sqrt{p_iq_i/c_i}}{\sum_{k=1}^{L} N_k\sqrt{p_kq_k/c_k}}\right) \qquad \textbf{(5.16)}$$

where N_i denotes the size of the ith stratum, p_i denotes the population proportion for the ith stratum, and c_i denotes the cost of obtaining a single observation from the ith stratum.

Example 5.13

- The data in Table 5.3 were obtained from a survey conducted last year. The advertising firm now wants to conduct a new survey in the same country to estimate the proportion of households viewing show X.
- Although the fractions p_1, p_2, p_3 that appear in Eqs. (5.15) and (5.16) are unknown, they can be approximated by the estimates from the earlier study, that is, $\hat{p}_1 = 0.80, \hat{p}_2 = 0.25, \hat{p}_3 = 0.5$
- The cost of obtaining an observation for each stratum is $c_1 = c_2 = \$9, c_3 = \$16$. The number of households within the strata are $N_1 = 155, N_2 = 62, N_3 = 93$.
- The firm wants to estimate the population proportion p with a bound on the error of estimation equal to 0.1. Find the sample size *n* and the strata sample sizes n_1, n_2, n_3, that will give the desired bound at minimum cost.

Solution

SOLUTION We first use Eq. (5.16) to find the allocation fractions a_i. Using $\hat{p}_i$ to approximate p_i, we have

$$\sum_{i=1}^{3} N_i \sqrt{\frac{\hat{p}_i \hat{q}_i}{c_i}} = N_1 \sqrt{\frac{\hat{p}_1 \hat{q}_1}{c_1}} + N_2 \sqrt{\frac{\hat{p}_2 \hat{q}_2}{c_2}} + N_3 \sqrt{\frac{\hat{p}_3 \hat{q}_3}{c_3}}$$

$$= 155 \sqrt{\frac{(0.8)(0.2)}{9}} + 62 \sqrt{\frac{(0.25)(0.75)}{9}} + 93 \sqrt{\frac{(0.5)(0.5)}{16}}$$

$$= \frac{62.000}{3} + \frac{26.846}{3} + \frac{46.500}{4}$$

$$= 20.667 + 8.949 + 11.625 = 41.241$$

and

$$n_1 = n \left(\frac{N_1 \sqrt{\hat{p}_1 \hat{q}_1 / c_1}}{\sum_{i=1}^{3} N_i \sqrt{\hat{p}_i \hat{q}_i / c_i}} \right) = n \left(\frac{20.667}{41.241} \right) = n(0.50)$$

Similarly,

$$n_2 = n\left(\frac{8.949}{41.241}\right) = n(0.22)$$

$$n_3 = n\left(\frac{11.625}{41.241}\right) = n(0.28)$$

Thus, $a_1 = 0.50$, $a_2 = 0.22$, and $a_3 = 0.28$.

The next step is to use Eq. (5.15) to find n. First, the following quantities must be calculated:

$$\sum_{i=1}^{3} \frac{N_i^2 \hat{p}_i \hat{q}_i}{a_i} = \frac{N_1^2 \hat{p}_1 \hat{q}_1}{a_1} + \frac{N_2^2 \hat{p}_2 \hat{q}_2}{a_2} + \frac{N_3^2 \hat{p}_3 \hat{q}_3}{a_3}$$

$$= \frac{(155)^2(0.8)(0.2)}{0.50} + \frac{(62)^2(0.25)(0.75)}{0.22} + \frac{(93)^2(0.5)(0.5)}{0.28}$$

$$= 18.686.46$$

$$\sum_{i=1}^{3} N_i \hat{p}_i \hat{q}_i = N_1 \hat{p}_1 \hat{q}_1 + N_2 \hat{p}_2 \hat{q}_2 + N_3 \hat{p}_3 \hat{q}_3$$

$$= (155)(0.8)(0.2) + (62)(0.25)(0.75) + (93)(0.5)(0.5)$$

$$= 59.675$$

To find D, we let $2\sqrt{V(\hat{p}_{st})} = 0.1$ (the bound on the error of estimation). Then

$$V(\hat{p}_{st}) = \frac{(0.1)^2}{4} = 0.0025 = D$$

and

$$N^2D = (310)^2(0.0025) = 240.25$$

Finally, from Eq. (5.15), n is given approximately by

$$n = \frac{\sum_{i=1}^{3} N_i^2 \hat{p}_i \hat{q}_i / a_i}{N^2D + \sum_{i=1}^{3} N_i \hat{p}_i \hat{q}_i} = \frac{18{,}686.46}{240.25 + 59.675} = 62.3 \text{ or } 63$$

Hence,

$$n_1 = na_1 = (63)(0.501) = 31.6 \approx 32$$
$$n_2 = na_2 = (63)(0.216) = 13.67 \approx 14$$
$$n_3 = na_3 = (63)(0.282) = 17.76 \approx 18$$

- A minor point of interest is that the three unrounded values add to 63, but the usual rounding rules lead to a total sample size of 64

Reference

- Scheaffer RL, William M, Lyman O. 2006. Elementary Survey Sampling sixth ed. PWS-KENT Publishing Company.

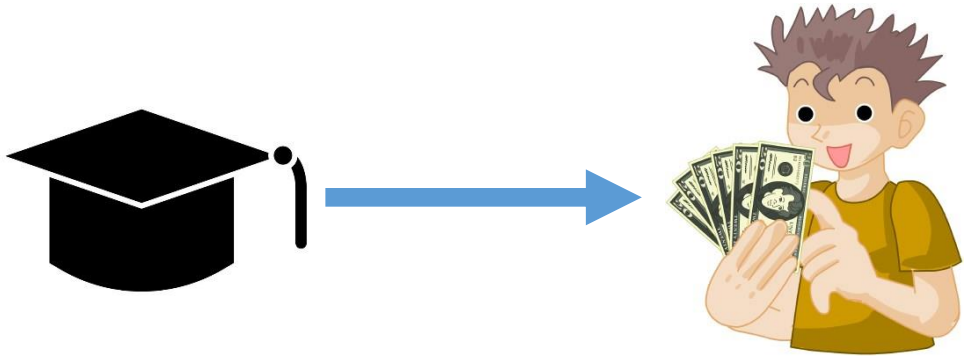
Simple Linear Regression

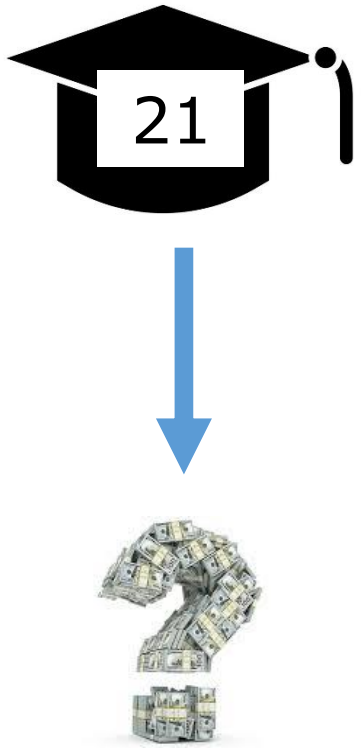
Metode kuantitatif

Tim pengajar

2022

Simple Linear Regression





education	11	12	11	15	8	10	11	12	17	11
salary	25	33	22	41	18	28	32	24	53	26

Simple Linear Regression

- Our objective is to study the relationship between two variables X and Y.
- One way is by means of **regression.**
- Regression analysis is the process of estimating a functional relationship between X and Y. A regression equation is often used to predict a value of Y for a given value of X.
- Another way to study relationship between two variables is **correlation.** It involves measuring the direction and the strength of the **linear** relationship.

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where

y = dependent variable

x = independent variable

b_0= y-intercept

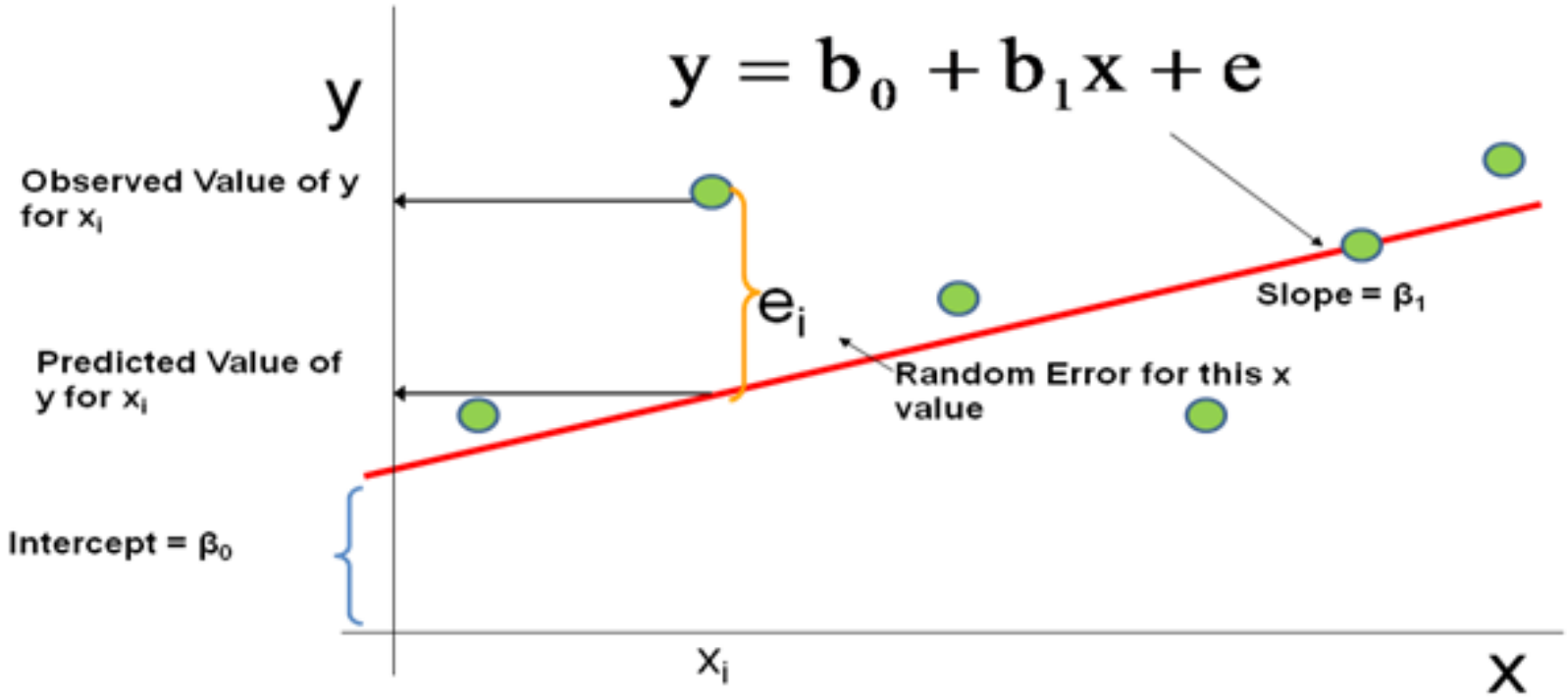
b_1= slope of the line

e = error variable

This model is

- **Simple**: only one X
- **Linear in the parameters**: No parameter appears as exponent or is multiplied or divided by another parameter
- **Linear in the predictor variable (X)**: X appears only in the first power.

Deterministic Component of Model



Sumber: https://www.econ-analysis.com

Error

- The scatterplot shows that the points are not on a line, and so, in addition to the relationship, we also describe the error:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ i=1,2,...,n}$$

- The Y's are the response (or dependent) variable. The x's are the predictors or independent variables, and the epsilon's are the errors. We assume that the errors are normal, mutually independent, and have variance σ^2.

Least Squares

- Minimize: $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2$

$$\Longrightarrow \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- The quantities $R_i = y_i - \hat{y}_i$ are called the residuals.
- If we make $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into matrix eq., then
- $\hat{Y} = \beta X$ with
- $\hat{Y}$ is$(1xn)$, β is $(1x2)$ and X is $(2xn)$

Least Squares

- $\hat{Y} = \beta X$ then $\varepsilon = Y - \hat{Y}$
- If $Q = \varepsilon^2$ $then$ $Q = \varepsilon\varepsilon^T$
- $Q = (Y - \beta X) * (Y - \beta X)^T$
- $Q = (Y - \beta X) * (Y^T - (\beta X)^T)$
- $Q = (Y - \beta X) * (Y^T - (X)^T(\beta)^T)$
- $Q = YY^T - YX^T\beta^T - \beta XY^T + \beta XX^T\beta^T$
- Actualy, $YX^T\beta^T$ is a scalar (note: check the dimension). So, $YX^T\beta^T = (YX^T\beta^T)^T$

Least Squares

- $Q = YY^T - YX^T\beta^T - \beta XY^T + \beta XX^T\beta^T$
- $Q = YY^T - 2\beta XY^T + \beta XX^T\beta^T$
- We want to find β so that Q is minimum
- $\frac{\partial Q}{\partial \beta} = 0, \frac{\partial (YY^T - 2\beta XY^T + \beta XX^T\beta^T)}{\partial \beta} = 0$
- We found that:$(-2XY^T + 2XX^T\beta^T = 0)$
- $XX^T\beta^T = 2XY^T$, let's multiply each side with $(XX^T)^{-1}$ then we get
- $\beta = (YX^T)(XX^T)^{-1}$

Least Squares

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$SS_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

What form does the error take?

- Each observation may be decomposed into two parts:

$$y = \hat{y} + (y - \hat{y})$$

- The first part is used to determine the fit, and the second to estimate the error.
- We estimate the standard deviation of the error by:

$$SSE = \sum (Y - \hat{Y})^2 = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}} \right)$$

Estimate of σ^2

- We estimate σ^2 by

$$s^2 = \frac{SSE}{n-2} = MSE$$

Example

- An educational economist wants to establish the relationship between an individual's income and education. He takes a random sample of 10 individuals and asks for their income (in $1000s) and education (in years). The results are shown below. Find the least squares regression line.

education	11	12	11	15	8	10	11	12	17	11
salary	25	33	22	41	18	28	32	24	53	26

Dependent and Independent Variables

- The dependent variable is the one that we want to forecast or analyze.
- The independent variable is hypothesized to affect the dependent variable.
- In this example, we wish to analyze income and we choose the variable individual's education that most affects income. Hence, y is income and x is individual's education

First Step:

- $\beta = (YX^T)(XX^T)^{-1}$
- $Y = [25 \quad 33 \;\; 22 \quad 41 \;\; 18 \quad 28 \;\; 32 \quad 24 \;\; 53 \quad 26] Y = (25, 33, 22, 41, 18, 28,$

The Least Squares Regression Line

- The least squares regression line is

$$\hat{y} = -13.93 + 3.74x$$

- Interpretation of coefficients:

*The sample slope $\hat{\beta}_1 = 3.74$ tells us that on average for each additional year of education, an individual's income rises by $3.74 thousand.

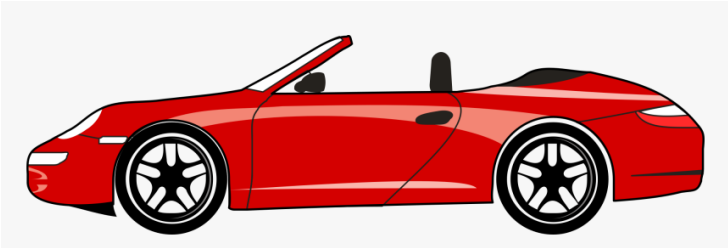
- The y-intercept is $\hat{\beta}_0 = -13.93$. This value is the expected (or average) income for an individual who has 0 education level (which is meaningless here)

Example

- Car dealers across North America use the red book to determine a cars selling price on the basis of important features. One of these is the car's current odometer reading.
- To examine this issue 100 three-year old cars in mint condition were randomly selected. Their selling price and odometer reading were observed.

Portion of the data file

Odometer	Price
37388	5318
44758	5061
45833	5008
30862	5795
…..	…
34212	5283
33190	5259
39196	5356
36392	5133



Example (Minitab Output)

Regression Analysis

The regression equation is
Price = 6533 - 0.0312 Odometer

Predictor	Coef	StDev	T	P	
Constant	6533.38	84.51	77.31	**0.000**	**(SIGNIFICANT)**
Odometer	-0.031158	0.002309	-13.49	**0.000**	**(SIGNIFICANT)**

S = 151.6 R-Sq = 65.0% R-Sq(adj) = 64.7%

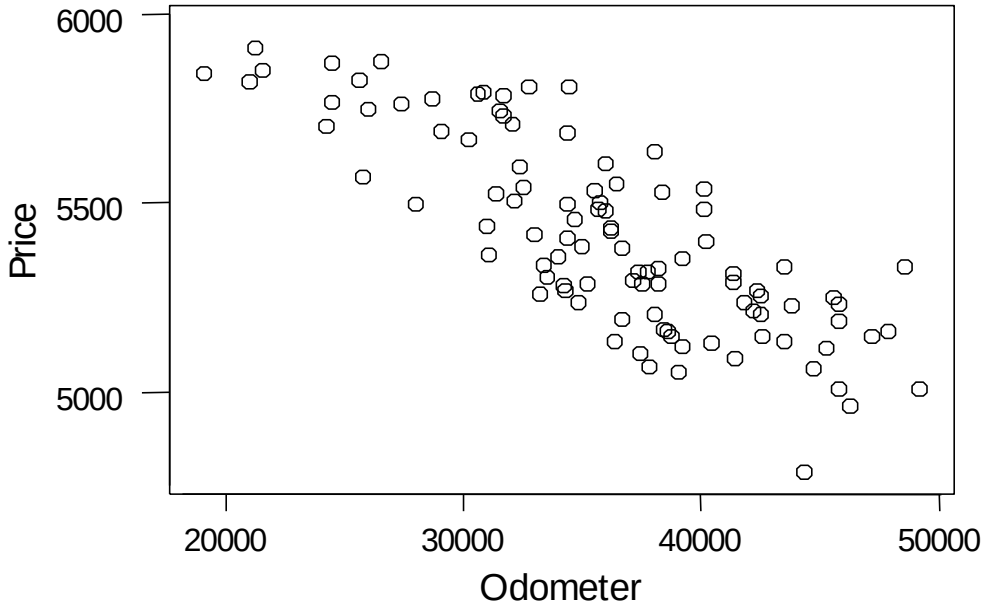
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4183528	4183528	182.11	0.000
Error	98	2251362	22973		
Total	99	6434890			

Example

- The least squares regression line is

$$\hat{y} = 6533.38 - 0.031158x$$

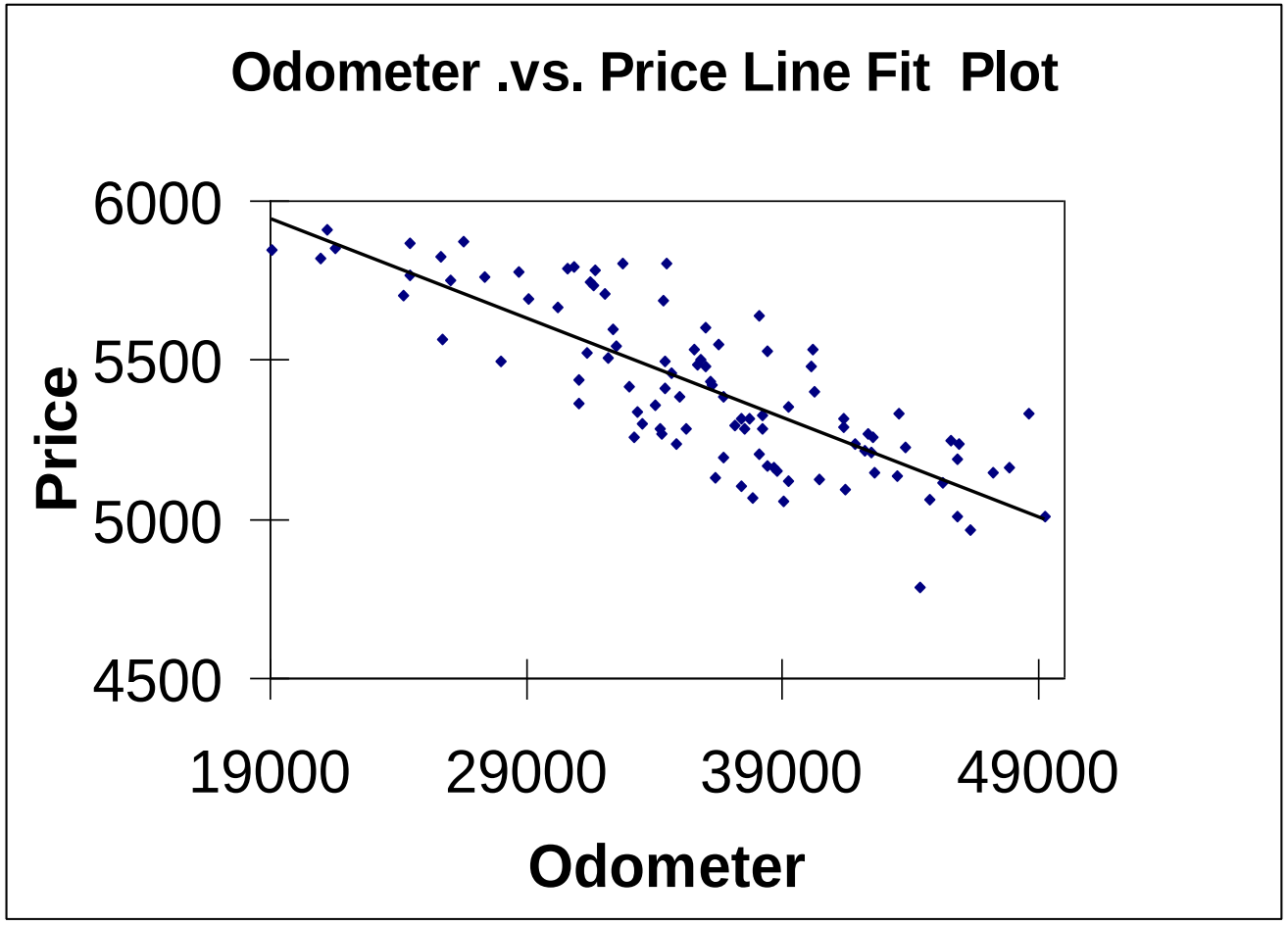


- $\hat{\beta}_1 = -0.031158$ means that for each additional mile on the odometer, the price decreases by an average of 3.1158 cents.
- $\hat{\beta}_0 = 6533.38$ means that when x = 0 (new car), the selling price is $6533.38 but x = 0 is not in the range of x. So, we cannot interpret the value of y when x=0 for this problem.
- R^2=65.0% means that 65% of the variation of y can be explained by x. The higher the value of R^2, the better the model fits the data.

R^2 and R^2 adjusted

- R^2 measures the degree of linear association between X and Y.
- So, an R^2 close to 0 does not necessarily indicate that X and Y are unrelated (relation can be nonlinear)
- Also, a high R^2 does not necessarily indicate that the estimated regression line is a good fit.
- As more and more X's are added to the model, R^2 always increases. R^2adj accounts for the number of parameters in the model.

Scatter Plot



Testing the slope

- Are X and Y linearly related?

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \quad \text{where} \quad s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{SS_x}}$$

$$SS_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SSE = \sum (Y - \hat{Y})^2 = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right)$$

Testing the slope (continue)

- The Rejection Region:
  - Reject H_0 if $t < -t_{\alpha/2,n-2}$ or $t > t_{\alpha/2,n-2}$.
- If we are testing that high x values lead to high y values, H_A: $\beta_1>0$. Then, the rejection region is $t > t_{\alpha,n-2}$.
- If we are testing that high x values lead to low y values or low x values lead to high y values, H_A: $\beta_1 <0$. Then, the rejection region is $t < - t_{\alpha,n-2}$.

Assessing the model

Example:

- Excel output

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	6533.4	84.512322	77.307	1E-89
Odometer	-0.031	0.0023089	-13.49	4E-24

- Minitab output

Predictor	Coef	StDev	T	P
Constant	6533.38	84.51	77.31	0.000
Odometer	-0.031158	0.002309	-13.49	0.000

Coefficient of Determination

$$R^2 = 1 - \frac{SSE}{SS_y}$$

$$SSE = \sum (Y - \hat{Y})^2 = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}} \right)$$

For the data in odometer example, we obtain:

$$SS_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i \right)^2}{n}$$

$$R^2 = 1 - \frac{SSE}{SS_y} = 1 - \frac{2{,}251{,}363}{6{,}434{,}890}$$

$$= 1 - 0.3499 = 0.6501$$

$$R^2_{\text{adj}} = 1 - \left(\frac{n-1}{n-p} \right) \frac{SSE}{SS_y}$$

where p is number of predictors in the model.

Using the Regression Equation

- Suppose we would like to predict the selling price for a car with 40,000 miles on the odometer

$$\begin{aligned}\hat{y} &= 6,533 - 0.0312x \\ &= 6,533 - 0.0312(40,000) \\ &= \$5,285\end{aligned}$$

Prediction and Confidence Intervals

- Prediction Interval of y for $x=x_g$: The confidence interval for *predicting the particular value of y* for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{1+\frac{1}{n}+\frac{(x_g-\bar{x})^2}{SS_x}}$$

- Confidence Interval of $E(y|x=x_g)$: The confidence interval for *estimating the expected value of y* for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n}+\frac{(x_g-\bar{x})^2}{SS_x}}$$

Solving by Hand (Prediction Interval)

- From previous calculations we have the following estimates:

$$\hat{y} = 5285, s_{\varepsilon} = 151.6, SS_x = 4309340160, \bar{x} = 36,009$$

- Thus a 95% **prediction interval** for x=40,000 is:

$$5,285 \pm 1.984(151.6)\sqrt{1 + \frac{1}{100} + \frac{(40,000 - 36,009)^2}{4,309,340,160}}$$

$$5,285 \pm 303$$

- The prediction is that the selling price of the car will fall between $4982 and $5588.

Solving by Hand (Prediction Interval)

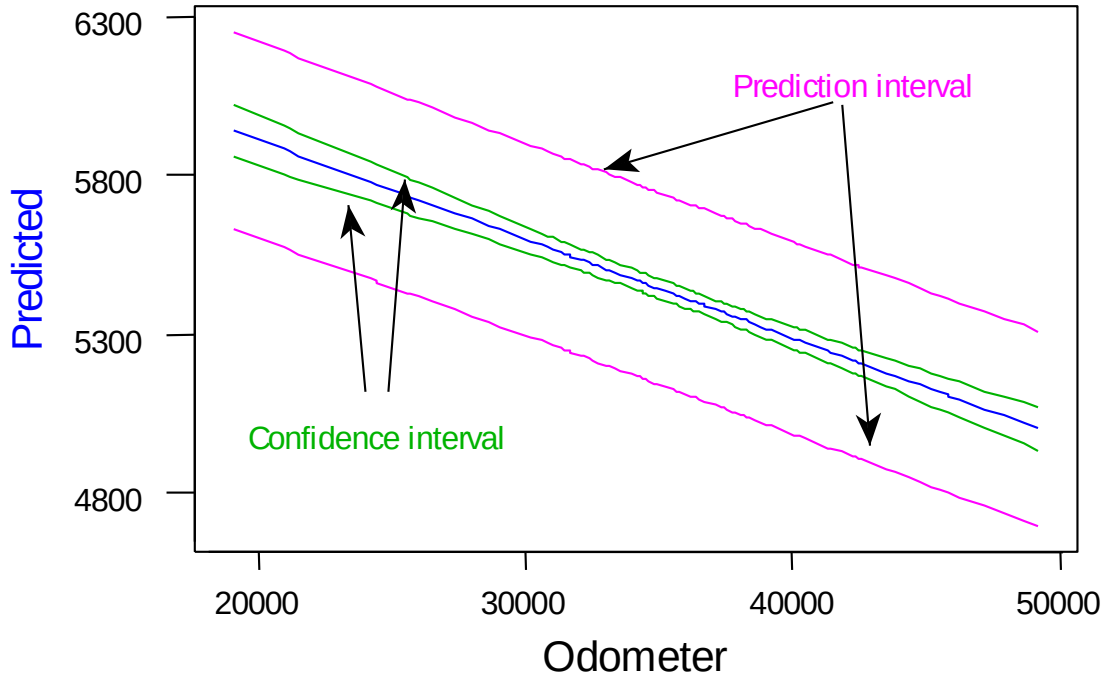
- A 95% **confidence interval** of E(y| x=40,000) is:

$$5{,}285 \pm 1.984(151.6)\sqrt{\frac{1}{100} + \frac{(40{,}000 - 36{,}009)^2}{4{,}309{,}340{,}160}}$$

$$5{,}285 \pm 35$$

- The mean selling price of the car will fall between $5250 and $5320.

Prediction and Confidence Intervals' Graph



Notes

- No matter how strong is the statistical relation between X and Y, no cause-and-effect pattern is necessarily implied by the regression model. Ex: Although a positive and significant relationship is observed between vocabulary (X) and writing speed (Y), this does *not* imply that an increase in X *causes* an increase in Y. Other variables, such as age, may affect both X and Y. Older children have a larger vocabulary and faster writing speed.

Regression Diagnostics

Residual Analysis:

- Non-normality
- Heteroscedasticity (non-constant variance)
- Non-independence of the errors
- Outlier
- Influential observations

Standardized Residuals

- The standardized residuals are calculated as

$$\text{Standardized residual} = \frac{r_i}{s_\varepsilon}$$

where $r_i = y_i - \hat{y}_i$.

- The standard deviation of the i-th residual is

$$s_{r_i} = s_\varepsilon \sqrt{1 - h_i} \text{ where } h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SS_x}$$

Non-normality:

- The errors should be normally distributed. To check the normality of errors, we use histogram of the residuals or normal probability plot of residuals or tests such as Shapiro-Wilk test.
- Dealing with non-normality:
  - Transformation on Y
  - Other types of regression (e.g., Poisson or Logistic …)
  - Nonparametric methods (e.g., nonparametric regression(i.e. smoothing))

Non-constant variance:

- The error variance σ_{ε}^{2} should be constant.
- To diagnose non-constant variance, one method is to plot the residuals against the predicted value of y (or x). If the points are distributed evenly around the expected value of errors which is 0, this means that the error variance is constant. Or, formal tests such as: Breusch-Pagan test

Dealing with non-constant variance

- Transform Y
- Re-specify the Model (e.g., Missing important X's?)
- Use Weighted Least Squares instead of Ordinary Least Squares

$$\min \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\mathrm{Var}(\varepsilon_i)}$$

Non-independence of error variable:

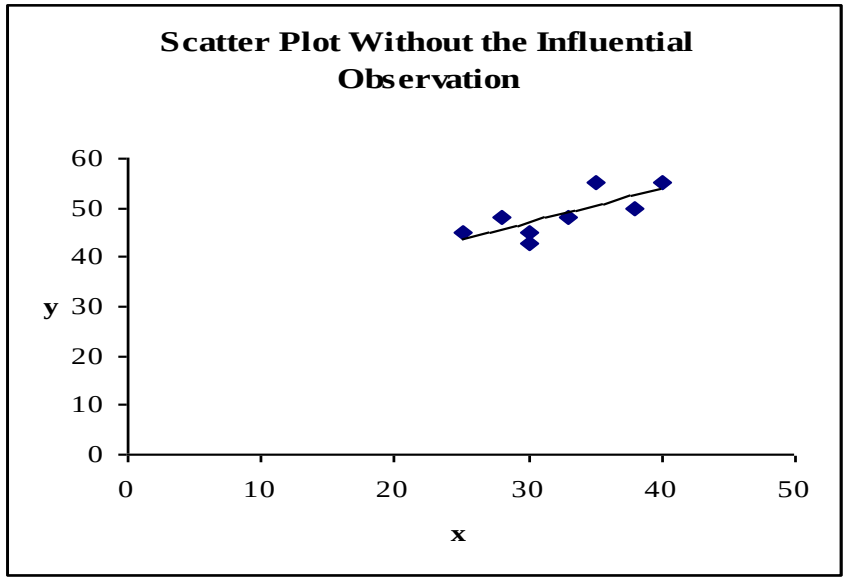
- The values of error should be independent. When the data are time series, the errors often are correlated (i.e., autocorrelated or serially correlated). To detect autocorrelation we plot the residuals against the time periods. If there is no pattern, this means that errors are independent. Or, more formal tests such as Durbin-Watson

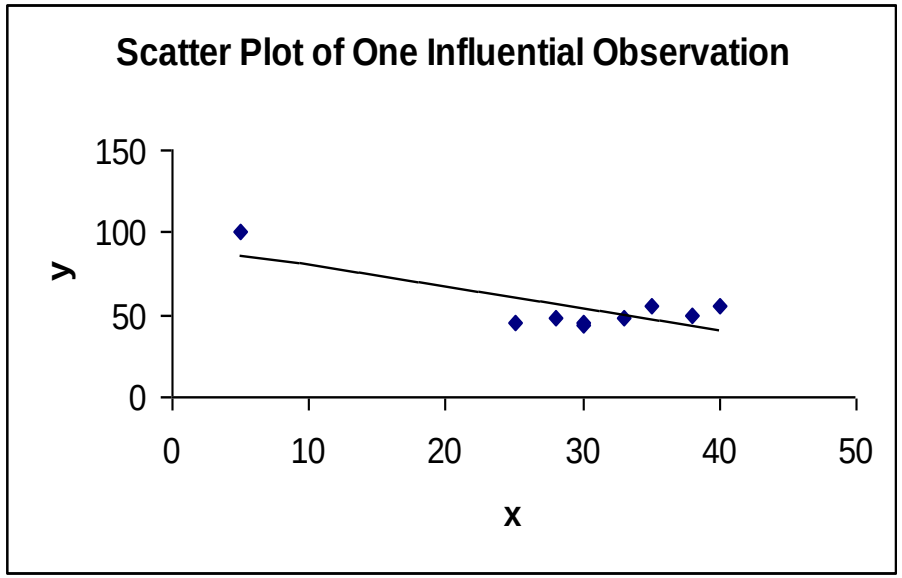
Outlier:

- An outlier is an observation that is unusually small or large. Two possibilities which cause outlier is

1. Error in recording the data. $\Rightarrow$ Detect the error and correct it
    - The outlier point should not have been included in the data (belongs to another population) $\Rightarrow$ Discard the point from the sample
2. The observation is unusually small or large although it belong to the sample and there is no recording error. $\Rightarrow$ Do NOT remove it

Influential Observations





- Detection:
  - Cook's Distance, DFFITS, DFBETAS (Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., (1996) Applied Linear Statistical Models, 4th edition, Irwin, pp. 378-384)

Multicollinearity

- A common issue in multiple regression is multicollinearity. This exists when some or all of the predictors in the model are highly correlated. In such cases, the estimated coefficient of any variable depends on which other variables are in the model. Also, standard errors of the coefficients are very high…

Multicollinearity

- Look into correlation coefficient among X's: If Cor>0.8, suspect multicollinearity
- Look into Variance inflation factors (VIF): VIF>10 is usually a sign of multicollinearity
- If there is multicollinearity:
  - Use transformation on X's, e.g. centering, standardization. Ex: $Cor(X,X^2)=0.991$; after standardization Cor=0!
  - Remove the X that causes multicollinearity
  - Factor analysis
  - Ridge regression
  - …

Exercise

- In baseball, the fans are always interested in determining which factors lead to successful teams. The table below lists the team batting average and the team winning percentage for the 14 league teams at the end of a recent season.

Team-B-A	Winning%
0.254	0.414
0.269	0.519
0.255	0.500
0.262	0.537
0.254	0.352
0.247	0.519
0.264	0.506
0.271	0.512
0.280	0.586
0.256	0.438
0.248	0.519
0.255	0.512
0.270	0.525
0.257	0.562

y = winning % and
x = team batting average

LS Regression Line

$$\sum x_i = 3.642, \sum x_i^2 = 0.949$$
$$\sum y_i = 7.001, \sum y_i^2 = 3.549$$
$$\sum x_i y_i = 1.824562$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 1.824562 - \frac{(3.642)(7.001)}{14} = 0.0033$$

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 0.948622 - \frac{(3.642)^2}{14} = 0.00118$$

LS Regression Line

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{0.003302}{0.001182} = 0.7941$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.5 - (0.7941)0.26 = 0.2935$$

- The least squares regression line is

$$\hat{y} = 0.2935 + 0.7941x$$

- The meaning $\hat{\beta}_1 = 0.7941$ is for each additional batting average of the team, the winning percentage increases by an average of 79.41%.

Standard Error of Estimate

$$SSE = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right) = \left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right) - \left(\frac{S_{xy}^2}{S_{xx}}\right)$$

$$= (3.548785 - \frac{7.001^2}{14}) - \frac{0.003302^2}{0.00182} = 0.03856$$

So, $s_\varepsilon^2 = \frac{SSE}{n-2} = \frac{0.03856}{14-2} = 0.00321$ and $s_\varepsilon = \sqrt{s_\varepsilon^2} = 0.0567$

- Since s_ε=0.0567 is small, we would conclude that “s” is relatively small, indicating that the regression line fits the data quite well.

Do the data provide sufficient evidence at the 5% significance level to conclude that higher team batting average lead to higher winning percentage?

$$H_0: \quad \beta_1 = 0$$
$$H_A: \quad \beta_1 > 0$$

$$\text{Test statistic: } t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = 1.69 \qquad \text{(p-value=.058)}$$

Conclusion: Do not reject H_0 at $\alpha = 0.05$. The higher team batting average does not lead to higher winning percentage.

Coefficient of Determination

$$R^2 = \frac{SS_{xy}^2}{SS_x - SS_y} = 1 - \frac{SSE}{SS_y} = 1 - \frac{0.03856}{0.04778} = 0.1925$$

The 19.25% of the variation in the winning percentage can be explained by the batting average.

Predict with 90% confidence the winning percentage of a team whose batting average is 0.275.

$$\hat{y} = 0.2935 + 0.7941(0.275) = 0.5119$$

$$\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS_x}} =$$

$$0.5119 \pm (1.782)(0.0567)\sqrt{1 + \frac{1}{14} + \frac{(0.275 - 0.2601)^2}{0.001182}}$$

$$0.5119 \pm 0.1134$$

90% PI for y: $(0.3985, 0.6253)$

The prediction is that the winning percentage of the team will fall between 39.85% and 62.53%.

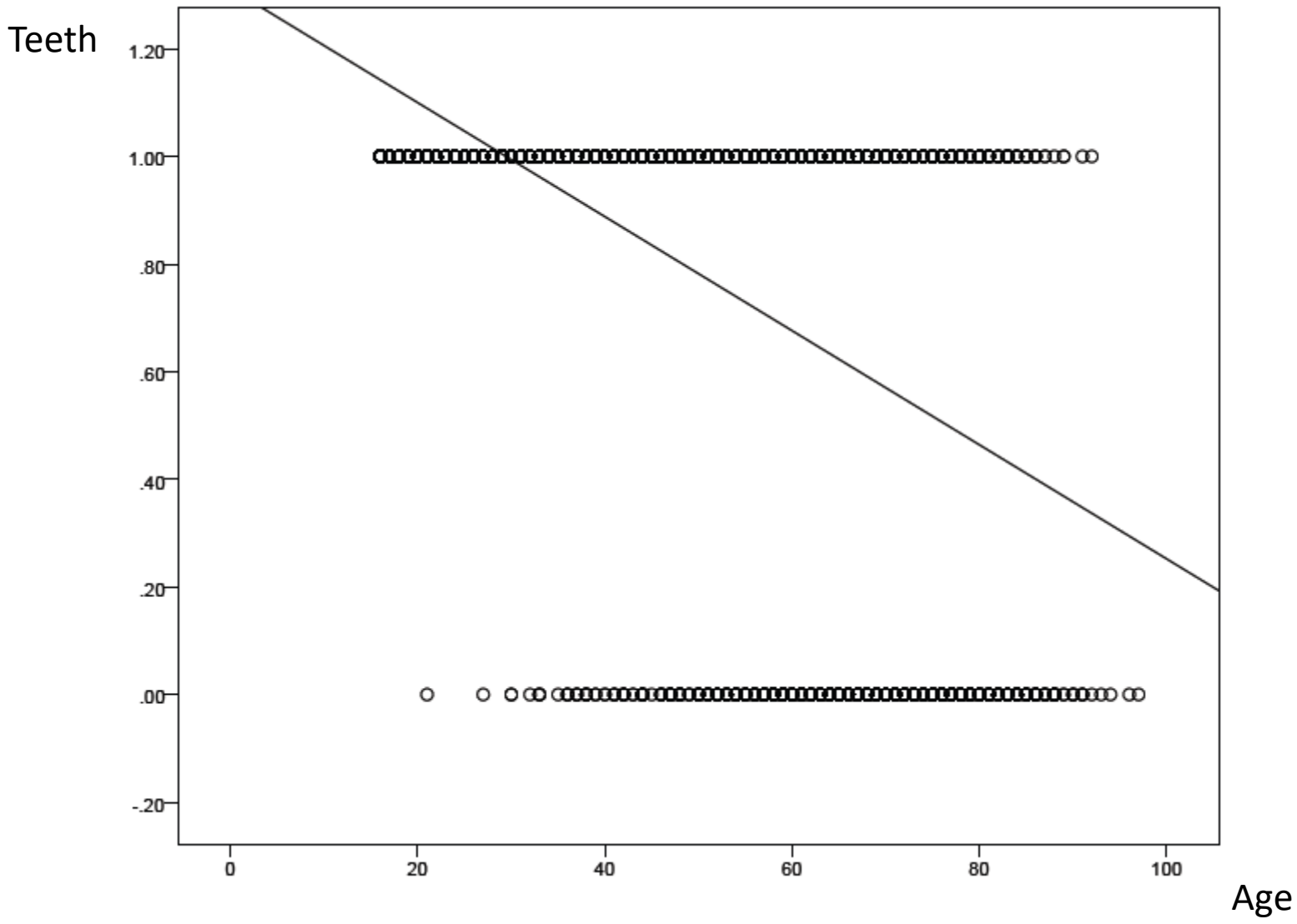
Logistic Regression
2022

What happens if we want to use a categorical dependent variable?

- More specifically, can we use (or adapt) linear regression if we want to look at an outcome which consists of discrete, unordered alternatives?
- It is tempting to think that since we can use two-category, 0/1 dummy variables as independent variables in a linear regression, we should also be able to use a 0/1 variable as the dependent variable.

Teeth!

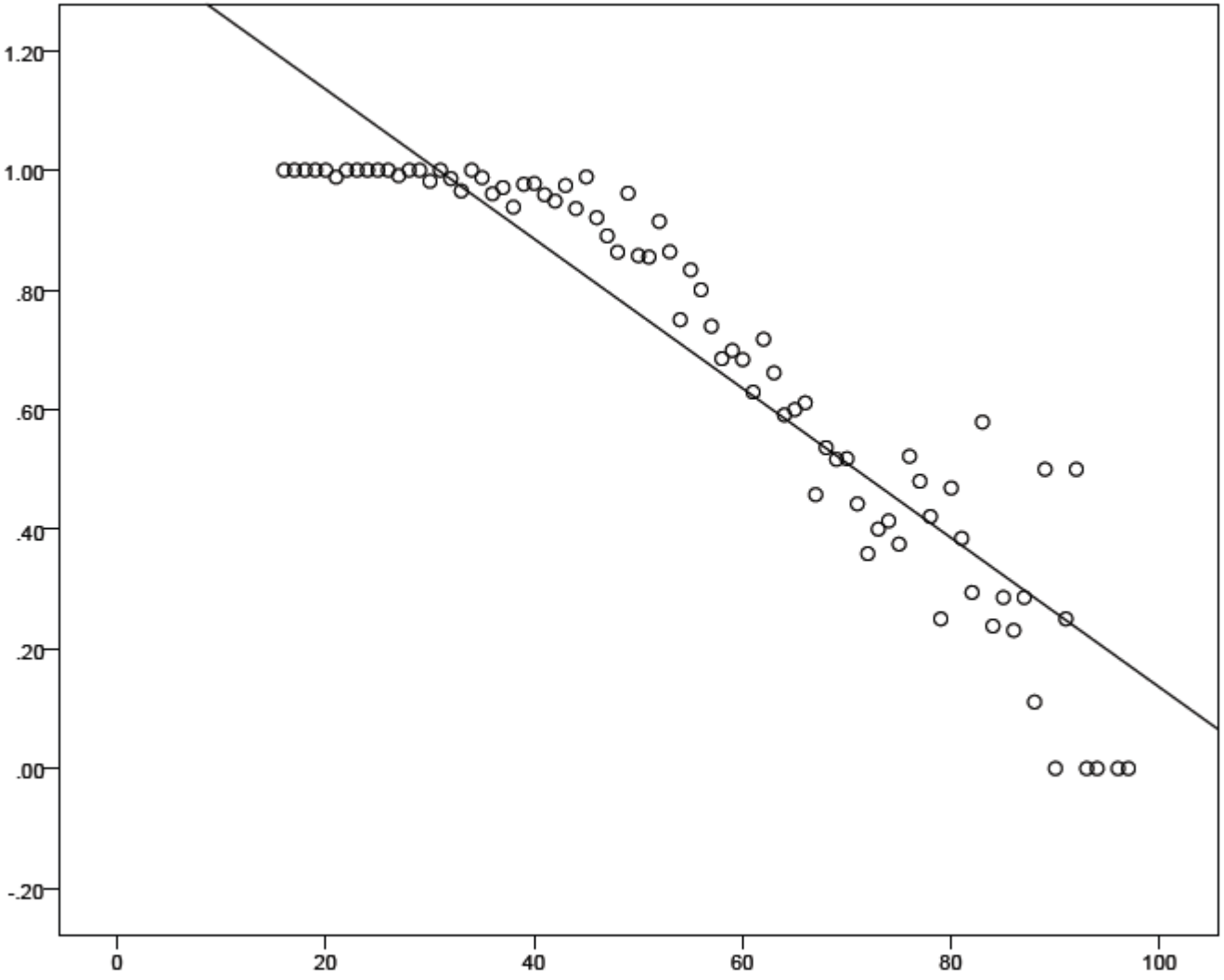
- Suppose that we are interested in whether individuals have any of their original, 'natural' teeth, and how this varies according to age.
- Note that this is, in effect, a simplification of considering *how many* teeth an individual has.
- On the next slide we can see what happens if we carry out a linear regression of teeth on age, coding (natural) teeth as 1 and no (natural) teeth as 0.


Teeth
1.20
1.00
.80
.60
.40
.20
.00
-.20
0
20
40
60
80
100
Age

- The regression line can be expressed as:

  TEETH = (*B* x AGE) + *C*
- However, the line is nowhere near the plotted cases, predicts values other than 0 or 1, and leaves 'residuals' which will more or less inevitably deviate from assumptions of normality and homoscedasticity.
- But the predicted values might perhaps be interpreted as *probabilities*. So what happens if we fit a line to the *proportions* of people of different ages who have any natural teeth?

Proportion With teeth (P)

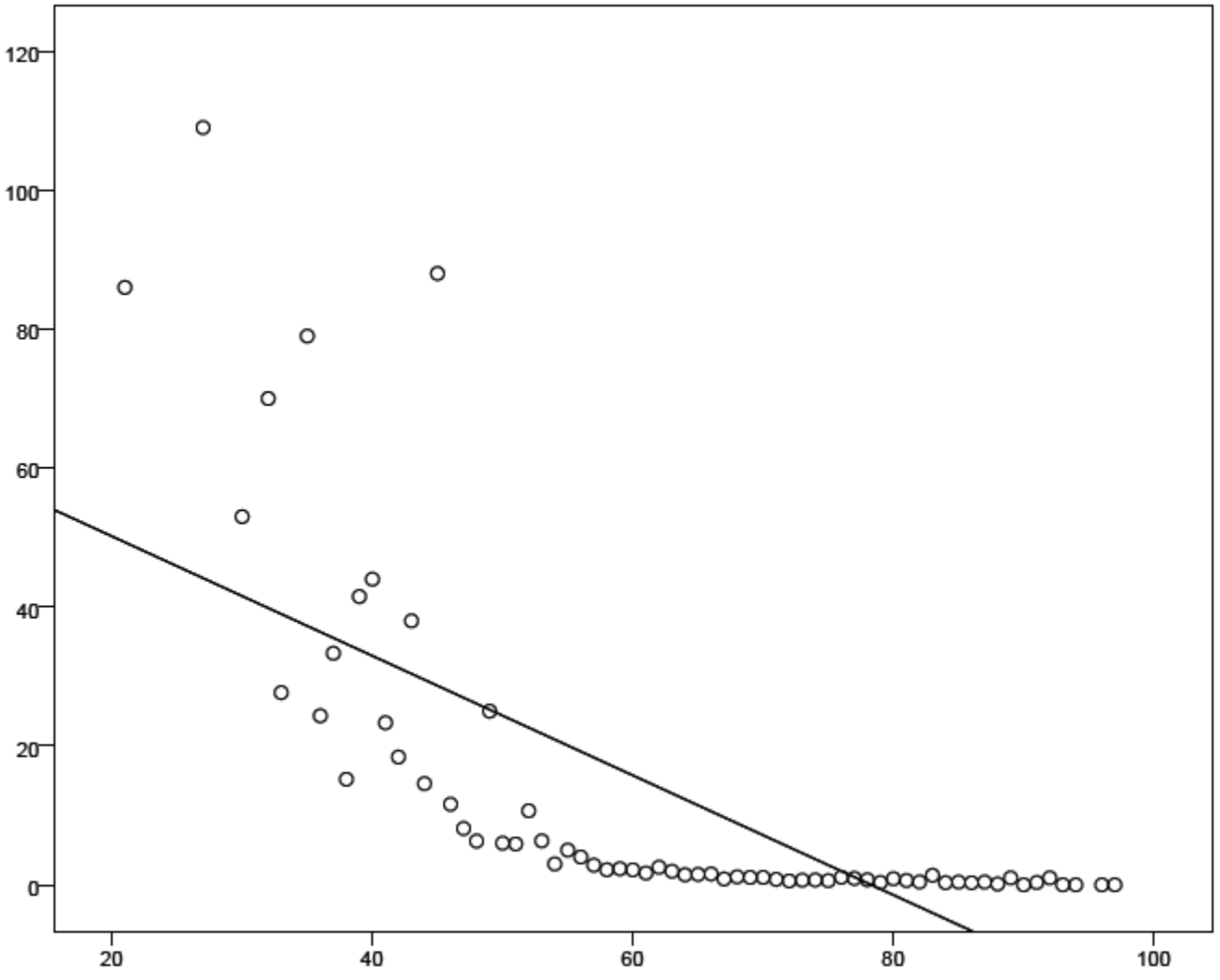


Age

- The ‘regression’ line can now be expressed as:

  $P = (B \times AGE) + C$
- However, it does not fit the reverse S-shape of the points that well, and predicts values of above 1 for lower values of age!
- An alternative to examining the *proportion* of people with teeth, i.e. the *probability* of teeth, is to examine the *odds* of teeth, i.e. P/(1 - P)

Odds of having teeth



Age

Regression Line

- The ‘regression’ line can now be expressed as:

  $P/(1 - P) = (B \times AGE) + C$

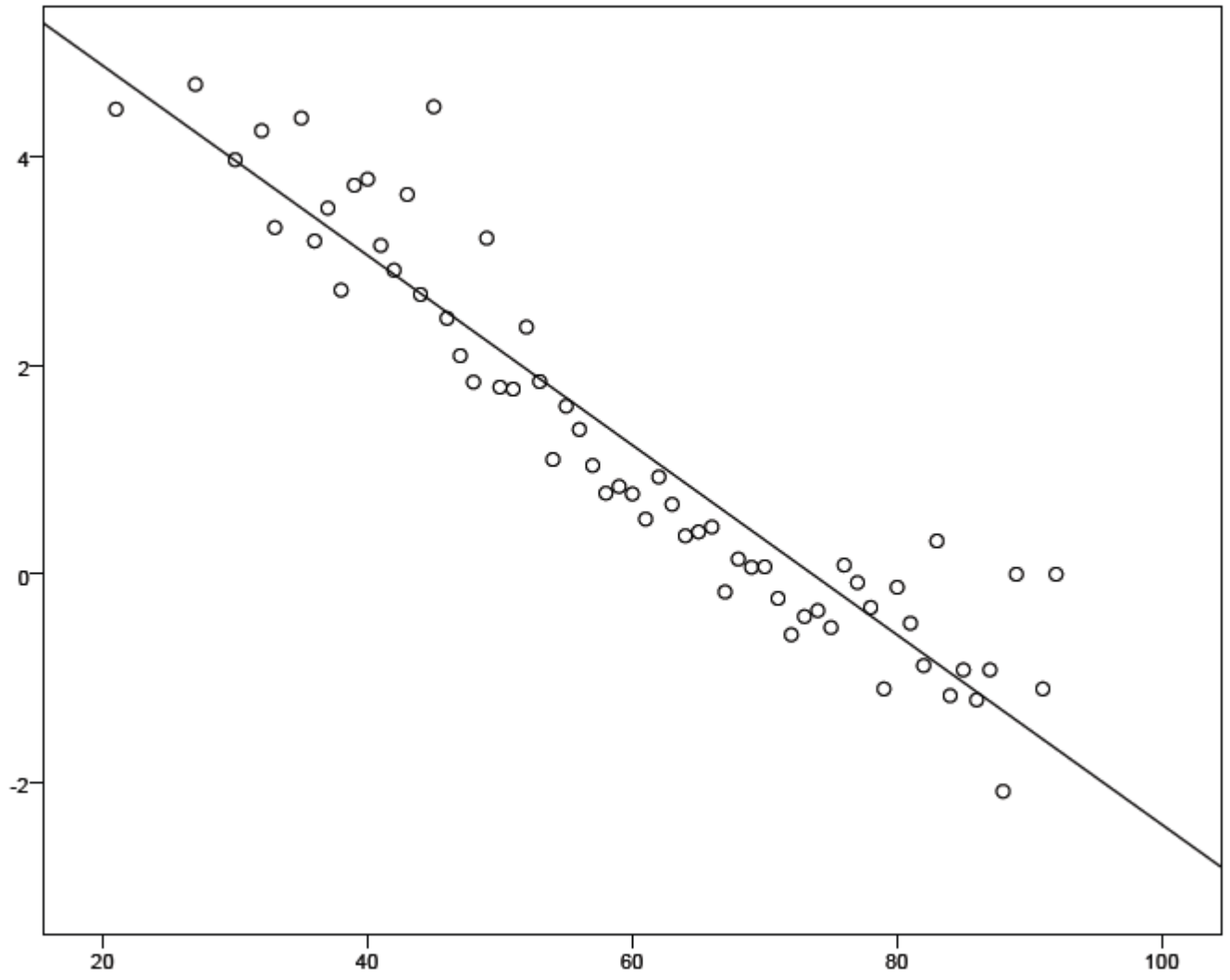
- Looking at the odds of teeth has solved the problem of predicted values of over 1 (since odds can take values between 0 and ∞ (i.e. infinity)
- However, there are now predicted values of less than zero at higher ages, and the summary line does not fit the curvature of the plotted values of the odds.

  (Note that some of the odds values are ∞, and hence not plotted, and the ‘regression’ line is thus also only based on the non-∞ values!)

Time for a transformation?

- But the shape of the curve looks a bit like a negative exponential plot, suggesting that a logarithmic transformation might be helpful?

Log odds of having teeth



Age

Logistic regression!

- The formula for the 'regression' line is now:

  $\log [P/(1 - P)] = (B \times AGE) + C$

- This is the formula for a *logistic regression*, as the transformation from P into log [P/(1 – P)] is called a *logistic transformation*.
- log [P/(1 – P)] is sometimes referred to as the *logit* of P, and hence logistic regressions are sometimes referred to as *logit models*.

In this case...

- Fitting the model to the data in an appropriate way (more details later!) gives the following formula:

  $\log [P/(1 - P)] = (-0.10 \times AGE) + 7.1$

- i.e. $B = -0.10$ and $C = 7.1$
- But we are now predicting the log odds of having teeth, which is at best difficult to grasp as a meaningful thing to do...

The solution!

- The solution is to take the predicted values and subject them to the reverse (inverse) transformation.
- The ‘opposite’ of the logarithmic transformation is exponentiation
- i.e. Exp (log [P/(1 – P)]) = P/(1 – P)
- It is quite straightforward to move back from odds to probabilities too!

$$P/(1 - P) = \text{Exp}\{7.1 - (0.10 \times \text{AGE})\}$$

Age	**Predicted value**	**Odds**	**Probability**
20	5.1	Exp(5.1) = 164.0	0.994
60	1.1	Exp(1.1) = 3.0	0.750
80	-0.9	Exp(-0.9) = 0.41	0.291
100	-2.9	Exp(-2.9) = 0.055	0.052

Odds ratios

- In fact, logistic regression analyses usually focus on the *odds* of the outcome rather than the *probability* of the outcome.
- So the focus in terms of effects is on the *B* values subjected to the exponentiation transformation, i.e. Exp(*B*) values.
- These are *odds ratios*, which have *multiplicative* effects on the odds of the outcome.

Logistic regression and odds ratios

- Men: 1967/294 = 6.69 (to 1)
- Women: 1980/511 = 3.87 (to 1)

- Odds ratio 6.69/3.87 = 1.73

- Men: P/(1 - P) = 3.87 x 1.73 = 6.69
- Women: P/(1 - P) = 3.87 x 1 = 3.87

Odds and log odds

- Odds = Constant x Odds ratio
- Log odds = log(constant) + log(odds ratio)

- Men

  $\log (P/(1 - P)) = \log(3.87) + \log(1.73)$

- Women

  $\log (P/(1 - P)) = \log(3.87) + \log(1)$

  $= \log(3.87)$

- $\log (P/(1 - P)) = \text{constant} + \log(\text{odds ratio})$

- Note that:

log(3.87) = 1.354

log(6.69) = 1.900

log(1.73) = 0.546

log(1) = 0

- And that the ‘reverse’ of the logarithmic transformation is exponentiation

- log (P/(1 - P)) = constant(*C*) + (*B* x SEX)

  where *B* = log(1.73)

  SEX = 1 for men

  SEX = 0 for women

- Log odds for men = 1.354 + 0.546 = 1.900
- Log odds for women

  = 1.354 + 0 = 1.354

- Exp(1.900) = 6.69 & Exp(1.354) = 3.87

Interpreting effects in Logistic Regression

- In the above example:

  Exp(*B*) = Exp(log(1.73)) = 1.73 (the odds ratio!)

- In general, effects in logistic regression analysis take the form of exponentiated *B*'s (Exp(*B*)'s), which are *odds ratios*. Odds ratios have a multiplicative effect on the (odds of) the outcome
- Is a *B* of 0.546 (= log(1.73)) significant?
- In this case $p = 0.000 < 0.05$ for this *B*.

Back from odds to probabilities

- Probability = Odds / (1 + Odds)

- Men: 6.69 / (1 + 6.69) = 0.870
- Women: 3.87 / (1 + 3.87) = 0.795

‘Multiple’ Logistic regression

- log odds = C + (B_1 x SEX) + (B_2 x AGE)

  = C + (0.461 x SEX) + (-0.099 x AGE)

- For B_1 = 0.461, p = 0.000 < 0.05
- For B_2 = -0.099, p = 0.000 < 0.05

- Exp(B_1) = Exp(0.461) = 1.59
- Exp(B_2) = Exp(-0.099) = 0.905

Other points about logistic regression

- Categorical variables can be added to a logistic regression analysis in the form of dummy variables. (SPSS automatically converts categorical variables into these!!)
- The model is fitted to the data via a process of *maximum likelihood estimation*, i.e. the values of *B* are chosen in such a way that the model is the one most likely to have generated the observed data.

Other points (continued)

- Instead of assessing the ‘fit’ of a logistic regression using something like *r*-squared, the *deviance* of a model from the data is examined, in comparison with simpler models.
- This is, in effect, a form of chi-square statistic, as are *changes* in deviance between models.
- Measures which do a broadly equivalent job to *r*-squared in terms of assessing variation explained are also available.

Assumptions

- Logistic regression assumes a linear relationship between the log odds of the outcome and the explanatory variables, so transformations of the latter may still be necessary.
- The assumption of independent error terms ('residuals') and the issue of collinearity are still of relevance.
- However, we do not need to worry about residuals being normally distributed or about homoscedasticity, since these linear regression assumptions are not relevant to logistic regression.

PERANCANGAN PERCOBAAN (EXPERIMENTAL DESIGN) : 2 kali pertemuan

Ambil kasus misalnya kita ingin membuat suatu software edukasi untuk siswa TK agar mencintai lingkungan. Inginnya software mempunyai kinerja baik. Oleh karena itu, apa saja yg perlu dirumuskan : (dalam hal ini konten materi yg akan sediakan dalam sofware sudah disediakan oleh konsultan lain).

Untuk menghasilkan kinerja software yg baik, apa saja yg mempengaruhi kinerja software. Apa yg dimaksud kinerja dalam hal ini?

Menentukan parameter cinta lingkungan

Menentukan hal2 agar mencintai lingkungan

Metode Pembelajaran yg disukai anak2 TK

Bentuk tampilan → disain

Cara penyajian → penyajian

Gaya bahasa penyampaian

Sistem penghargaan untuk anak2 yg belajar mengenai lingkungan → adanya hadiah

Dsb.

Tujuan : agar anak2 TK mencintai lingkungan → kinerja software dikaitkan peningkatan kecintaan dari anak TK thd lingkungan setelah menggunakan SW. Maka selanjutnya kita perlu mengeksplore hal2 apa yg harus ada pada software sehingga setelah menggunakan SW tsb kecintaan anak2 pd lingkungan meningkat.

Kita perlu tahu hal2 apa yg ada pada SW yg sekiranya bisa meningkatkan kinerja→ melalui pakar dunia anak, survey ke anak, dsb.

- ➔ Game
- ➔ Lagu anak2
- ➔ Animasi → cara penyajian
- ➔ Hadiah
- ➔ disain
- ➔ dsb

jadi banyak kemungkinan yg mempengaruhi kinerja SW. Selanjutnya kita perlu menentukan bagaimana cara mengukur cinta lingkungan → parameter apa yg dikatakan cinta lingkungan. Misalnya : kecintaan pada lingkungan ditunjukan alokasi waktu menyiram tanaman di halaman dalam 1 minggu, atau dalam satu bulan, berapa hari dialokasikan anak untuk kegiatan2 yg terkait dengan lingkungan, dsb.

Hal2 yg mungkin mempengaruhi tk kecintaan pd lingkungan dikenal sebagai peubah bebas/independent/penyebab. Sedangkan kecintaan pada lingkungan sebagai akibatnya, dikenal peubah responds/dependent. Dalam ini peubah responds diukur/diamati, sedangkan peubah bebas tidak diukur/tidak diamati. Dalam suatu penelitian, kita hanya memperhatikan beberapa peubah saja (tidak semuanya), dikarenakan berbagai hal. Jadi beberapa peubah bebas yg diduga berpengaruh pada respons di sebut faktor. Misalkan pada kasus di atas, faktor yg ingin diamati

adalah : Tampilan sistem (T) dan cara penyajian (P). Selain faktor, maka peubah-peubah bebas dikenal dengan lingkungan.

Faktor adalah peubah bebas yg pengaruhnya ingin dipelajari (apakah ada atau tidak, jika ada, bagaimana pengaruhnya). Karena faktor itu adalah peubah, maka dia mempunyai beberapa kemungkinan nilai (bisa numerik atau kategori).

Misal untuk contoh di atas ada 2 faktor :

Faktor pertama adalah Tampian (T). Berdasar pakar dunia anak serta bentuk2 interaksi manusia komputer, misal ada 3 pilihan yg bisa dipilih, yaitu t1, t2, dan t3.

Faktor kedua cara penyajian (P). Berdasar referensi dikenal 3 juga jenis penyajian, yaitu p1, p2, dan p3.

Oleh karena itu, kita perlu menentukan tampilan seperti ada dan penyajian seperti apa agar sistem yg dibuat efektif bisa meningkatkan kecintaan lingkungan dari anak TK setelah menggunakan sistem tersebut.

Faktor2 tsb mempunyai nilai2 yg bisa dipilih oleh pengembang sistem. Nilai-nilai faktor disebut level/taraf. Dalam sistem yg akan dibuat itu ada kombinasi2 pilihan dari level-level faktor yg ada. Dalam hal ini pilihannya adalah :

T1p1, t1p2, t1p3, t2p1, t2p2, t2p3, t3p1, t3p2, dan t3p3.

Jadi ada 3x3 = 9 kombinasi level2 dari faktor. Kombinasi level2 dari faktor dikenal dengan nama perlakuan/treatment.

Selanjutnya kita mengukur kecintaan?. Misalnya kecintaan thd lingkungan diukur dengan peningkatan waktu (jam per minggu) yg dialokasikan oleh anak2 TK dalam kegiatan2 yg berkaitan dengan lingkungan. Dalam hal perlu ditetapkan apakah yg diukur itu “anak” atau “kelompok anak/ atau kelas”. Jadi jika “anak”, maka pada satu anak diukur lama waktu untuk kegaitan terkait lingkungan selama 1 minggu, sebelum menggunakan sistem dengan perlakuan tertentu. Hal ini dilakukan kembaik setelah tsb mengunakan sistem dengan satu perlakuan tertentu. Repsondsnya adalah selisih antara setelah dengan sebelum.

Atau kalau yg dilakukan adalah : mengamati satu kelompok anak atau sebagai dari suatu kelas/kelompok setelah itu kelas ini diminta menggunakan sistem yg dibuat dengan perlakuan tertentu, dan pengamatan dilakukan setelah kelompok anak ini menggunakan sistem dan respondsnya adalah selisih antara setelah dengan sebelum.

Dalam hal ini, “anak” atau “kelas/kelompok anak” yg diminta mengggunakan satu sistem yg dibuat dengan perlakuan tertentu dikenal degan nama satuan percobaan. Untuk kasus “anak” sebagai satuan percobaan, maka pengamatannya juga anak itu. Sedangkan kalau pada “kelas” yg dikenai satu perlakuan maka pengamatan tidak harus 1 kelas, tetapi bisa 1 kelas atau sebagian dari kelas. Sebagain anak dalam kelas ini yg diamati disebut sebagai satuan pengamatan. Dalam kasus “kelas” sebagai satuan percobaan, maka satuan pengamatan adalah sebagian atau bisa juga sama dengan kelas.

selanjutnya adalah melakukan percobaan, yaitu bagaimana mencobakan perlakuan2 pada satuan percobaan. Dalam hal ini dikenal prinsip2 dalam percobaan :

a. ulangan : percobaan pemberian perlakuan pada satuan percobaan harus diulang pada satuan percobaan yg lain. Jika ulangan diinginkan r kali, maka banyaknya satuan percobaan yg diperlukan adalah r*banyaknya perlakuan. Untuk contoh di atas ada 9 perlakuan. Jika ingin diulang 3 kali, maka diperlukan 3*9=27 satuan percobaan.
b. Randomisasi : pemberian perlakuan dilakukan secara random, artinya : setiap satuan percobaan akan menerima suatu perlakuan tertentu dengan peluang yg sama.
c. Kontrol lingkungan : lingkungan dikontrol agar pengaruh ligkungan bisa dipisahkan dengan pengaruh dari faktor. Hal ini agar pengaruh faktor bisa dipelajari. Kalau tidak terpisah (confounding) maka pengaruh faktor tidak bisa dipelajari. Untuk itu perlu lingkungan yg homogen. Jika tdk bisa mendapatkan satuan percobaan yg homogen, maka perlu perancangan lingkungan untuk mengotrol lingkungan.

Sebgaai ilustrasi untuk kasus di atas, misalkan kita ingin ulangan 3 kali (setiap perlakuan akan dicobakan 3 kali). Oleh karean itu diperlukan 3*9=27 satuan percobaan.

Cara mengalokasikan 9 perlakuan ke setiap satuan percobaan dilakukan secara random (misalnya dengan cara mengunakanprogram membangkitkan bilangan random dari 1 sd 9). Jika muncul bilangan 1, maka satuan percobaan tsb diberikan perlakuan 1, jika muncul 2 maka stauan percobaan tsb diberikan perlakuan 2, dsb.

Perlakuan 1 : t1p1,
Perlakuan 2 : t1p2,
Dsb. t1p3, t2p1, t2p2, t2p3, t3p1, t3p2, dan t3p3

Mulai dibangkitkan bil random dari 1 sd 9 :
Yg muncul pertama adalah 3, maka satuan percobaan 1 diberi perlakuan 3 → t1p3
Dst, sehingga : semua satuan percobaan telah dimendapatkan perlakuan, dan semua perlakuan diulang 3 kali.

1 T1p3	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27

Proses ini diteruskan sd semua perlakuan telah dialokasikan kepada 27 satuan percobaan , dan masing2 perlakuan tepat diulang 3 kali. Misalnya hasil pengacakan perlakuan ke dalam 27 satuan percobaan adalah sbb:

1 t1p3	2 T3p2	3 T2p2	4 T1p3	5 T3p2	6 T3p2	7 T2p3	8 T1p1	9 ~~T3p2~~ t2a2
10 T2p1	11 T2p2	12 T3p1	13 T2p1	14 T1p1	15 T1p3	16 T2p3	17 T1p2	18 T1p1
19 ~~T1p3~~t3p1	20 T2p1	21 ~~T2p1~~t2p3	22 ~~T1p1~~t1p2	23 T3p3	24 T1p2	25 T3p3	26 T3p1	27 T3p3

Setelah semua satuan percobaan menerima perlakuan tertentu, maka selanjutnya dilakukan pengamatan untuk mendapatkan respons dari setiap satuan percobaan. Misalnya hasil observasinya adalah sebagai berikut :

1 T1p3 19	2 T3p2 21	3 T2p2 24	4 T1p3 18	5 T3p2 23	6 T3p2 20	7 T2p3 20	8 T1p1 23	9 ~~T3p2t3p2~~t2p2 22
10 T2p1 20	11 T2p2 25	12 T3p1 18	13 T2p1 22	14 T1p1 20	15 T1p3 21	16 T2p3 22	17 T1p2 22	18 T1p1 21
19 ~~T1p3~~t3p1 18	20 T2p1 19	21 ~~T2p1~~t2p3 19	22 ~~T1p1~~t1p2 20	23 T3p3 20	24 T1p2 19	25 T3p3 22	26 T3p1 16	27 T3p3 24

Berdasar data observasi di atas, kita ingin tahu :

Apakah perlakuan yg ada mempunyai pengaruh yg berbeda terhadap efektifitas sistem (artinya : apakah antara t1p1, t1p2, …, t3p3 pengaruhnya sama semua, atau tidak semua sama/dengan kata lain setidaknya ada satu pasang yg pengaruhnya berbeda).

Untuk mengtehaui hal ini, data perlu dianalisis, dengan menggunakan ANOVA (Analysis of Varaince). Analysis ini menguraikan ragam/variance dari pengamatan menjadi ragam/variance penyusunnya. Dalam hal ini, keragaman/variasi data karena 2 hal : perbedaan perlakuan juga karena perbedaan satuan percobaan (tetapi karena satuan percobaan dianggap homogen, maka variance dari satuan percobaan ini sebagai error/galat/sisaan). Dengan demikian, maka disusun tabel anova sbb:

Sumber keragaman (Source)	**Derajad bebas (degree of freedom)**	**Jumlah kuadrat (sum of square)**	**Rataan jumlah kuadrat RJK→ mencerminkan variance**	**F hitung**	**F Tabel (teoritis)**
Perlakuan	9-1=8	SS perlakuan	RJKP=SSP/dB=SSP/8 besar apa kecil? Jika besar maka minimal ada 1 pasang perlakuan yg	Hasil bagi antara var perlakukan dibagi dengan variance karena error, yaitu : RJKP/RJKE	F(taraf nyata uji (5%), dF yg dibagi, dF dari

			berbeda pengaruhnya. Kalau kecil: tidak ada perbedaan pengaruh secara nyata dari perlakukan thd efektifitas sistem. Bagaimana menentukan besar atau kecil? Perlu nilai batas. Batas tersebut direlatifkan terhadap errornya	Variance dibagi dengan variance (dalam teori statistika) berdistribusi F dengan derajad (dF yg dbagi, dF dari pembagi). Hasil bagi ini sebagai kriteria apakah besar atau kecil. Untuk menentukan besar atau kecil, maka dilihat dalam ditribusi F nilai hasil bagi ini ada di mana, apakah di sebelah kanan (lebih besar) atau ada di sebelah kiri (lebih kecil) dari nilai F teoritis. F teoritis didasarkan pada toleransi kesalahan kesimpulan yg masih kita tolerir (biasanya 5%). Untuk kasus, maka batas tsb adalah F(5%,8,18)=2.51	pembagi), pada contoh ini adalah F(5%,8,18), yaitu 2.51
Error	26-8=18	SS Error=SS Total-SS perlakuan	RJKE=SSE/dB=SSE/18		
Total	**27-1=26**	**SS Total**			

Bagaimana cara menghitung anova ini?

1 T1s3 19	2 T3s2 21	3 T2s2 24	4 T1s3 18	5 T3s2 23	6 T3s2 20	7 T2s3 20	8 T1a1 23	9 ~~T3s2t3s2~~t2s2 22
10 T2s1 20	11 T2s2 25	12 T3s1 18	13 T2s1 22	14 T1s1 20	15 T1s3 21	16 T2s3 22	17 T1s2 22	18 T1s1 21
19 ~~T1s3~~t3s1 18	20 T2s1 19	21 ~~T2s1~~t2s3 19	22 ~~T1s1~~t1s2 20	23 T3s3 20	24 T1s2 19	25 T3s3 22	26 T3s1 16	27 T3s3 24

Data di atas perlu disusun dalam struktur data seperti dibawah ini :

		Sajian			Jumlah
		p1	p2	p3	
Tampilan	t1	23 20 21	22 20 19	19 18 21	
		64	**61**	**58**	183
	t2	20 22 19	24 22 25	20 22 19	
		61	**71**	**61**	193
	t3	18 18 16	21 23 20	20 22 24	
		52	**64**	**66**	182
Jumlah		177	196	185	**558**

		Sajian			Jumlah
		p1	p2	p3	
Tampilan	t1	23	22	19	
		20	20	18	
		21	19	21	
		64	**61**	**58**	183
	t2	20	24	20	
		22	22	22	
		19	25	19	
		61	**71**	**61**	193
	t3	18	21	20	
		18	23	22	
		16	20	24	
		52	**64**	**66**	182
Jumlah		177	196	185	**558**

Dari data pengamatan/percobaan di atas, kita ingin tahu :

1. Apakah kinerja sistem dengan perlakuan tertentu, berbeda atau tidak dengan kinerja sistem kalau menggunakan perlakuan yg lain?
   ➔ Apakah perbedaan perlakuan memberikan dampak pada perbedaan kinerja

Jawaban ini diperoleh dari analisis Anova:

Sumber keragaman (Source)	**Derajad bebas (degree of freedom)**	**Jumlah kuadrat (sum of square)**	**Rataan jumlah kuadrat RJK→ mencerminkan variance**	**F hitung**	**F Tabel (teoritis)**
Perlakuan	9-1=8				
Error	26-8=18				
Total	**27-1=26**				

Untuk menjawab hal ini, data perlu diolah/dianalisis?. Teknik analisisnya menggunakan analisis sidik ragam (Analysis of Variance→ Anova). Berdasar data di atas, dilakukan perhitungan2 dan disajikan dalam tabel, disebut tabel Anova.

Sumber keragaman (Source)	**Derajad bebas (degree of freedom)**	**Jumlah kuadrat (sum of square)**	**Rataan jumlah kuadrat RJK→ mencerminkan variance**	**F hitung**	**F Tabel (teoritis)**
Perlakuan	9-1=8	74,67	74,67/8=9,33	9,33/2,41=3,88	
Error	26-8=18	118-74,67=43,33	43,33/18=2,41		
Total	**27-1=26**	**118**			

Faktor koreksi=FK= (jumlah obs)^2 dibagi banyaknya observasi = (558)^2/27

Jumlah kuadrat total = [jumlah (obs^2)] – Faktor Koreksi = 11650 - 11532 = 118

Jumlah Kuadrat perlakuan = {jumlah [(juml perlakuan)^2]/banyaknya obs tsb} – FK

= 11606.67 – 11532 = 74.67 → RJKP=74.67/8=9.33

Jumlah kuadrat error = JKT – JKP = 118 – 74.67 = 43.33 → RJKE=43.33/18=2.41

Sumber keragaman (Source)	**Derajad bebas (degree of freedom)**	**Jumlah kuadrat (sum of square)**	**Rataan jumlah kuadrat RJK→ mencerminkan variance**	**F hitung**	**F Tabel (teoritis)**
Perlakuan	9-1=8	74.67	9.33	9.33/2.41=3.88	F(5%,8,18)=2.51
Error	26-8=18	43.33	2.41		
Total	**27-1=26**	**118**			

Oleh karena F hitung > F tabel, maka dikatakan bahwa variance karena perlakuan besar. Artinya tidak semua perlakuan pengaruhnya sama, artinya minimal ada 1 pasang yg berbeda.

Perlakuan yg memberikan pengaruh berbeda ini apakah karena faktor T atau faktor p atau ada interaksi antara T dengan p. Dengan demikian maka anova ini perlu dideailkan lagi, menjadi sebagai berikut :

Sumber keragaman (Source)	**Derajad bebas (degree of freedom)**	**Jumlah kuadrat (sum of square)**	**Rataan jumlah kuadrat RJK→ mencerminkan variance**	**F hitung**	**F Tabel (teoritis)**
Perlakuan	9-1=8	74.67	9.33	9.33/2.41=3.88	F(5%,8,18)=2.51
T=tampilan	3-1=2	JKTam=8.22	4.11		
P=penyajian	3-1=2	JKSaj=20.22	10.11		
TxS (interaksi antara T dan S)	2x2=4	JKTamxSaj=74.67 -8.22-20.22=46.22	11.55	11.55/2.41=4.79	F(5%,4,18)=2.93
Error	26-8=18	43.33	2.41		
Total	**27-1=26**	**118**			

Cara menghitung masing2 jumlah kuadrat adalah :

JK Tampilan = jumlah[(tmp^2)/#obs] – FK = - 11532 = 8.22

JK Sajian = = - 11532 = 20.22

Yg diperhatikan adalah interaksi dahulu, sebab jjika ada interkasi, maka kita tdk bisa melihat masing-masing secara sendiri, tetapi harus bersamaan.

F hitng untuk interaksi adalah 4.79 > F tabel 2.93, artinya ada interaksi antara tampilan dengan sajian. Oleh karena itu yg perlu dilihat adalah secara bersamaan. Untuk itu perlu di hitung beda nyata terkecilnya (BNT) yaitu nilai beda pg paling kecil yg bisa dikatakan berbeda.

BNT= $t_{5\%/2,dB\ error}$ *s*sqrt(2/ulangan)

T tabel : t (0.05/2, dF 18)= 2.101

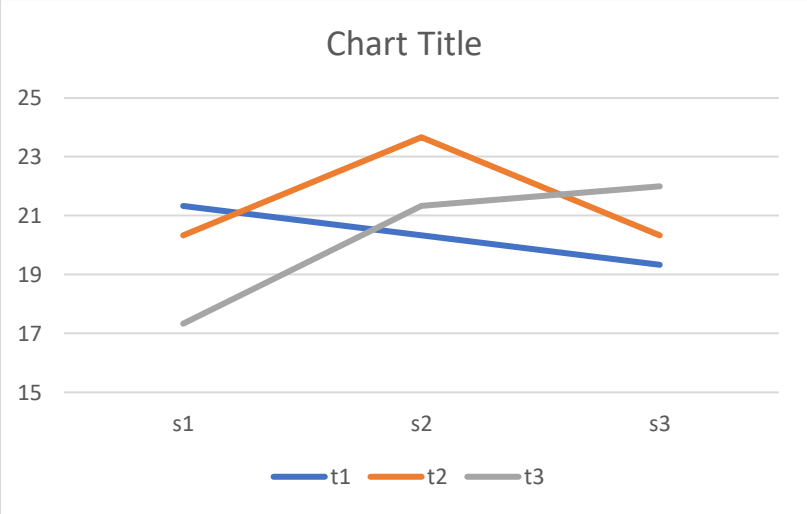
S=sqrt(2.41)

Ulangan=3

BNT=2.101*sqrt(2.41)*sqrt(2/3) = 2.66

Jika selisih antar rataan lebih besar dari 2.66 maka dikatakan berebda nyata. Jika tidak maka dikatakan tidak berbeda nyata. Interaksi antar faktor tsb bisa ditunjukkan dengan grafik di bawah, yg saling memotong (tidak sejajar). Dengan demikian, kita tidak bisa membandingkan t1 dengan t2 dan dengan t3 , bagus mana ? ini tidak bisa dilakukan secara sendiri. Tetapi harus dalam kaitannya dengan sajiannya pakai yg mana.

perl	rataan
t3p1	17,33
t1p3	19,33
t1p2	20,33
t2p1	20,33
t2p3	20,33
t1p1	21,33
t3p2	21,33
t3p3	22,00
t2p2	23,67



Sumber Keragaman	Derajad bebas	Jumlah kuadrat	Rataan Jumlah Kuadrat	F hitung	F Tabel
1. Perlakuan	a=14	175	…175/14= 12,5	5	F(…c,…d)
a. Algoritma (A)	4	…e….	20	…f….	F(…g,…h)
b. Tampilan (T)	…2….	…j….	…k….	…l….	F(…m,…n)
c. AxT	8	…o….	…p….	0.5	F(…q,…r)
2. Error/Galat	30	…t….	2.5		
Total	44	…u….			

PR : dikumpulkan sebelum UTS (dukumpul hari selasa) dan dengan tulis tangan, dikerjakan dengan kalkulator (Excell). File : metkuan_PR1_NRP), di email : agusbuono@apps.ipb.ac.id

Two-way ANOVA: obs versus t; s

Source	DF	SS	MS	F	P
t	2	8,222	4,1111	1,71	0,209
s	2	20,222	10,1111	4,20	0,032
Interaction	4	46,222	11,5556	4,80	0,008
Error	18	43,333	2,4074		
Total	26	118,000			

S = 1,552 R-Sq = 63,28% R-Sq(adj) = 46,96%

PR :

29. The article "An Analysis of Variance Applied to Screw Ma-chines" (*Industrial Quality Control,* 1956: 8–9) describes an experiment to investigate how the length of steel bars was affected by time of day (A), heat treatment applied (B), and screw machine used (C). The three times were 8:00 A.M., 11:00 A.M., and 3:00 P.M., and there were two treatments and four machines (a $3 \times 2 \times 4$ factorial experiment), resulting in the accompanying data [coded as 1000(length − 4.380), which does not affect the analysis].

B_1

	C_1	C_2	C_3	C_4
A_1	6, 9, 1, 3	7, 9, 5, 5	1, 2, 0, 4	6, 6, 7, 3
A_2	6, 3, 1, −1	8, 7, 4, 8	3, 2, 1, 0	7, 9, 11, 6
A_3	5, 4, 9, 6	10, 11, 6, 4	−1, 2, 6, 1	10, 5, 4, 8

B_2

	C_1	C_2	C_3	C_4
A_1	4, 6, 0, 1	6, 5, 3, 4	−1, 0, 0, 1	4, 5, 5, 4
A_2	3, 1, 1, −2	6, 4, 1, 3	2, 0, −1, 1	9, 4, 6, 3
A_3	6, 0, 3, 7	8, 7, 10, 0	0, −2, 4, −4	4, 3, 7, 0

Sums of squares include SSAB = 1.646, SSAC = 71.021, SSBC = 1.542, SSE = 447.500, and SST = 1037.833.

a. Construct the ANOVA table for this data.

b. Test to see whether any of the interaction effects are significant at level .05.

c. Test to see whether any of the main effects are significant at level .05 (i.e., H_{0A} versus H_{aA}, etc.).

d. Use Tukey's procedure to investigate significant differences among the four machines.