4 Producing Data

4.1 First Steps

Anecdotal Evidence:
selected individual cases that are striking in some way — however, these cases often are not representative for a larger group of cases

Available Data:
data produced in the past for some specific purpose that is now generally available (and perhaps useful to answer a present question)

Sample:
a (randomly selected) part of the population — a sample (when properly obtained) allows us to draw conclusions on the entire population

Census:
obtaining information on every subject (or object) in the population

Advantages of Sample vs Census:

- sample usually costs less
- sample faster to obtain
- census sometimes impossible to obtain (e.g., entire population is unknown, e.g., all fish in Bear Lake or all trees in Utah)
- when obtaining data, we may destroy the subject (or object) of interest (e.g., when determining the life expectancy of light bulbs) — in case of a census, there might be no operational unit (e.g., light bulb) left after we are done with collecting the data and the statistical result would be immediately meaningless

Exploratory Data Analysis: (now also called Visual Data Mining)
the data analyst often has no specific question in mind; he/she uses available data and tries to find structure, patterns, and relationships in the data

Sources for Data:
Internet, Libraries, Publications of Federal Agencies, etc. — see Murdoch handout for data sources on the Web

4.2 Design of Experiments

Observational Study: observes individuals and measures variables of interest but does **not** influence the responses

Experiment: imposes some influence (called **treatment**) on individuals to observe their responses

Experimental Units: individuals on which an experiment is conducted (**subjects** if humans, or objects if particular items)

Treatment: specific experimental condition applied to the experimental units

Factor: explanatory variable in an experiment

Level: a specific value that can be taken by one of the factors

Placebo Effect: **placebos** are often used in medical experiments; placebos are dummy drugs that have no physical effect; however, often people feel better if they take any drug even if it only contains water or sugar ... — when this happens, this is called **placebo effect**; in a medical experiment, one half of the patients gets the new drug and the other half gets the placebo; the question of interest is if there is any difference in the recovery process of the two groups of patients

Control Group: group of patients that receives a sham (placebo) treatment to control the effects of lurking variables on the outcome of an experiment

Randomization: use of chance (e.g., random digit tables or computer software) to divide experimental units into two (or more) groups

Completely Randomized Experiment (or **Completely Randomized Experimental Design**): all experimental units are allocated at random among all the treatments

Statistical Significance: an outcome too large (or too small) which cannot be explained by pure chance

Replication: repetition of treatments on a large number of experimental units to allow the systematic effect of the treatments to be seen

Principles of Experimental Design:

(i) Control effects of lurking variables on the response, most simply by comparing several treatments.

(ii) Use randomization to assign experimental units by chance to treatments.

(iii) Use replication of the experiment on many experimental units to reduce variation by chance in the results.

Bias: a study is biased if it *systematically* favors certain outcomes, e.g., if we are likely to have people with a particular income, gender, or ethnic background in out study, it is very likely that this study will be biased towards this group of people and will not represent the true population

Hidden Bias: no experimental unit should be influenced in either way; a person that receives a placebo should be handled in exactly the same way as a person that receives the real treatment

Double–blind Experiment: neither the subjects nor the people conducting the experiment know which treatment is assigned to which subject; this information is only known to the people that analyze the experiment and is revealed to the subjects only after the experiment is over

Block: a group of experimental units that are similar in such a way that we can expect that these similarities will effect the response to treatments (e.g., gender)

Block Design: random assignment of units to treatments is carried out separately within each block

Matched Pairs Design: comparison of just two treatments based on almost identical experimental units (e.g., twins)

Use of Random Digit Tables

Random Digit Tables contain digits from 0 to 9 in a "random" arrangement. To make use of random digit tables, we first number the individuals in the population in any order (e.g., alphabetically if names). Then we use a random digit table (such as Table B in Moore/McCabe) to select individuals at random.

Example: Stat 2000 Students

Assume a large section of Stat 2000 has 68 students named Aaa to Zzz. We first label the students in alphabetical order:

Number	Name
01	Aaa
02	Bbb
⋮	⋮
68	Zzz

We want to interview 5 randomly selected students, i.e., we want to obtain a sample of 5 students.

Do achieve this goal, we select any line of a random digit table as the starting point (e.g., by blindly pointing with our finger on the table). Then we split off 2 digit numbers, reading the digits row–wise from left to right and starting at the beginning of the next row once we reached the end of a row. If a number is bigger than the number of individuals in the population, we just take the next number. We also take the next number if a number has been used before.

In the current case, finger got me to line 129:

36759 58984 68288 22913 18638 54303

Based on this starting row, we would select the following 5 students: ...

Note that a different starting row would give us a different sample.

4.3 Sampling Design

Voluntary Response Sample: based on people that decide that they want to answer to a general appeal they received by mail or on the phone (e.g., shopping behavior, customer satisfaction, etc.); often, only people respond that have a strong (negative) opinion; therefore, this sample is usually biased, i.e., it does not correctly represent the entire population; a 10% reply quota is often a success for these samples

Undercoverage: (some) group(s) in the population are left out from the sampling process

Nonresponse: an individual chosen for the sample cannot be contacted or refuses to reply

Simple Random Sample (SRS): a sample of size n (i.e., a sample that consists of n different individuals from the population) that is chosen in such a way that every set of n individuals has the same chance to be the sample actually selected

Example: these are **NO** SRS:

- pick the first student from a class list at random, then take the next 10 names in alphabetical order
- pick the first student randomly, then take all students that have a lag of 3 in the internal numbering (e.g., if we pick # 7 first, we would also select # 10, # 13, # 16, etc.)

Probability Sample: gives each member of the population a known chance (> 0) to be selected; these chances may be different for each individuum; therefore, a SRS is a special case of a probability sample since in a SRS, each individuum has a chance of $1/N$ to be selected

Stratified Random Sample: first we divide the population into similar subsets, called **strata**; then we choose a separate SRS in each stratum and combine the individual results to form the full sample

Example:
in the US, the 50 (+1) states are natural strata; usually, characteristics of the population such as education level, health, income, employment, etc., are very similar in each state (i.e., stratum), but quite different among states

4.4 Toward Statistical Inference

Statistical Inference: based on data, usually a sample (or the outcome of an experiment), we want to draw conclusions about the entire population

Parameter: a number that describes the population; in practice, the numeric value of a parameter usually is not known

Statistic: a number that can be computed from the sample data (or from the outcome of an experiment) without making use of any unknown parameter; in practice, we use a statistic to estimate an unknown parameter

<u>Example:</u>

μ and σ are unknown parameters; $\overline{x}$ and s are statistics calculated from the sample data

<u>Sampling Distribution and Sampling Variability:</u>

Let us take a sample of n objects and look for a particular criterion, e.g., people that have blue eyes. We define

$$\hat{p} = \frac{\text{\# objects thet meet the criterion}}{n}.$$

The **sample proportion** $\hat{p}$ is a statistic. If we take a second sample, $\hat{p}$ most likely will be different. The fact that the value of a statistic varies in repeated random sampling is called **sampling variability**. The **sampling distribution** of a statistic is the distribution of values that can be taken by the statistic in all possible samples of the same size from the same population.

Example: "The Cost of Victories"

In our initial sample of $n = 8$ teams, we determined the average number of wins

$$\overline{x} = \frac{155}{8} = 19.375.$$

Now, each student in class should take another simple random sample (SRS) of size $n = 8$ teams and recalculate $\overline{x}$ based on this his/her new sample.

For the entire population (i.e., all teams), we know that

$$\begin{aligned} N &= 30 \\ \mu &= \frac{565}{30} = 18.83 \\ \sigma^2 &= 18.94 \\ \sigma &= 4.35 \end{aligned}$$

Here are the ... sample means for the samples of size 8 taken in class:

We round each of ... the sample means to 1 decimal digit and draw a stem–and–leaf plot (with split stems) of the data:

We calculate the mean $\overline{x}$ and the standard deviation s of the ... sample means:

$\overline{x} =$

$s =$

Note that a sample of size ... is relatively small. So it is somewhat justifyable that this data is approximately Normal distributed.

Recall that we know from the empirical rule that 68% (i.e., $0.68 \times ... = ...$) of all samples should fall into the interval $\overline{x} \pm 1 \cdot s = [\quad , \quad]$; in our case, of the sample means fall into this interval.

We also know from the empirical rule that 95% (i.e., $0.95 \times ... = ...$) of all samples should fall into the interval $\overline{x} \pm 2 \cdot s = [\quad , \quad]$; in our case, of the sample means fall into this interval.

Take a closer look at the stem–and–leaf plot. How does the distribution look like?

(i) Is it (almost) symmetric? — yes/no

(ii) Is it (approximately) Normal distributed? — yes/no

(iii) Are there any outliers? — yes/no

(iv) Is the center close to μ? — yes/no

Ideally, we should be able to answer questions (i), (ii), and (iv) with yes and question (iii) with no when the sample size is large enough. For relatively small sample sizes, this may not always be the case.

Unbiased Statistic:
a statistic that is used to estimate an unknown parameter is unbiased if the mean of its sampling distribution is equal to the true value of this (unknown) parameter; as an example $\overline{x}$ is an unbiased estimate of the population mean μ and $\hat{p}$ (obtained from a SRS) is an unbiased estimate of the population parameter p

Variability of a Statistic:
the variability of a statistic is described by the spread of its sampling distribution; this spread is determined by the sampling design and the sample size n; larger samples result in a smaller spread