Sample Surveys
Financial Time Series I

Topic 2: Data Collection

Hung Chen

Department of Mathematics

National Taiwan University

10/17/2002

OUTLINE

1. Data Collection
   - experiment
   - survey
2. Motivated Example
   - nonresponse bias
   - Literary Digest poll of 1936
   - adjust for nonresponse
   - interviewer bias
3. Probability Sampling
   - Simple Random Sampling
     * probability analysis on the behavior of sample mean
   - Stratified Random Sampling
   - Cluster Sampling
   - Systematic Sampling
4. Experiment Design
5. Randomization
6. Observational Studies
7. Fishing Expeditions

Sampling from Populations: Sample Surveys

Where does the data come from?

- Much empirical data arises from *experiments*, in which the investigator interacts in some way with the units of observation and actually influences the conditions of the units leading to the measurements.
  Clinical trial on comparing effectiveness of treatment is such an example.
- Many other sets of data result from simply *observing*, that is, making a survey.
  - Usually one cannot observe every individual in the population, and often this would not even be desirable, for many individuals are similar.
  - Sampling is needed because of limited resources (time, money, etc).
    Gallup polls before elections, telephone surveys about commercial products
  - One does not need to eat the whole bowl to learn how the soup tastes; a spoonful will suffice, provided that the soup has been adequately stirred.

    - The "spoonful" is a sample from the bowl (population).
    - The "stirring" corresponds to drawing a random sample.

Sample surveys are especially useful in business and economics, market research, sociology, and industrial quality control.

- A survey often involves an *interview*, a *questionnaire*, or some sort of *inspection*.
  - Care must be taken in working out the details of administration of the interview, questionnaire, or inspection.
  - The planning of a survey should involve two kinds of professionals working together - a subject-matter specialist (economist, sociologist, production engineer) and a statistician.

Motivated Example

Suppose that one wants to draw a random sample from among the households in a county to find out how many hours a day children watch television.

- Individual children are the "units of observation."
  Question: How do we sample it?
- Households are the "sampling units."
  - The frame is a physical list of the sampled population.
    In this example, we have the household registration record.
  - Complication arising from the sampling unit is different from units of observation: Each sampling unit contains one, more than one, or no units of observation.
- The list of household registration records on the county is a"frame."
- The"population sampled" consists of the children in the county who live in households.

- The "target population" is the set of all children who live in the county.
  In general, the target population is the population about which it is desired to make inferences.
- Any difference between the population sampled and the target population is a potential source of **bias**.
- Question: Suppose that the sampling unit is households on the county tax roll. Do you forsee any *bias*?

Many problems would arise in carrying out the survey of TV-viewing habits.
How should the survey be designed?

- For how many days should each child be observed?
  Suppose that it is decided to monitor the TV viewing of the children for two-week periods be chosen.
  The first question to be settled is *Which two week should be chosen?*
  - Should different two-week periods be chosen for different subsamples of children?

- How can one ensure that the parents keep an accurate record of how long their children watch TV?
- If there is more than one child in a household, should all of them be included in the sample?
- Should some inducement to participate be offered?

Nonresponse bias

- When a survey involves an expenditure of time or effort on the part of those selected to be in the sample, there are almost always some individuals who refuse to respond.
- When the sampling unit is a household, and interviewer may find no one home when he or she calls a household designated to be in a sample.
- Such failures to obtain information from each unit meant to be sampled can introduce some bias onto the results when those not responding differ systematically from those responding.

Why do we need to introduce the term *nonresponse bias*?

- It means, in effect, that the population sampled differs from the target population, inasmuch as the population sampled consist only of those individuals in the target population who are willing to respond or can be induced to do so.
- If, with respect to the characteristic of interest in the study, the population of persons who do not respond differs from those who do, there is a nonresponse bias.

The Literary Digest poll of 1936

- most famous flawed surveys
- It predicted a 57% to 43% victory for Republican Alfred Landon over incumbent president Franklin Roosevelt.
- Questionnaires were limited to about 10 million voters, who were selected from lists such as telephoon books and club memberships.
- Approximately 2.4 million of the questionnaires were returned.
- Two intrinsic problems:
  - noneresponse: those who did not respond may have voted differently from those who did
  - selection bias: even if all 10 million voters had responded, they would not have constituted a random sample
    Those in lower socioeconomic classes(who were more likely to vote for Roosevelt) were less likely to have telephone service or belong to clubs and thus less likely

to be included in the sample than were wealthier voters.

Adjust for nonresponse bias: statistical method

- Nonresponse cannot be avoided!
- Can we *reduce* the impact of nonresponse?
- Idea: modelling
  - In order for any such adjustment to be possible, however, there must be a second try, in which at least some of those who did not respond the first time do in fact respond.

Now we discuss how a nonresponse can be converted into a response when a second try is being done.

- A second try for getting response:
  - If no one in a household is at home at the time the interviewer calls, the interviewer can return (a "call-back").
  - If a questionnaire mailed to a respondent is not returned, the respondent can be tried by telephone.
  - Typically, the second attempt is more expensive than the first (for example, the

households may be more scattered), and the response on the second round may not be complete.

Now we use an example of using modeling to adjust for nonresponse.
Suppose that of a sample of 100 taken in March to estimate the unemployment rate in the country there were 10 nonrespondents.

- Of the 90 respondents, 5 were unemployed, and the unemployment rate was estimated as $\hat{p}_1 = 5/90 = 0.056$. This is an estimate of the unemployment rate in the population of respondents.
- The 10 nonrespondents were revisited in April and asked of they were employed in March; 8 of them responded, 3 saying that they were unemployed in March.
- The estimate of the unemployment in the population of nonrespondents is $\hat{p}_2 = 3/8 = 0.375$.

Idea:

- Classify the population in two subpopulations: the population of respondents and the

population of nonrespondents.

- Let p_1 and p_2 denotes the unemployment rates of the population of respondents and the population of nonrespondents.
- Set π to be the proportion of the population of respondents.
- The overall unemployment rate p is then defined as $\pi p_1 + (1 - \pi)p_2$.
- An estimate of π is 90/10.
  The corresponding estimate of unemployment rate is $0.9\hat{p}_1 + 0.1\hat{p}_2 = 0.088$.
  This is considerably higher than the initial estimate of 5.6%, due to the high rate of unemployment among the nonrespondents.

Interviewer bias

It results when different persons in the sample are questioned by different interviewers.

- This may occur because interviewers interpret the instructions given by the director of the survey in slightly different ways, or simply because people react in different ways to different interviewers.
- Care should be taken to minimize these effects beforehand.
- Sometimes statistical methods, such as *analysis of variance*, can be employed after the fact to assess the extent of interviewer bias and adjust for it.
- One-way ANOVA

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

  - α_i is used to model the effect due to the interviewers.
  - Suppose that each interviewer is only interview two peoples. Can we still estimate μ correctly?

Probability Sample

- How do we assess the accuracy and precision of estimators resulting from the various sampling methods?
- For such assessment to be possible, the sample drawn must be a probability sample, a sample drawn in such a way that he investigator knows the probability that any individual unit of observation will be included in the sample.
- Probability sampling methods: simple random sampling, stratified random sampling, cluster sampling, and systematic subsampling with random starts.
  In each case some random device plays a role in determining which members of the population shall be included in the sample.

Simple Random Sampling

- Aim: Obtain information about large populations by examining only a portion. Examples include traffic, tax audits, quality control, census preparation, and etc.
- To minimize sampling bias, it is impartial and objective to use probability methods to sample from a population.
  - Random sampling guards against investigator bias, consider election polls.
  - Above all, this randomness, allows an estimate of the error, (we can even design the sample size necessary to obtain a given precision).
- A simple random sample of size n is a sample of n drawn in such a way that all sets of N units in the population have the same chance of being in the sample.
- There are $C(N, n)$ such sets where N is the population size.
  - Sampling without Replacement: Consider taking a random sample such that all the

individuals in the sample are different.
If we want a sample of 2 individuals from a population of 4, the first is selected randomly. Suppose 3 appears.
That leaves 1, 2, and 4 in the population. Then one of these three is drawn randomly. The result determines the second member of the sample.

- The fact that all these sets are equally likely to be chosen implies in particular that all units have the same chance of being included.
- Implementation:
  - A frame, a list of all units with one of the numbers from 1 to N assigned to each, is constructed.
  - A random sample can then be drawn.
- How is it being done? Need random devices.
  - If the sample size is not very large, random numbers can be taken from a table of random numbers such as in your high school textbook.
  - We can obtain a random sample from a

deck of 52 playing cards by shuffling the deck adequately and dealing a specified number of cards from the top of the deck. A random sample from the population of 52 weeks in the year can be obtained by identifying each week with a different card and then dealing a sample of the desired size. Unfortunately, the just described mechanical devices for drawing at random is subject to biases. A deck of cards may be shuffled inadequately.

– Consider computer random number generator, based on uniform random number generator. A computer can be used to obtain a stream of numbers that behave like random numbers (often called pseudorandom numbers).

– Since the frame of N units is finite, the sampling is done without replacement. The set of n random numbers determines the n sampling units to be included in the sample. Characteristics of the sampled units are then obtained by interview, written questionnaire, of direct mea-

surement.

- Composition of the sample is random (the labels are random) implies that the sample mean, the sample total... are random variables.
- The population mean is a number, the sample mean is a random variable whose accuracy as an estimate can be evaluated by a probabilistic analysis.

Behavior of Sampling Mean under SRS

- Set-up
  - In the case of a numerical variable, let the numbers $y_1, y_2, \ldots, y_N$ represent the values of the variable for the N individuals in the population.
  - The population mean μ is $\Sigma_k\, y_k/N$, and the population variance is $\sigma^2 = \Sigma_k (y_k - \mu)^2/N$.
  - Let the values of the variable for the n individuals in the sample be $x_1, x_2, \ldots, x_n$, where x_i is the value of the variable for the ith individual in the sample, $i = 1, 2, \ldots, n$.
  - The symbol x_1 bears no special relation to y_1; if $n = 3$ and the sample of three individuals consists of the individuals numbered 13, 17, and 8 in the frame, then $x_1 = y_{13}$, $x_2 = y_{17}$, and $x_3 = y_8$.)
- How do we describe x_1?
  - It is a random variable.
  - X_1 can take values $y_1, y_2, \ldots, y_N$ each with probability N^{-1}.

– X_i's distribution is called the sampling distribution.

- Is $\bar{x}$ close to μ?
- Standard analysis: Find the expectation and variance of $\bar{X}$.
- The sample mean $\bar{x}$ is an unbiased estimate of the population mean μ, that is, $E(\bar{X}) = \mu$ (whether sampling is with or without replacement).
- The variance of the sample mean for sampling without replacement is

$$\sigma^2_{\bar{X}} = \frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

  The finite population correction factor for the variance $(N-n)/(N-1)$ is approximately $1-n/N$. When the fraction sampled n/N is small, this correction is negligible.
- What is the definition of mean and variance?

Stratified Random Sampling

- Randomness in drawing a sample, which is essential to obtaining unbiased estimates, results in sampling variability.
- In some situations the variability can be reduced without introducing bias by using other information about the population.
- Suppose that sample of engineers employed in a large corporation is to be drawn to estimate the mean salary of all engineers.
  - An individual's salary depends heavily on his or her corporate function-whether the position is supervisory or nonsupervisory.
  - If the listing of engineers (a frame) is such that the function of each is identified, it is possible to draw a random sample of each type of engineer and estimate the average salary of each type of engineer in the entire corporation.
  - A weighted average of these two estimates yields an estimate of the average salary of all engineers.

  - The sampling variability of this procedure is less than that of simple random sampling because the variability within the set of supervisory engineers and with in the set of nonsupervisory engineers is less than the variability among all the engineers.

- A stratum is a subpopulation.
- A set of strata is a collection of subsets of individuals in the population such that each individual belongs to one and only one such subset.
- To use "stratified sampling," it is essential that a frame be available for each stratum; this implies that the sizes if the strata are known.

Consider the example with two strata.

- Two strata: supervisory and nonsupervisory.
- Let N denote the size of the total population of engineers and μ the mean salary in this population.

- Suppose that some N_1 of these N have supervisory positions; let μ_1 denote the mean income in this stratum.
  The other N_2 $(= N - N_1)$ engineers have nonsupervisory positions, and their mean income is μ_2.
- The parameter μ is equal to $(N_1\mu_1+N_2\mu_2)/N$.
- The parameters μ, μ_1 and μ_2 are, of course, unknown to the investigator, but N, N_1, and N_2 are known from information on the sampling frame (the list of engineers).
- The population has been stratified by occupational position (supervisory versus nonsupervisory).
  - One takes a random sample of specified size n_1 from the N_1 supervisory engineers and a random sample of size n_2 from the N_2 nonsupervisory engineers.
  - The estimates of the strata μ_1 and μ_2 are the sample means $\hat{x}_1$ and $\hat{x}_2$ of the samples from the two strata.
  - As an estimate of μ, we take the weighted

average, namely

$$\frac{N_1}{N}\hat{x}_1 + \frac{N_2}{N}\hat{x}_2.$$

- $E(\hat{\mu}) = \mu$ and

  $$Var(\hat{\mu}) = \left(\frac{N_1}{N}\right)^2 \frac{N_1 - n_1}{N_1 - 1}\frac{\sigma_1^2}{n_1} + \left(\frac{N_2}{N}\right)^2 \frac{N_2 - n_2}{N_2 - 1}\frac{\sigma_2^2}{n_2}.$$

- What can we say about σ_1^2 and σ_2^2 for this particular example?

Cluster Sampling

Cluster sampling refers to sampling "cluster" of potential respondents and then sampling respondents in the clusters in the sample.

- To determine the total number of unemployed in a city, for example, one might consider city blocks as the clusters and households as the "respondents."
- A sample of city blocks is taken, using a map of the city to number the blocks.
- In each block the households are enumerated, and a random sample of households is taken in each block in the sample.
- The total number of unemployed in a sampled block is estimated from the sample of households in that block.
- In turn, the total number of unemployed in the city is estimated from these estimated block totals.

How do we estimate the population mean?

- An unbiased estimate of the population mean is obtained by dividing the estimated total by the number of units in the population.
- Sampling variability arises from two source: the sampling of clusters and the sampling of units with in clusters.
- The formula for the variance of an estimate depends on the rule for sample size within the sampled cluster. For example, the sample sizes may be the same in all clusters or they may be a fixed proportion of the cluster sizes. Refer to Cochran (1977) for details.

Advantages of cluster sampling

- When the clusters represent geographically compact sets of units, as in the above illustration, with cluster sampling the interviewers may spend more time in interviewing than traveling.
- Also, a frame for clusters (for instance, blocks) may be available, making enumeration of clusters feasible, while preparation of a frame for the entire population of units is not practical.

- In a sense the clusters are strata, since each individual or unit belongs to one and only one cluster. The greater flexibility here results because clusters (or strata) are themselves sampled.

Systematic Sampling with a Random Start

The idea of systematic sampling is to take every tenth name on a list, or check every fifth car passing a toll booth, or review every twentieth file folder in a drawer.

- The method is appealing because it is easy to carry out and it spreads the "sample" out through the population.
- It is clearly not random.
- To add an element of randomness-which is necessary to obtain unbiased estimates-one may select the starting point at random.

The sample mean is unbiased (because the start is random), but the precision of $\bar{x}$ depends on how the characteristic under observation varies as we go through the frame.

- If the population is the 365 days of the year, the frame is the calendar.
- When the sampling interval is 7 and $n = 52$, we get a systematic sample that is based on the same day of the week over the entire year.

- The method is good for estimating the average hours of daylight per day over the year.
- But poor for estimation the average hours of work per day over the year.

- Suppose again that we are sampling households and that the fame lists houses in the following order:
  - Think of houses in a particular community which are located on a chess board arrangement.
    There are three avenues and four streets. Avenue A from west to east; Avenue B from west to east on the north side of the street, Avenue B from west to east on the south side of the street, etc.
    First Street from north to south; Second Street
  - The blocks are oblong and narrow between avenues, so that all houses face on avenues.
  - Within each block, there are six houses among which three houses face the same

avenue.

- Suppose that the ordering in the frame corresponds to the cost of the houses; house 1 is least expensive and house 36 most expensive.

  Then systematic sampling gives us a sample of households that is varied and representative as far as cost of house (and consequently family income) is concerned. In this case, systematic sampling performs toughly likely stratified sampling, where the strata are defined in terms of income.

- On the other hand, if the variation in the population is related to the sampling interval, then systematic sampling can be much less precise than simple random sampling; at worst it can be equivalent to having only a sample of size 1.

  If the corner houses are most expensive, and our sample interval is 3, then we get a very nonrepresentative sample: either all corner houses or all middle-of-the-block houses.

How do we use a probabilistic analysis on this

sampling scheme?

- Denote by m the starting number, the integer chosen at random from $1, 2, \ldots, \ell$.
- The quantity m is a random variable that takes on the values $1, 2, \ldots, \ell$, each with probability $1/\ell$.
  - Only ℓ different samples are possible.
  - Although all individuals have the same chance of being included in the sample, not all possible sets of n have the same chance.
  - ℓ sets have probability $1/\ell$ each. All the other sets have zero probability.
- The observed $\bar{x}$ is a random sample of 1 from the population of ℓ values, $\bar{y}_1, \ldots, \bar{y}_\ell$.
- The variance is $\Sigma_b(\bar{y}_b - \mu)^2/\ell$ which is unknown.
  It cannot be estimated from the sample because the sample is effectively a sample of 1 from the population $\bar{y}_1, \ldots, \bar{y}_\ell$.
- The investigator cannot assess variability, carry out tests of significance, or construct

confidence intervals.

Experiment Design

Example 1. Mammary Artery Ligation

- A person with coronary disease suffers from chest pain during exercise because the constricted arteries cannot deliver enough oxygen to the heart. The treatment of ligating the mammary arteries enjoyed a brief vogue; the basic idea was that ligating these arteries forced more blood to flow into the heart. This procedure had the advantage of being quite simple surgically, and it was widely publicized in an article Reader's Digest (Ratcliffe 1957).
- Two years later, the results of a more careful study (Cobb et al. 1959) were published.
- In this study, a control group and an experimental group were established in the following way.
  - When a prospective patient entered surgery, the surgeon made the necessary preliminary incisions prior to tying off the mammary artery. At that point, the surgeon

opened a sealed envelope that contained instructions as to whether to *complete the operation* by tying off the artery.

  - Neither the patient nor his attending physician knew whether the operation had actually been carried out.
  - The study showed essentially no difference after the operation between the control group (no ligation) and the experimental group (lifation), although there was some suggestion that the control group had done better.

- The Ratcliffe and Cobb studies differ in that in the earlier one there was no control group and thus no benchmark by which to gauge improvement.
- The reported improvement of the patients in this earlier study could have been due to the placebo effect, which we discuss next.
- The design of the later study protected against possible unconscious biases by randomly assigning the control and experiment groups and by concealing from the patients and their

physicians the actual nature of the treatment.

- Such a design is called a double-blind, randomized controlled experiment.

The Placebo Effect: The placebo effect refers to the effect produced by any treatment, including dummy pills (placebo), when the subject believes that he or she has been given an effective treatment.

- The possibility of a placebo effect makes the use of a blind design necessary in many experimental investigations.
- The placebo effect may not be due entirely to psychological factors, as was shown in an interesting experiment by Levine, Gordon, and Fields (1978).
  - A group of subjects had teeth extracted.
  - During the extraction, they were given nitrous oxide and local anesthesia.
  - In the recovery room, they rated the amount of pain they were experiencing on a numerical scale.

- Two hours after surgery, the subjects were given a placebo and were again asked to rate their pain.
- An hour later, some of the subjects were given a placebo and some were given naloxone, a morphine antagonist.
- It is known that there are specific receptors to morphine in the brain and that the body can also release endorphins that bind to these sites.
  Naloxone blocks the morphine receptors.
- In the study, it was found that when those subjects who responded positively to the placebo received naloxone, they experienced an increase in pain that made their pain levels comparable to those of the patients who did not respond to the placebo.
- The implication is that those who responded to the placebo had produced endorphins, the actions of which were subsequently blocked by the naloxone.

• The use of controls does not in itself ensure

a valid experimental design; the allocation of subjects to treatment and control groups should be done by randomization.

- Wilson (1952) relates a story of a test of a pill to prevent seasickness. Prior to the voyage, the use of controls was carefully explained to the captain. On returning, he reported that the pills had been a marvelous success?most of the controls had been sick and the treatment group had fared well. But when further questioned, he revealed that he had given the pills to his crew and had used the passengers as controls.

Example 2: The Lanarkshire Milk Experiment

- The importance of the randomized assignment of individuals (or other experimental units) to treatment and control groups is illustrated by a famous study known as the Lanarkshire milk experiment.
- In the spring of 1930, an experiment was carried out in Lanarkshire, England to determine the effect of providing free milk to

schoolchildren.

- In each participating school, some children (treatment group) were given free milk and others (controls) were not.
- The assignment of children to control or treatment was initially done at random; however, teachers were allowed to use their judgment in switching children between treatment and control to obtain a better balance of undernourished and well-nourished individuals in the groups.
- A paper by Gosset (1931), who published under the name Student (as in Student's t test), is a very interesting critique of the experiment.
- An examination of the data revealed that at the start of the experiment the controls were heavier and taller.
- Student conjectured that the teachers, perhaps unconsciously, had adjusted the initial randomization in a manner that placed more of the undernourished children in the treatment group.

- A further complication was caused by weighing the children with their clothes on.
- The experiment data were weight gains measured in late spring relative to early spring or late winter.
- The more well-to-do children probably tended to be better nourished and may have had heavier winter clothing than the poor children. Thus.
- The well-to-do children's weight gains were vitiated as a result of differences in clothing, which may have influenced comparisons between the treatment and control groups.

Example 3. The Portocaval Shunt

- Cirrhosis of the liver, to which alcoholics are prone, is a condition in which resistance to blood flow causes blood pressure in the liver to built up to dangerously high levels.
- Vessels may rupture, which may cause death. Surgeons have attempted to relieve this condition by connecting the portal artery, which feeds the liver, to the vena cava, one of the

main veins returning to the heart, thus reducing blood flow through the liver. This procedure, called the Portacaval shunt, had been used for more than 20 years when Grace, Muench, and Chalmers (1966) published and examination of 51 studies of the method.

- They examined the design of each study (presence or absence of a control group and presence or absence of randomization) and the investigators' conclusions (categorized as markedly enthusiastic, moderately enthusiastic, or not enthusiastic).
- The results are summarized in the following table, which speaks for itself:

Design	Ma
No Controls	24
Nonrandomized Controls	10
Randomized Controls	0

- The differences between the experiments that used controls and those that did not is not entirely surprising, since the placebo effect was probably operating.
- The importance of randomized assignment

to treatment and control groups is illustrated by comparing the conclusions for the randomized and nonrandomized controlled experiments.

- Randomization can help to ensure against subtle unconscious biases that may creep into an experiment. For example, a physician might tend to recommend surgery for patients who are somewhat more robust than the average. Articulate patients might be more likely to have an influence on the decision as to which group they are assigned to.

Example 4. FD&C Red No. 40

- This discussion follows Lagakos and Mosteller (1981).
- During the middle and late 1970's, experiments were conducted to determine possible carcinogenic effects of a widely used food coloring, FD&C Red No. 40.
- One of the experiments involved 500 male and 500 female mice.

- Both genders were divided into five groups: two control groups, a low-dose group, a medium-does group, and a high-does group.
- The mice were bred in the following way: Males and females were paired and before and during mating were given their prescribed dose of Red No. 40.
- The regime was continued during gestation and weaning of the young.
- From litters that had at least three pups of each sex, three of each sex were selected randomly and continued on their parents' dosage throughout their lives.
- After 109-111 weeks, all the mice still living were killed.
- The presence or absence of reticuloendothelial tumors was of particular interest.
- Although there were significant differences between some of the treatment groups, the results were rather confusing.
  - For example, there was a significant difference between the incidence rates for

the two male control groups, and among the males the medium-does group had the lowest incidence.

- Several experts were asked to examine the results of this and other experiments.
- Among them were Lagakos and Mosteller, who requested information on how the cages that housed the mice were arranged.
  - There were three racks of cages, each containing five rows of seven cages in the front and five rows of seven cages in the back.
  - Five mice were housed in each cage.
  - The mice were assigned to the cages in a systematic way: The first male control group was in the top of the front of rack 1; the first female control group was in the bottom of the front of rack 1; and so on, ending with the high-does females in the bottom of the back of rack 3.
  - Lagakos and Mosteller showed that there were effects due to cage position that could

not be explained by gender of by dosage group.

  - A random assignment of cage positions would have eliminated this confounding.
  - Lagakos and Mosteller also suggested some experimental designs to systematically control for cage position.

- It was also possible that a litter effect might be complicating the analysis, since littermates received the same treatment and littermates of the same sex were housed in the same or contiguous cages.
- In the presence of a litter effect, mice from the same litter might show less variability than that present among mice from different litters.
- This reduces the effective sample size?in the extreme case in which littermates react identically, the effective sample size is the number of litters, not the total number of mice.
- One way around this problem would have been to use only one mouse from each litter.

- The presence of a possible selection bias is another problem.
  Since mice were in the experiment only if they came from a litter with at least three males and three females, offspring of possibly less health parents were excluded.
  This could be a serious problem since exposure to Red No. 40 might affect the parents' health and the birth process.

- If, for example, among the high-does mice, only the most hardy produced large enough litters, their offspring might be hardier than the controls' offspring.

Randomization

- As well as guarding against possible biases on the part of the experiment, the process of randomization tends to balance any factors that may be influential but are not explicitly controlled in the experiment.
- Time is often such a factor; background variables such as temperature, equipment calibration, line voltage, and chemical composition can change slowly with time.
- In experiments that are run over some period of time, therefore, it is important to randomize the assignments to treatment and control over time.
- Time is not the only factor that should be randomized, however. In agricultural experiments, the positions of test plots in a field are often randomly assigned. In biological experiments with teat animals, the locations of the animals' cages may have an effect, as illustrated in the preceding section.

Observational Studies, Confounding, and Bias in Graduate Admission

- It is not always possible to conduct controlled experiments or use randomization.
- In evaluating some medical therapies, for example, a randomized, controlled experiment would be unethical if one therapy was strongly believed to be superior.
- For many problems of psychological interest (effects of parental modes of discipline, for example), it is impossible to conduct controlled experiments.
  In such situations, recourse is often made to observational studies.
- Hospital records may be examined to compare the outcomes of different therapies, or psychological records of children raised in different ways may be analyzed.
- Although such studies may be valuable, the results are seldom unequivocal.
- Since there is no randomization, it is always possible that the groups under comparison

differ in respects other than their "treatments."

- As an example, let us consider a study of gender bias in admissions to graduate school at the University of California at Berkeley (Bickel and O' Connell 1975).
  - In the fall of 1973, 8442 men applied for admission to graduate studies at Berkeley, and 44% were admitted; 4321 women applied, and 35% were admitted.
  - If the men and women were similar in every respect other than sex, this would be strong evidence of sex bias.
  - This was not a controlled, randomized experiment, however; sex was not randomly assigned to the applicants.
  - As will be seen, the male and female applicants differed in other respects, which influenced admission.
  - In the percentages admitted are compared, women do not seem to be unfavorably treated.
  - But when the combined admission rates

for all six majors are calculated, it is found that 44% of the men and only 35% of the women were admitted, which seems paradoxical.

- The resolution of the paradox lies in the observation that the women tended to apply to majors that had low admission rates (C through F) and the men to majors that had relatively high admission rates (A and B).

  This factor was not controlled for, since the study was observational in nature; it was also "confounded" with the factor of interest, sex; randomization, had it been possible, would have tended to balance out the confounded factor.

- Confounding also plays an important role in studies of the effect of coffee drinking.
  - Several studies have claimed to show a significant association of coffee consumption with coronary disease.
  - Clearly, randomized, controlled trials are not possible here?

A randomly selected individual cannot be told that he or she is in the treatment group and must drink 10 cups of coffee a day for the next five years.

- It is known that heavy coffee drinkers also tend to smoke more than average, so smoking is confounded with coffee drinking.
- Hennekens et al. (1976) review several studies in this area.

Fishing Expeditions:

- Another problem that sometimes flaws observational studies, and controlled experiments as well, is that they engage in "fishing expeditions."
- For example, consider a hypothetical study of the effects of birth control pills.
  - In such a case, it would be impossible to assign women to a treatment or a placebo at random, but a nonrandomized study might be conducted by carefully matching controls to treatments on such factors as age and medical history.
  - The two groups might be followed up on for some time, with many variables being recorded for each subject such as blood pressure, psychological measures, and incidences of various medical problems.
  - After termination of the study, the two groups might be compared on each of these variables, and it might be found, say, that there was a "significant finding" in the incidence of melanoma.

- The problem with this "significant finding" is the following.
  - Suppose that 100 independent two sample t tests are conducted at the 0.05 level and that, in fact, all the null hypotheses are true.
  - We would expect that five of the test would procedure a "significant" result.
  - Although each of the tests has probability 0.05 of type I error, as a collection they do not simultaneously have $\alpha = 0.05$.
  - The combined significance level is the probability that at least one of the null hypotheses is rejected:

$$\begin{aligned} & P(\text{at least one } H_0 \text{ rejected}) \\ &= 1 - P(\text{no } H_0 \text{ rejected}) \\ &= 1 - 0.95^{100} = 0.994 \end{aligned}$$

    Thus, with very high probability, at least one "significant" result will be found, even if all the null hypotheses are true.

There are no simple cures for this problem.

In general, we consider the following three possibilities:

- Regard the results of a fishing expedition as merely providing suggestions for further experiments.
- The data could be split randomly into two halves, one half for fishing in and the other half to be locked safely away, unexamined. "Significant" results from the first half could then be tested on the second half.
- Conduct each individual hypothesis test at a small significance level.
  To see how this works, suppose that all null hypotheses are true and that each of n null hypotheses is tested at level α.
  - Let R_i denote the event that the ith null hypothesis is rejected.
  - Let α^* denote the overall probability of a type I error.
  - Then

$$\begin{aligned} \alpha^* &= P\{R_1 \text{ or } R_2 \text{ or } \cdots \text{ or } R_n\} \\ &\leq P(R_1) + P(R_2) + \cdots + P(R_n) = n\alpha \end{aligned}$$

– Thus, if each of the n null hypotheses is tested at level α/n, the overall significance level is less than or equal to α. This is often called the Bonferroni method.