7. Data Collection

7.1. Introduction

The set of all statistical units of interest is called population (populaatio/ perusjoukko). A subset of a population is called a sample (näyte). It is called a random sample (satunnaisotos), if every element of the population has a positive probability of being selected.

In a census (kokonaistutkimus) we collect information about all elements of the population, whereas in a sample survey (otantatutkimus) we collect information only from the elements of the sample.

Much statistical data is available from Statistics Finland (Tilastokeskus) at www.stat.fi. If ready-made data is not available, you must produce it yourself, either in an observational study (havaintotutkimus) which measures variables without attempting to influence their values, or in an experiment (koe) which deliberately imposes some treatment on the objects in order to observe their responses.

7.2. Design of Experiments

Some Terms

The objects on which the experiment is done are the experimental units (koeyksiköt). A specific experimental condition applied to the units is called a treatment (käsittely). The explanatory variables in an experiment are called factors (tekijät), the values of which are called factor levels (tekijän taso). Then a treatement is defined by the levels of each of the factors. The dependent variables in an experiment are called response variables (vastemuuttujat).

Example. A television program is interrupted either 1, 3 or 5 times with a commercial for a certain product of either 30 or 90 seconds. Then there are two factors: length of the commercial with 2 levels, and repetition with 3 levels. The 6 combinations of one level of each of the two factors form 6 treatments. A potential response variable would be the inclination of viewers of the TV programme to buy the product.

Principles of Experimental Design (koesuunitelman periaatet)

After deciding on the factors, the response variables and the layout of the treatment, one has to ensure that no lurking variables outside the control of the experimenter influence the outcome of the experiment. The design of a study is called <u>biased</u> (harhainen) if it systematically favours certain outcomes. In order to avoid such a <u>bias</u> (harha) in simple experiments with only one treatment, one introduces a so called <u>control group</u> (vertailu-/ kontrolliryhmä), which is not exposed to the treatment in question.

<u>Example.</u> Medical experiments without control group tend to report curative effects even for medication which is known to have no effect. This is the so called <u>placebo effect</u> (lumevaikutus). Introducing a control group exposed to placebo treatment alone makes sure that any measured response to medication is not due to the placebo effect.

Comparison of the effects of several treatments is only valid when all treatments are applied to similiar groups of experimental units. Therefore the experimental units must be assigned to the treatments in a way that does not depend on any of their characteristics nor on the judgement of the experimenter in any way. Doing this by impersonal chance is called randomization (satunnaistaminen).

Example. (TV commercial continued)
If 60 people are available for testing the commercial we may randomly assign them to the six treatments e.g. by letting them tossing a dice. If we wish groups of equal size they may toss again if their group is already full.

In practice, randomization may be either done with a table of random digits (satunnaisluvuntaulukko) or a random number generator (satunnaislukugeneraattori).

A table of random digits is a list of the digits $0, 1, 2, \ldots, 9$ with the following properties:

1. The digit in any position has the same chance of being any one of $0, 1, 2, \ldots, 9$.
2. The digits in different positions are independent in the sense that the value of any one has no influence on the value of any other.

This implies:

Any pair has the same chance of being any of the 100 possible pairs: $00, 01, \ldots, 98, 99$; any triple has the same chance of being any of the 1000 possible triples: $000, 001, \ldots, 999$; ... and so on for groups of four or more random digits.

A random number generator is a computer algorithm that generates a random number usually between 0 and 1 (implemented in Excel as the function RAND()). You can get an integer random number x with the property $a \leq x \leq b$ by calling RANDBEWTEEN(a,b) in Excel.

Example. (TV commercial continued)

In order to assign people randomly to the treatments with a table of random digits, enumerate each person, and assign it to the random digit at his position, disregarding the digits $0, 7, 8, 9$ and those with a group that is already full.

In order to assign people randomly to the treatments with the random number generator in Excel, generate for each of them a number x in the range $1 \leq x \leq 6$ with the command RANDBETWEEN(1,6) (drawing again for groups that are already full).

Note. We may avoid extra drawings for full groups by assigning remaining persons to treatments rather than treatments to persons.

7.3. Sampling Methods *(Otantamenetelmät)*

Simple Random Sampling (SRS)
(Yksinkertainen satunnaisotanta)

Simple Random Sampling (SRS) should be used when there is no prior information available about the structure of the population. In SRS, every statistical unit has the same probability to be included into the sample, which implies that also every sample of the same size has the same chance of being drawn.

To choose a simple random sample of size n out of a population of N elements, enumerate the elements of the population from 1 to N, generate n random number within the same range (=RANDBETWEEN(1,N)), and select those statistical units with their order numbers chosen by the random number generator into the sample.

If a table of random digits is to be used, the elements of the population must be numbered from 0 to $N-1$. Then one considers numbers made up of the same number of digits k as N.

Example.
Suppose we want to select a simple random sample of size $n = 5$ from a population of $N = 600$ elements using a random digit table starting with the digits

59426 45792 78799 15803.

After enumerating the elements of the population from 0 to 599, we select the elements with the numbers 594, 264, 579, 278, and 158. (The number 799 is discarded since it is larger than 599.)

Note. SRS need not necessarily yield a representative sample if the population is made up of very heterogeneous groups.

Systematic Sampling (Systemaattinen otanta)

Systematic Sampling is applicable when the population cannot be precisely determined (e.g. customer research) or when the elements of the population are arranged in some order (e.g. existing register). In systematic sampling one unit is chosen from the whole population in regular distances of length $k = N/n$, where N denotes the size of the population and n the size of the sample. The first sample unit is chosen randomly from the first k elements of the population, and after that every k'th element is selected.

Note. Systematic Sampling is inappropriate when the statistical units have been ordered according to their values, or the distance k in which units are selected coincides with some systematic periodicity in the data.

Cluster Sampling (ryväsotanta)

In cluster sampling the population may be divided into groups or clusters based upon some characteristics, such as county, education institution, etc.

Cluster Sampling consists of selecting m out of M clusters, and then either using all elements within the m clusters as the sample (single-stage cluster sampling) or further selecting a random sample of n elements out of the m clusters selected (two-stage cluster sampling).

Unlike stratified sampling (ositetussa otanta) to be discussed below, cluster sampling does not assume that the units within the same cluster have more similiar characteristics regarding the topic under investigation then those between different clusters.

In Stratified Sampling (ositetussa otanta) one tries to exploit some available background information about the properties of the population. If the population may be split up into heterogeneous groups in such a way that that elements within the same group are similar, but different between different groups, then stratified sampling should be applied. Each group is called stratum (osite) and one takes seperate random samples from each group, that is, one first decides upon L strata $O_1, O_2, \ldots, O_L$ with respective numbers of observations $N_1, \ldots, N_L$ and then takes random samples of each of them using either SRS or systematic sampling.

The splitup of the population into strata is called Allocation (kiintiöinti). In a uniform allocation (tasainen kiintiöinti) the sample gets equally many elements from each stratum. In proportional allocation (suhteellinen kiintiöinti) the number of elements from each stratum is proportional to its size. The allocation is called optimal (optimaalinen) if it results from minimzing e.g. the costs of the survey.

7.4. Sampling Distributions (otantajakaumat)

Every function of sample observations alone is called a (sample) statistic (tunnusluku). The sample mean, variance, and standard deviation are examples of (sample) statistics. Numerical measures of the population are called parameters (parametrit).

Now selecting a sample is a random event, which implies that sample statistics are random variables with associated probability distributions. The probability distribution of a statistic is called a sampling distribution (otantajakauma). Sampling distributions will soon be valuable to us when we want to assess whether we can reasonably expect our sample (the only thing we do observe) to have been taken from a population with a hypothesized prespecified distribution (statistical inference/ tilastollinen päättely).

The Sampling Distribution of the Mean (Otoskeskiarvon otosjakauma)

Let $X_1, \ldots, X_n$ be a random sample of a distribution with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Then:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\overbrace{E(X_i)}^{\mu} \\ &= \frac{1}{n}\cdot(n\cdot\mu) = \mu \end{aligned}$$

and

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\overbrace{V(X_i)}^{\sigma^2} \\ &= \frac{1}{n^2}\cdot(n\cdot\sigma^2) = \frac{\sigma^2}{n}. \end{aligned}$$

Furthermore, recall from our discussion of sums of normally distributed variables that

$$\begin{aligned} &\bar{X} \sim N(\mu, \sigma^2/n) \quad \text{for } X_i \sim N(\mu, \sigma^2), \text{ and} \\ &\bar{X} \overset{as.}{\sim} N(\mu, \sigma^2/n) \quad \text{otherwise,} \end{aligned}$$

since by the central limit theorem

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \overset{as.}{\sim} N(0, 1).$$

The Standardized Sampling Distribution of $\bar{X}$ when σ is not known: <u>The t-distribution</u>

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables (denoted as $X_i \sim$ *NID*(μ, σ^2) "Normally and Independently Distributed") and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, of which we now wish to find the standardized sampling distribution also for the case that σ is not known. We then use the sample standard deviation,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

in the place of σ. The standardized sampling distribution of $\bar{X}$ is now no longer normal, but

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1),$$

where $t(n–1)$ denotes the <u>Student t-distribution</u> with degrees of freedom (df) $n-1$.

The Student t-distributions are symmetrical around the origin and similiar in shape to the normal distribution, but have wider tails. They approach N(0,1) as $n \to \infty$.

Table. Tail fractiles (t_α) of the t-distribution: $P(T > t_\alpha(df)) = \alpha$.

	t_α					
df	0.100	0.050	0.025	0.010	0.005	0.001
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
40	1.303	1.684	2.021	2.423	2.704	3.307
50	1.299	1.676	2.009	2.403	2.678	3.261
60	1.296	1.671	2.000	2.390	2.660	3.232
70	1.294	1.667	1.994	2.381	2.648	3.211
80	1.292	1.664	1.990	2.374	2.639	3.195
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174

Example. $T \sim t(10) \Rightarrow$ 0.025 $= P(T \geq$ 2.228).

Implementation in Excel:

$P(T > t_\alpha(df)) = \alpha \;\Rightarrow\; t_\alpha(df)$=TINV($2 * \alpha; df$), Note.

The symmetry of the Student t-distribution implies that

$$P(T > t_\alpha(df)) = \alpha \;\Leftrightarrow\; P(|T| > t_\alpha(df)) = 2\alpha.$$

Now renaming α as $\alpha/2$ we obtain

$$P(|T| > t_{\alpha/2}(df)) = \alpha.$$

That is, we may also use the preceeding table in order to find the critical value $t_{\alpha/2}$ such that $P(|T| > t_{\alpha/2}) = \alpha$ simply by looking up in the column corresponding to $\alpha/2$.

Example.

$T \sim t(10) \Rightarrow\; 0.05 = P(|T| \geq 2.228)$.

Implementation in Excel:

$P(|T| > t_\alpha(df)) = \alpha \;\Rightarrow\; t_\alpha(df)$=TINV$(\alpha; df)$.

The Sampling Distribution of the Sample Proportion (Prosenttisen osuuden otosjakauma)

Let Π denote the percentage of type A elements in a random sample with elements $X_1, \ldots, X_n$. Π may then be estimated from:

$$\widehat{P} = \frac{100}{n}\sum_{i=1}^{n} I_i, \qquad I_i = \begin{cases} 1 & \text{for } X_i \text{ of type A} \\ 0 & \text{otherwise.} \end{cases}$$

Now $I_i \sim$Ber(p) with $p = \Pi/100$, such that $\sum_{i=1}^{n} I_i \sim$Bin(n, p) with expected value np and variance $np(1-p)$, such that by the central limit theorem for large n and $\Pi \approx 50$:

$$\sum_{i=1}^{n} I_i \overset{as.}{\sim} N(np, npq)$$

$$\Rightarrow\ \widehat{P} = \frac{100}{n}\sum_{i=1}^{n} I_i \overset{as.}{\sim} N\left(100p, \frac{100p \cdot 100(1-p)}{n}\right)$$

$$\Rightarrow\ \widehat{P} \overset{as.}{\sim} N\left(\Pi, \frac{\Pi(100-\Pi)}{n}\right).$$