Index of 2-L

Practical Information

WELCOME back — any new students today? ...

Overview of today's lecture:

- more on descriptive analysis with brief Minitab demonstration,
- presentation of ideas about statistical planning/ experimentation and epidemiological reasoning:
  - causation: confounding, bias,[1]
  - design of experiments: control, randomization and replication,[2]
  - random selection, random numbers,[3]
  - surveys: sampling, stratification,[4]
- in addition, some exercises to complete "together",
- brief overview of ethics related to data collection,
- references to skipped chapters (e.g., the correlation r): you may skip over for now, we'll return to those sections later.

Next lab session: Monday $1-4$pm in 218S, our standard time.

[1] PSLS 3e/4e: brief discussion (Chapter 7/6); IPS7e: Section 2.6.
[2] PSLS 3e/4e: Chapter 8/7; IPS 7e: Section 3.1.
[3] PSLS 3e/4e: Chapter 7/6; IPS 7e: Section 3.1.
[4] PSLS 3e/4e: Chapter 7/6; S: Section 1.2; IPS 7e: Section 3.2.

More about Descriptive Statistics and Outliers

Determining shape for distributions of quantitative (continuous) variables:

- graphically explore shape by stemplot, dotplot and/or histogram (relevant distribution curves, e.g. normal, may be overlaid),
- further assess symmetry by descriptive measures (median $\approx$ mean, Q1 and Q3 symmetrical around median),
- further assess shape by computing
  - skewness: $<0, =0, >0 \sim$ left-skewed, symmetrical, right-skewed, respectively,
  - kurtosis: $=0 \sim$ normal, $>0 \sim$ (often) heavy tails (for data: outliers), resp.,[5]
- beware that distribution shapes may appear irregular in small samples (say, $n \leq 15$).

Outlier = observation that does not belong with the others, typically by being extreme,

- visual assessments from stemplot, dotplot or histogram,
- subject-matter knowledge may deem value implausible or an outright error,
- "suspected outlier" rule based on 5-number summary:
  - screening tool for (extreme) values that may be worth inspecting,
  - cannot be expected to correctly identify real outliers (see 1L−14).

[5] For skewed distributions, kurtosis is of no interest! In Stata, kurtosis values are 3 units larger: kurtosis=3 for normal. Controversy exists about (in)correct interpretations of kurtosis (e.g., Westfall (2014), *Amer. Statist.* 68, 191−195).

Association and Causation

Generally in statistics, when talking about interpretations and relationships, these always refer to a population, ideally the population the data are thought to represent.

Association between two variables x and y = a certain pattern in the combined distribution of the two, e.g. explored by a scatterplot for two quantitative variables:

- positive association: high (low) values of x and y appear together,
- negative association: high (low) values of x together with low (high) values of y.

Causation = direct link between variables whereby one (say x) causes the other (y).

Fundamental caution for interpretations: association does not always imply causation.

Example 7.2 of PSLS 3e (6.2 of PSLS 4e): alcohol type (wine vs. beer & spirits) and health in UNC (University of North Carolina) Alumni Heart Study,[6]

- apparent health benefits of wine (compared with other alcohol) found, but ...
- "may be due to confounding by dietary habits and other lifestyle factors".

Definition (IPS, PSLS): two variables are confounded when their effects on a response variable cannot be distinguished from each other.

[6] Barefoot et al. (2002), *Amer. J. Clin. Nutr.* 76, 466−72.

Experimental and Observational Data

Experiment versus Observational study:

- an experiment deliberately imposes some treatments on individuals in order to observe their responses,
- an observational study observes individuals and variables of interest, but does not attempt to influence responses; its results/interpretations may be subject to bias:
  - systematic favour of certain outcomes (IPS/PSLS) ⇒ potentially false conclusions,

Ideal method of establishing causation: experiments, because they allow comparisons all other things being equal, however often (depending on research field) not feasible:

- unethical to carry out experiments (e.g., on humans),
- impractical (due to cost or logistics).

Guidelines exist for establishing causation from association without experiments, e.g.:[7]

- strong and consistent association (several data sources point in the same direction,
- gradual association (stronger exposure ⇒ stronger response, dose-response),
- time consistency (exposure before response; changes in exposure ⇒ subsequent changes in response),
- plausible cause (e.g., established in similar setting, such as another species).

[7] Discussed in Epidemiology; a brief summary is included in IPS7e, Section 2.6.

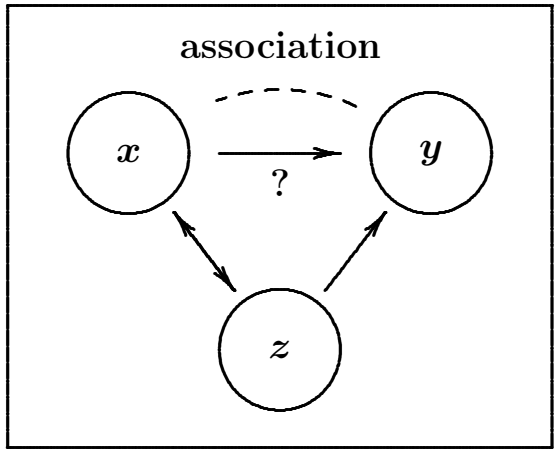
Exercise 3.6

National Halothane Study[8]: a U.S. epidemiological study from the late 1960s to evaluate halothane[9] toxicity.

(a) The anesthetic used was not imposed, but rather chosen by the doctors caring for each patient.

(b) The higher death rate for anesthetic C could be due to confounding, possibly unobserved (IPS, PSLS: lurking), variables z; some possibilities: nature and seriousness of condition/surgery, the patient's overall physical condition and age.

Schematic for potential confounding scenario by variable z:



x = type of anesthetic

y = mortality

[8] Bunker JP, Forrest WH, Mosteller F, et al (Eds). National Halothane Study. A study of the possible association between halothane anesthesia and postoperative hepatic necrosis. U.S. Government, Washington, DC, 1969.

[9] Halothane is an inhaled anesthetic, introduced in the 1950s and in the early years suspected to be associated with increased risk of hepatitis (liver disease).

Planning a Study

Data sources:

- anecdotal evidence: rarely useful, unrepresentative of population,
- available data, often in registers:
  - produced/collected for other purposes $\Rightarrow$ quality and usefulness not evident!
- designed sample surveys (samples): address a subset of entire population, as opposed to censuses,
- designed experiments may use data generated exclusively for study or routinely recorded data (e.g. use of register data for clinical field trials).

Statistical design = procedures for collecting data:

(1) Which individuals to be studied? and how many?[10]

(2) What variables to record?

(3) What patterns to explore and hypotheses to test?

(4) Method af analysis.

Many statistical designs include only (1)−(2); however the statistical analysis and any statistical assessment of necessary sample size (to be discussed later in the course) benefit from or require information about (3)−(4).

[10] Recall that "individuals" (or observational/experimental units, next slide) can also be samples.

Experimental Design

2 examples of statistical designs (both completely randomized; 2L−9):

Parasite exposure in Lithuania		
Calves	Pasture	
	safe	infected
	10	10

Advertising study (Exercise 13.22)			
Subjects	Repetitions (times)		
Familiarity	1	2	3
familiar	15	15	15
unfamiliar	15	15	15

Some terminology for experimental design:

- treatment: specific experimental condition applied by the experimenter,
- experimental units: subjects/individuals/samples to which treatments are applied,[11]
- factor: (controlled) explanatory variable in experiment,
- level: specific value of factor/treatment.

Three basic principles of experimental design:

(1) control of lurking variables, by control/placebo[12] group and/or blinding[13],

(2) randomization: random assignment of units to treatments (hence sometimes called a "randomized comparative experiment"; IPS, PSLS),

(3) replication: sufficient number of units to "drown out" randomness.

[11] Not necessarily the same as the measurement units on which measurements are taken.

[12] Placebo: control treatment that is "fake" but otherwise indistinguishable from real treatment; placebo effect: apparent positive effect of placebo treatment.

[13] Blinding: subject and/or experimenter are not aware of the identity of treatment groups (both → double-blind).

Random Selection

In practice: how to select individuals for a treatment group?
— e.g., 10 (m) calves for safe pasture out of 20 (n).

First number individuals 1,. . . ,20 (n). Then, employ one of the following methods,

- use one's own random generation, e.g. draw 10 (m) cards from a pile of 20 (n),
- using random digits (Table A in PSLS; Table B in IPS):
  - choose starting point in table (arbitrarily),
  - read off digits (along rows) to form numbers $1-20$ (n), until 10 (m) different numbers encountered,
- using Minitab:
  - generate column with numbers 1,. . . ,20 (n) (`Calc-Make Patterned Data-Simple...`),
    - (easy way): sample from the column without replacement[14],
    - (flexible way): generate a column of the same length with random numbers[14], and sort both columns by the random numbers (`Data-Sort`),
  - use the first 10 (m) numbers in new or sorted column,
  - procedure is reproducible using a seed (Minitab terminology: "base"),
- using Simple Random Sample applet (not reproducible).

[14] **Minitab menus:** `Calc-Random Data-Sample from Columns` **and** `Calc-Random Data-Uniform`, **respectively.**

Completely Randomized Design

In a completely randomized design (or trial), all treatments are allocated at random among the experimental units (using a method for random selection). In all other respects, the units are treated as equally as possible.

$\Rightarrow$ (idea/rationale:)

Differences in response must be due to either treatment effects or play of chance in random assignment of units.

Comments:

+ simple/easy to understand, carry out and analyze,
+ flexible (allows any number of levels and replications),
+ randomization as safeguard against systematic errors (bias), e.g. by randomly re-ordering the experimental units (easy in statistical software),
− the experimental units need to be "homogeneous", otherwise the random variation will be large,
− if a good grouping of experimental units exists (either in their state before treatments are applied or in general conditions during the experiment), other designs will be more efficient (give better precision),
∗ primarily for small designs with no obvious grouping (as described above) that could be used as blocks ($\rightarrow$ next slide).

Block Design

Blocks = groups of homogeneous experimental units, i.e., units are more alike within than between groups, before and during experiment.

In a (randomized) block design, treatments are assigned randomly to the units within each block, typically such that each treatment occurs once in each block[15] ⇒ idea/rationale:

> it should be more accurate to compare similar units (within same block), and to aggregate such comparisons across blocks.

Some special cases:

- matched pairs design: blocks of size 2,
- cross-over design: multiple treatments on same subject (= block) in successive periods.

Examples of factors used to form blocks:

- animal science: litters, groups (age/weight/sex), environment (herd),
- human medicine: twins, family, groups (as above + condition, social status, lifestyle...),
- agriculture: areas in fields; general: time, operator (surgeon, technician).

Comments on block designs:

\+ improvement in precision if efficient groups available,

+/− minor added complexity in design and analysis,

− less flexible (block size should match number of treatments),

∗ very useful and very much used.

[15] Not a strict requirement, and even incomplete blocks are possible; → VHM 802 for discussion of block designs.

Observational Study Types and Terminology

Brief overview of observational study types:[16]

- cross-sectional study: based on survey at a single point in time,
- cohort study: study groups (e.g. defined by exposure) followed over time,
- case-control study: cases and control selected separately ($\Rightarrow$ each individual's status as a case or control is predetermined), and their characteristics are compared,
  - main advantage when cases (or events) are rare,[17]
  - needs special analytical methods to get proper inference about the population,
- controlled study: an experiment carried out on subjects in their usual environment (and possibly utilizing routinely recorded data),
- retrospective study: using past data (often as a case-control study), contrasting a prospective study (often a cohort study).

Concepts and terminology from epidemiology:

- Studies do not always use individuals (subjects, samples) that are drawn directly from the population of interest (PoI); this raises the issue of whether the study findings are representative for the PoI (in epidemiology, termed external validity),
- Different bias types distinguished, e.g. selection, information, and confounding bias.

[16] Detailed coverage can be found in courses in Epidemiology.
[17] If cases are rare, it may be impractical to get enough cases by sampling randomly from the population.

Data Ethics

Aim: brief overview of concepts and questions
— more detailed coverage should be available for your program.[18]

4 major principles of research involving humans:

* An institutional review board must review all planned studies in advance in order to protect the study subjects.[19]
* All individuals must give their informed content in writing before data are collected.
* All individual data must be kept confidential, meaning for example that only summaries for groups of subjects may be made public.
* A guiding principle for ethics questions is: "The interests of the subject must always prevail over the interests of science and society".

Further questions involving ethics:

- Is it ethical to include a placebo group when safe and efficient treatments exist?
- Must a trial be stopped or altered if early results indicate adverse effects or clear superiority of one group?[20]

[18] In the past, ethics was part of the mandatory course VBS 803 (Principles of Biomedical Research); as this course is undergoing changes, ethics may be covered in a workshop.
[19] UPEI has separate bodies for human and animal research: Research Ethics Board (human participants), and Animal Care Committee (animal participants).
[20] Research in biostatistics deals with both stopping rules and methodology for analysis of altered trials.

Surveys and Sampling

In surveys, we select a subset of individuals from a population to draw inference about the entire population,

- population: all the individuals of interest,
- sample: subset from population,
- sample design: method to extract sample; major distinction: random/probability sampling vs. non-probability sampling (e.g., convenience and purposive samples[21]),
- common example: opinion polls.

Main reasons to prefer survey over census ($\sim$ entire population) are costs and feasibility.

Some common causes of selection bias by favouring certain parts of the population in the sampling:

- voluntary response sample (respondents rarely representative),
- non-random selection (often a convenience sample: the individuals/samples at hand),
- undercoverage (some parts of population left out of sampling process),
- non-response (non-response may be more likely for certain parts of population).

Response bias = answers incorrect due to "circumstances",

- particular example: wording of questions.

[21] A convenience sample is chosen because easy to obtain, a purposive sample in order to "fit" to the targeted population.

(Random) Sampling Schemes

Simple random sampling (often SRS):

- choose n individuals from population such that every subset of size n has same chance of selection,[22]
- \+ simplest to understand and analyze,
- − impractical if entire population cannot be enumerated.

Systematic random sampling:

- assume population (units) ordered (say $1, \ldots, N$), and choose a sampling interval I, typically to achieve a desired sample size $n = N/I$,
- select the first sample randomly among units $1, \ldots, I$, and thereafter select every Ith unit,
- a (logistically) simple probabilistic sampling method, but not SRS $\Rightarrow$ biases may occur.

Stratified sampling:

- split population into homogeneous groups (strata, e.g. geographical), then use SRS (or other approaches) within each group,[23]
- similar to a block design (strata $\sim$ blocks), and with similar advantages and disadvantages.

Multistage sampling: (including cluster sampling[23])

- sampling in several stages, often corresponding to population's hierarchical structure; for example, sampling of cows in two stages — first herds, then cows within selected herds,
- practical and economical advantages (but more complex analysis).

[22] The same principle as with random selection in completely randomized designs.

[23] The strata need not be represented equally in data which would then be accounted for in analysis by a weighting procedure; note that all strata need to represented, as opposed to cluster sampling with only a subset of clusters included.

Exercise 3.52

Word lengths in writings of Tom Clancy.

Answers:

- population: words in Tom Clancy novels,
- sample: 250 words on one page in one novel,
- subject/experimental unit: each word,
- variable measured: length (number of letters in word).

Do you think the sample is representative for the population?

Answer:

* maybe need more pages in same book,
* certainly need more books,
* all in all too small.

Summary Notes

2 aims of statistical methods:

- detect pattern(s) in a data set, without prior knowledge about which patterns the analysis will focus on ⇒ exploratory data analysis,
- confirm or disprove certain theories (hypotheses) about relationships in the population (ideally population of interest) the data are thought to represent
  ⇒ formal statistical inference.

Any generalization from specific (sample) to common (population) relies on assumptions! (e.g., representativity and ability to avoid/control bias).

Key words and concepts:

- descriptive statistics to quantify distribution shape,
- causation, confounding/lurking variable, bias,
- experimental design terminology, control, randomization (including methods for), replication,
- completely randomized design, block design,
- survey and sampling terminology, simple random sample, stratification,
- individual, experimental unit, population (of interest).