

Index of 2-L

Page	Title
1	Practical information
2	More about descriptive stats and outliers
3	Association and causation
4	Experimental and observational data
5	Exercise 3.6
6	Planning a study
7	Experimental design
8	Random selection
9	Completely randomized design
10	Block design
11	Surveys and sampling
12	Sampling schemes
13	Exercise 3.52
14	Study and experimental design terminology
15	Summary notes

PRACTICAL INFORMATION

WELCOME to all newly arrived students...

Overview of today's lecture:

- more on descriptive analysis (1L–20, 2L–2), with further Minitab demonstrations,
- summary worksheet (from Stephens book; Chapter 3) for descriptive statistics,
- 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,
- no discussion of ethics in VHM 801 (\Rightarrow VBS 803),
- references to skipped chapters (e.g., the correlation r): skip over for now, we'll return to those sections later.

Next lab session: on Monday (Sep 9) 1-4pm; it will include time to catch up from the first lab (if needed).

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 STATS AND OUTLIERS

Determining shape for distributions of continuous variables:

- graphically explore shape by stemplot and/or histogram (relevant distrib. 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-skew, symm., right-skew,
 - * kurtosis: $= 0 \sim$ normal, $> 0 \sim$ (often) heavy tails (in data distributions: outliers),⁵
- beware that shape may appear irregular for 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,
- “suspected outlier” rule based on 5-number summary:
 - * screening tool for (extreme) values worth inspecting,
 - * relies on assumptions of symmetry and “moderate” data size in order to correctly identify real outliers.

⁵ Note: *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

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 appear 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: alcohol type (wine vs. beer) and health in UNC (Univ. North Carolina) Alumni Heart Study,⁶

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

Definition: 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 no attempt to influence responses.

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/criteria exist for establishing causation from association without experiments, e.g.:⁷

- strong association,
- consistent association (several data sources point in the same direction),
- gradual association (stronger exposure \Rightarrow stronger response), often as a dose-response relation,
- time consistency (exposure before response; changes in exposure \Rightarrow subsequent changes in response),
- plausible cause (e.g., established in similar setting).

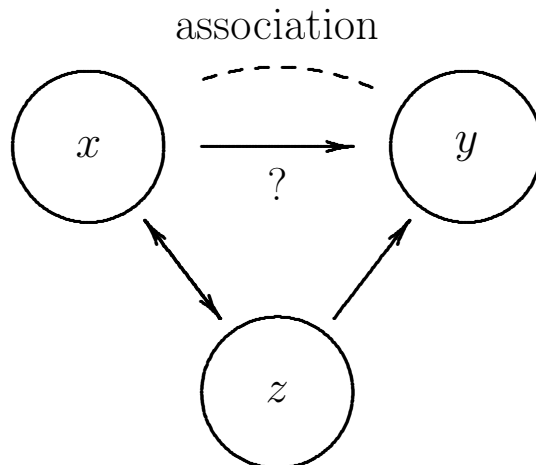
⁷ Discussed in Epidemiology; a brief summary is included in IPS7e, Section 2.6.

EXERCISE 3.6

National Halothane Study⁸: a U.S. epidemiological study from the late 1960s to evaluate halothane⁹ 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 *lurking*/unobserved) variables; 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 Printing Office, 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,
- available data, often in registers:
 - * produced/collected for other purposes,
 - * 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 of 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 course) benefit from or require information about (3)–(4).

¹⁰ Recall that observational/experimental units can also be samples.

EXPERIMENTAL DESIGN

2 examples of statistical designs (completely randomized):

Parasite exposure in Lithuania			Advertising study (Exercise 13.22)			
Calves	Pasture		Repetitions (times)			
	safe	infected	Familiarity	1	2	3
	10	10	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.

3 basic principles of experimental design:

- (1) control of lurking variables, by control/placebo group,
- (2) randomization: random assignment of units to treatments to avoid unexpected patterns,
- (3) replication: sufficient number of units to “drown out” randomness, or repetitions of experiment.

Violation of (1) or (2) may result in bias = systematic favour of certain outcomes (\Rightarrow potentially false conclusions).

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

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 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 column without replacement (Calc-Random Data-Sample from Columns),
 - *other way*: generate column of same length with random numbers (Calc-Random Data-Uniform), and sort both columns by the random numbers (Data-Sort),
 - * use first 10 (m) numbers in new or sorted column,
 - * procedure is reproducible using a *seed* (Minitab: base),
- using Simple Random Sample applet (not reproducible).

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.

⇒ (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 reordering the experimental units (easy in statistical software),
- all 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 → 2L–10.

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 (*incomplete blocks* also possible \Rightarrow VHM 802)

\Rightarrow (idea/rationale:)

More accurate to compare similar units (within same block). and to aggregate such comparisons across blocks.

Special cases:

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

Examples of factors used to form blocks:

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

SURVEYS AND SAMPLING

In surveys, we select a subset of individuals from a population to draw inference *about populations*,

- population: entire group of individuals of interest,
- sample: subset from population,
- sample design: method to extract sample,
- common example: opinion polls.

The main reasons for preferring survey to census (for entire population) are costs and feasibility.

Some common causes of bias (systematic errors) by favouring certain parts of the population:

- voluntary response sample (respondents rarely representative),
- in general, 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.

SAMPLING SCHEMES

Simple random sampling:

- choose n individuals from population such that every subset of n individuals has equal chance of selection,¹²
- + simplest to understand and analyze,
- impractical if entire population cannot be enumerated, so often statistically and practically inefficient.

Systematic random sampling:

- assume samples ordered (say $1, \dots, N$), and choose a *sampling interval* I , typically to achieve a desired sample size $n = N/I$,
- select the first sample *randomly* among samples $1, \dots, I$, and thereafter select every I th sample,
- often the logistically simplest probabilistic sampling method, but not a simple random sample \Rightarrow biases may occur.

Stratified sampling:

- split population into homogeneous groups (strata, e.g. geographical), use simple random sampling in each group,¹³
- similar to a block design (strata \sim blocks), and with similar advantages and disadvantages.

Multistage sampling: (including cluster sampling)

- 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).

¹² The same principle as with random selection in completely randomized designs.

¹³ Note: The strata need not be represented equally in data; this would then be accounted for in analysis by a weighting procedure.

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,
- variable measured: length (number of letters).

Do you think the sample is representative for the population?

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

STUDY AND EXPERIMENTAL DESIGN TERMINOLOGY

- 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).
- Observational study types (in epidemiology):
 - * cross-sectional study: based on survey at single point in time,
 - * cohort study: study groups (e.g. exposure) followed over time,
 - * case-control study: cases and control selected separately, and their characteristics are compared,
 - * retrospective study: using past data (often as a case-control study), contrasting a prospective study.
- Randomized comparative designs involve several (≥ 2) treatment groups and random allocation of subjects to treatments (\Rightarrow includes completely randomized designs).
- Placebo: control treatment that is “fake” but otherwise indistinguishable from real treatment; placebo effect: apparent positive effect of placebo treatment.
- Blinding: subject and/or experimenter are not aware of the identity of treatment groups (both \rightarrow double-blind).

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 data, typically with the aim of generalizing the conclusions to a more general context,
⇒ formal statistical inference.

Any generalization from specific (sample) to common (population) relies on assumptions! (e.g., representativity).

Key words and concepts:

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