

## Index of 2-L

Page	Title
1	Practical information
2	More about descriptive statistics 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	Observational study types and terminology
12	Data ethics
13	Surveys and sampling
14	Random sampling schemes
15	Exercise 3.52
16	Summary notes

## 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,<sup>1</sup>
  - \* **design of experiments**: control, randomization and replication,<sup>2</sup>
  - \* **random selection**, random numbers,<sup>3</sup>
  - \* **surveys**: sampling, stratification,<sup>4</sup>
- 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 1–4pm in 218S, our standard time throughout the semester.

---

<sup>1</sup> PSLS 3e/4e: brief discussion (Chapter 7/6); IPS7e: Section 2.6.; comprehensive discussions in epidemiology.

<sup>2</sup> PSLS 3e/4e: Chapter 8/7; IPS 7e: Section 3.1.

<sup>3</sup> PSLS 3e/4e: Chapter 7/6; IPS 7e: Section 3.1.

<sup>4</sup> 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.,<sup>5</sup>
- 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).

---

<sup>5</sup> 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,<sup>6</sup>

- 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.<sup>7</sup>

<sup>6</sup> Barefoot et al. (2002), *Amer. J. Clin. Nutr.* 76, 466–72.

<sup>7</sup> This is confounding in **statistical** sense, not in epidemiological sense.

## 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)  $\Rightarrow$  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.:<sup>8</sup>

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

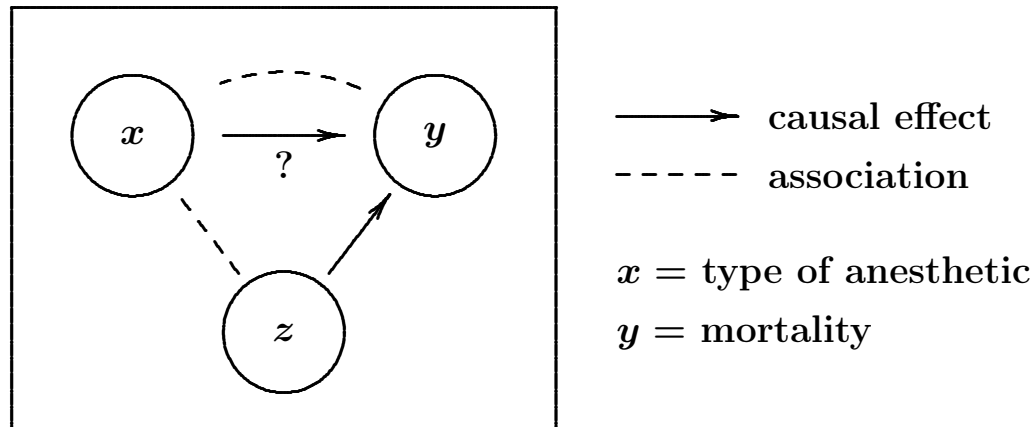
<sup>8</sup> Discussed in epidemiology; a brief summary is included in IPS7e, Section 2.6.

### EXERCISE 3.6

**National Halothane Study**<sup>9</sup>: a U.S. epidemiological study from the late 1960s to evaluate halothane<sup>10</sup> toxicity.

- The anesthetic used was not imposed, but rather chosen by the doctors caring for each patient.
- 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  $z$ :



<sup>9</sup> 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.

<sup>10</sup> 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?<sup>11</sup>
- (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 the course) benefit from or require information about (3)–(4) as well.

<sup>11</sup> Recall that “individuals” (or observational/experimental units, next slide) can also be samples.

## EXPERIMENTAL DESIGN

2 examples of statistical designs (both **completely randomized**  $\rightarrow 2L - 9$ ):

Parasite exposure in Lithuania			Advertising study (Exercise 13.22)			
Calves	Pasture		Subjects	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,<sup>12</sup>
- **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**<sup>13</sup> group and/or **blinding**<sup>14</sup>,
- (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.

<sup>12</sup> Not necessarily the same as the **measurement units** on which measurements are taken.

<sup>13</sup> **Placebo**: control treatment that is “fake” but otherwise indistinguishable from real treatment; **placebo effect**: apparent positive effect of placebo treatment.

<sup>14</sup> **Blinding**: subject and/or experimenter are not aware of the identity of treatment groups (both  $\rightarrow$  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<sup>15</sup>,
    - (**flexible way**): generate a column of the same length with **random numbers**<sup>15</sup>, 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).

<sup>15</sup> 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**.

⇒ (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 (→ 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<sup>16</sup> ⇒ **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.

---

<sup>16</sup> 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**:<sup>17</sup>

- **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**,<sup>18</sup>
  - \* 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.

---

<sup>17</sup> Detailed coverage can be found in courses in epidemiology.

<sup>18</sup> 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 involved in data collection and analysis.

**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 — at UPEI:
  - \* **Research Ethics Board** (human participants),
  - \* **Animal Care Committee** (animal participants),
- \* All individuals must give their **informed consent** 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?<sup>19</sup>

---

<sup>19</sup> 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-random/non-probability** sampling,
- **common example**: opinion polls.

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

Some common causes of **selection bias**<sup>20</sup> by favouring certain parts of the population in the sampling:

- **voluntary response sample** (respondents rarely representative),
- **non-random selection** (e.g., a **convenience** or **purposive** sample<sup>21</sup>),
- **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**<sup>20</sup> = answers incorrect due to “circumstances”,

- particular **example**: wording of questions.

---

<sup>20</sup> Detailed discussions in epidemiology.

<sup>21</sup> 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,<sup>22</sup>
- + **simplest** to understand and analyze,
- **impractical** if entire population cannot be enumerated.

### Systematic random sampling:

- assume population (units) 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 units  $1, \dots, I$ , and thereafter select every  $I$ th 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,<sup>23</sup>
- **similar** to a block design (**strata  $\sim$  blocks**), and with similar advantages and disadvantages.

### Multistage sampling: (including cluster sampling<sup>23</sup>)

- 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).

---

<sup>22</sup> The same principle as with random selection in completely randomized designs.

<sup>23</sup> 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 be 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  $\Rightarrow$  **exploratory data analysis**,
- **confirm or disprove certain theories** (hypotheses) about relationships in the population (ideally population of interest) the data are thought to represent  $\Rightarrow$  **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).