I.3. INFERENTIAL STATISTICS : TESTS OF HYPOTHESES AND SIGNIFICANCE

I.3.1. Sampling

Consider a card index with 10,000 entries. We wish to draw a sample from this population of 10,000 cards to find the age distribution of books in the library. To be able to make reliable inferences about the population as a whole, this sample must be large enough and unbiased. In the next section we will discuss some methods to attain this objective.

Let X be the stochastic variable which associates the book's age with every card. This stochastic variable is termed the 'population stochastic variable'. The distribution of X is the population (frequency) distribution.

A sample of size N from this population is then a finite sequence of random variables $X_1, X_2, \ldots, X_N$ with the same distribution as X. We moreover require the X_i's to be independent, i.e. (cf. formula [I.2.5])
$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_N \leq x_N) = P(X_1 \leq x_1).P(X_2 \leq x_2) \ldots P(X_N \leq x_N)$.
For a sample $X_1, \ldots, X_N$ the *sample mean*, $\bar{X}$, is defined as

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \quad ; \qquad [I.3.1]$$

and the *sample variance* is :

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2 \; . \qquad [I.3.2]$$

S is then called the sample standard deviation.

Theorem. If the population stochastic variable X has a mean μ and a variance σ^2 then

$$E(\bar{X}) = \mu \quad \text{and} \quad Var(\bar{X}) = \frac{\sigma^2}{N} \quad ; \qquad [I.3.3]$$

moreover :

$$E(S^2) = \sigma^2 \; . \qquad [I.3.4]$$

Proof. The proof of the formulae [I.3.3] follows immediately from formulae [I.2.14] through [I.2.17]. The proof of formula [I.3.4] is omitted. □

The relations $E(\bar{X}) = \mu$ and $E(S^2) = \sigma^2$ express the fact that $\bar{X}$ and S^2 are *unbiased estimators* for the population mean and variance.

We finally note the following result (the proof is omitted) :

Proposition. *If* $X \sim N(\mu;\sigma^2)$, *then* $\bar{X} \sim N(\mu ; \frac{\sigma^2}{N})$.

I.3.2. General remarks on hypothesis testing

In attempting to reach decisions on statistical grounds, it is useful to make assumptions about the population or populations involved. Such assumptions, which may or may not be true, are called 'statistical hypotheses'.

In many instances we formulate a statistical hypothesis for the sole purpose of rejecting it. For example, if we want to decide whether a given coin is loaded, we formulate the hypothesis that the coin is fair, i.e. $p = 0.5$, where p is the probability of heads. Similarly, if we want to decide whether one procedure is better than another, we formulate the hypothesis that there is no difference between the procedures (i.e. observed differences are merely due to chance fluctuations). Such hypotheses are called '*null hypotheses*', denoted by H_0. Any hypothesis which differs from a stated null hypothesis is termed an '*alternative hypothesis*', denoted by H_1.

Let us next take a closer look at the two kinds of errors that can be made when statistical decisions are taken. If the null hypothesis is rejected when it should be accepted, we say that a *type I error* has been made. If the null hypothesis is accepted when it should be rejected, we make a *type II error*.

Clearly a decision procedure should be such that it eliminates or at least reduces both kinds of error. However, an attempt to decrease one type of error usually increases the other type of error. In practice, one type of error may be more serious than the other, and so the null hypothesis is chosen in such a way that a type I error is worse than a type II error. The only way to reduce both types of error is to increase the sample size, but this is not always possible.

In testing a given hypothesis, the maximum probability with which we are willing to risk a type I error is called the '*level of significance*' of the test. This probability, denoted by α, must be specified before samples are drawn, so that the results obtained will not influence our choice. Once this significance level has been determined, we are saying that we will accept H_0 unless we witness some event which has a sufficiently small probability (α) of occuring when H_0 is true. This is what some people refer to as 'the principle of the disbelief in tall stories'. In practice a level of significance of 0.1, 0.05 or 0.01 is customary. If a hypothesis is accepted at a 0.1 level, it is automatically also accepted at the 0.05 and the 0.01 level.

I.3.3. Central limit theorem

The following *central limit theorem* (given without proof) forms the basis for several statistical tests.

I.3.3.1. Central limit theorem (Lindeberg, 1922)

Let $X_1,\ldots,X_N$ be independent random variables which are identically distributed and have a finite mean μ and a variance σ^2. Then, if

$$Y_N = X_1 + X_2 + \ldots + X_N,$$

$$\lim_{N\to\infty} P(a \leq \frac{Y_N - N\mu}{\sigma\sqrt{N}} \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{t^2}{2}} dt \quad , \qquad [I.3.5]$$

that is, the random variable $(Y_N - N\mu)/\sigma\sqrt{N}$, which is the standardised variable corresponding to Y_N, is asymptotically normal.

For several illustrations of this theorem we refer the reader to Feller (1968, p.244-246).

This theorem, together with the definition of the sample, implies that *if N is large enough (in practical situations $N \geq 30$), the sample mean $\bar{X}$ is normally distributed, even if X is not!* So, we have

$$\bar{X} \sim N(\mu ; \frac{\sigma^2}{N})$$

and hence

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim N(0;1) \; . \qquad [I.3.6]$$

I.3.3.2. A simple verification of the central limit theorem

The validity of the central limit theorem, which strikes most students as odd, can be experimentally verified as follows. Students (at least 30) go to the library and count the number of books on a shelf. Each student counts 10 or 20 shelves and computes the average number of books on one shelf. All these averages graphed together on a histogram will give a distribution which will resemble the normal distribution. Note that in this situation the distribution of the number of books per shelf is not known (and does not have to be known).

I.3.4. Tests of means

I.3.4.1. First test of the population mean

Let us take $H_0 : \mu = \mu_0$, i.e. we want to test whether the population mean is μ_0. If $N \geq 30$,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} \sim N(0;1) \ . \qquad [I.3.7]$$

and if σ is not known - as is usually the case - we estimate σ by s (the observed value of the sample standard deviation) and consider

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{N}} \sim N(0;1) \ . \qquad [I.3.8]$$

Note that the fact that Z is standard normally distributed follows from the central limit theorem.

If $N < 30$ and if X is known to be normally distributed, then we have, according to the proposition in Section I.3.1,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{N}} \sim N(0;1) \ , \qquad [I.3.9]$$

if σ is known, and

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{N}} \sim t(N-1) \ , \qquad [I.3.10]$$

if we use s as an estimate for σ.

In the first case the stochastic variable Z used for the test is standard normally distributed. For this reason such tests are often referred to as 'z-tests'. In the latter case we use a Student's distribution : this test is called a 't-test'. We note that the t-test may always be applied and that the z-test is a good approximation of it for large N. However, we recommend the use of samples of which the size is larger than thirty.

There are three possible forms for the alternative hypothesis H_1 :

1) $H_1 : \mu \neq \mu_0$ (leading to a two-sided test)
2) $H_1 : \mu > \mu_0$
3) $H_1 ; \mu < \mu_0$ (both leading to a one-sided test).

In the first case we will reject H_0 if the value of Z (denoted by z) is larger than y or smaller than -y, where y is the critical point yielding an area under the distribution curve to the right of y, equal to $\alpha/2$. For example,

if $\alpha = 0.05$ and $N > 30$, we find in the table for the standard normal distribution (Table A.1) that $y = 1.96$. So, in this case the acceptance region for H_0 is $]-1.96,+1.96[$. Figure I.3.1 illustrates this situation.

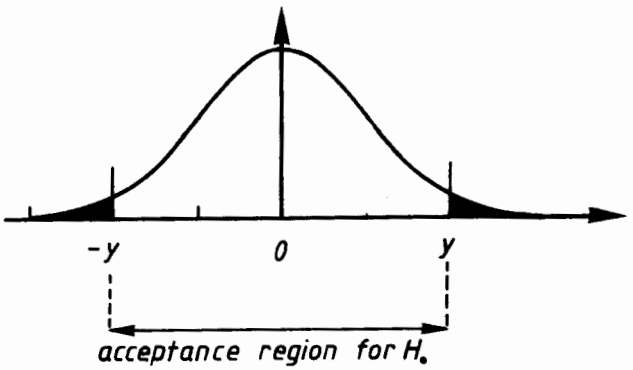

Fig.I.3.1 Acceptance region and critical regions for a two-sided test on the normal distribution

In the second case ($H_1 : \mu > \mu_0$) we will reject H_0 if z is too large. It only makes sense to do such a one-sided test if it is known for certain that μ is at least as large as μ_0 or if it makes no difference whether $\mu = \mu_0$ or $\mu < \mu_0$. The critical point y is here chosen in such a way that the area under the distribution curve to the right of y is equal to α.

Similarly, in the third case ($H_1 : \mu < \mu_0$), H_0 is rejected if z lies outside the acceptance region $]-y,+\infty]$, where y is the same number as in the preceding case.

I.3.4.2. Examples

1) Suppose that we know that an abstract in the ECONOMIC LITERATURE INDEX contains an average of $\mu = 79.56$ words, with a standard deviation of $\sigma = 24.80$. When examining forty abstracts written in German, we observe an average of 67.47 words. Is there a significant difference in the number of words between German abstracts and the general average?

We decide to do a test on the 1 % level. As there is no a priori reason to think that German abstracts are shorter or longer than the average, we

will do a two-sided test. We take :

$$H_0 : \mu = 79.56 ,$$

$$H_1 : \mu \neq 79.56 .$$

For a two-sided test on the 1 % level, the region of acceptance, based on the standard normal distribution, is :]-2.576,+2.576[. Now, using [I.3.7], $z = \frac{(67.47 - 79.56)}{24.8/\sqrt{40}} = -3.08$. So, we reject H_0 : meaning that we reject the hypothesis that German abstracts do not differ from the average case, on the 1 % level (hence also on the 5 % or the 10 % level).

2) We use the same data as in the preceding case but now we assume, more realistically, that we do not known the standard deviation σ. Assume, however, that we do know the sample variance $s^2 = 669$. We again perform a two-sided test on the 1 % level : $H_0 : \mu = 79.56$; $H_1 : \mu \neq 79.56$, with the same region of acceptance as in the first sample. In this case, the value of Z is, using [I.3.8] :

$$\frac{67.47 - 79.56}{\sqrt{669/40}} = -2.96 .$$

We again reject H_0 and conclude that the average number of words in German abstracts is different from the overall average.

I.3.4.3. A test for fractions

Based on the equations of Subsection I.3.4.1, we can also test *fractions*. Let N denote the number of items in the sample studied, p the proportion of the sample possessing the characteristic under study, and P the underlying, unknown proportion of the population which possesses the characteristic. When N is large, the sample proportion is approximately normally distributed with a mean of μ equal to P and a variance of σ^2 equal to PQ/N (where Q = 1-P). When (as is usually the case) we only know p, the observed fraction, we use $\frac{p(1-p)}{N-1}$ instead of $\frac{PQ}{N}$ (as in Subsection I.3.4.1 - see also formula [I.3.2]). The equations $\mu = P$ and $\sigma^2 = \frac{PQ}{N}$ can be proved, using the binomial distribution.

In this case the null hypothesis H_0 is $\mu = P$ and the alternative hypothesis is $H_1 : \mu \neq P$. We then consider

$$\frac{P - p}{\sqrt{\frac{p(1-p)}{N-1}}} \sim N(0;1) \quad \text{or} \quad t(N-1) \qquad [I.3.11]$$

according to $N \geq 30$ or $N < 30$.

In accordance with Fleiss (1981, p.13) a correction for continuity bringing binomial probabilities closer to normal curves, is needed when $1/2N < |P - p|$. We then use

$$\frac{|P - p| - 1/2N}{\sqrt{\frac{p(1-p)}{N-1}}} \quad . \qquad [I.3.12]$$

This test was used, for example, by Buckland et al. (1975) to study the overlap in bibliographic files and library holdings.

I.3.4.4. Confidence intervals for the population mean μ

Based on the results obtained in Section I.3.4.1, we can find a solution for the following problem. Suppose we draw a sample yielding a sample mean $\bar{x}$. Construct an interval $[\bar{x}-a,\bar{x}+a]$ such that the population mean μ belongs to this interval, with a confidence level of 100 $(1-\alpha)$ %.

We will give the solution based on the case of large samples ($N \geq 30$) with an unknown variance and $\alpha = 0.05$. It is then easy to find the solution in other cases. In the situation described above we know that

$$P(-1.96 \leq \frac{\bar{x} - \mu}{s/\sqrt{N}} \leq 1.96) = 0.95 \ .$$

Consequently :

$$P(\bar{x} - 1.96 \frac{s}{\sqrt{N}} \leq \mu \leq \bar{x} + 1.96 \frac{s}{\sqrt{N}}) = 0.95 \ .$$

Hence we have found the following 95 % *confidence interval* for the population mean μ :

$$[\bar{x} - 1.96 \frac{s}{\sqrt{N}} \ , \ \bar{x} + 1.96 \frac{s}{\sqrt{N}}] \quad . \qquad [I.3.13]$$

Note that confidence intervals depend on sample size N : the larger N is, the smaller the length of the confidence interval will be.

Confidence intervals are usually visualised by using 'error bars', i.e. an interval positioned vertically, with the observed mean in the centre and where its length is equal to that of the corresponding confidence interval (cf. Fig.I.3.2).

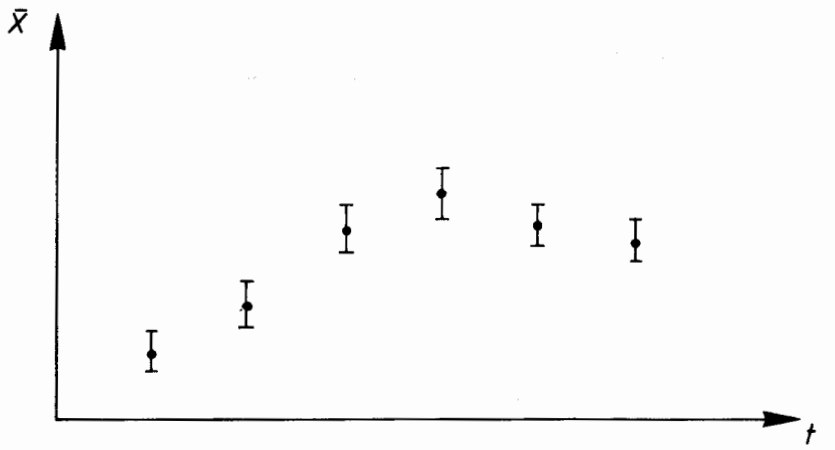

Fig.I.3.2 Data and confidence intervals

I.3.4.5. Second test on the mean : two measurements on the same sample

An example will clarify this case. Suppose we want to investigate the retrieval time of two online catalogues A and B. We determine a number of books belonging to both systems and measure the times needed by the online catalogue systems to retrieve them. Let X be the time (in seconds) needed by system A and Y the time needed by system B to retrieve the same book. Table I.3.1 gives the result for a sample of size 14.

Table I.3.1. Retrieval time of two online catalogues

X	Y	Y-X
6	21	15
12	13	1
8	72	64
28	13	-15
13	40	27
12	51	39
48	34	-14
14	32	18
17	28	11
21	43	22
24	33	9
10	24	14
6	21	15
3	21	18

When Y-X is considered, this situation is immediately reduced to the case of the mean of one sample and $H_0 : \mu_{Y-X} = 0$. In this example N is small and σ is unknown, so that we will have to perform a t-test. The sample mean of Y-X is 16.0 and the sample variance is $(19.93)^2$.

The t-value $\frac{\bar{x} - \bar{y} - 0}{s/\sqrt{14}}$ equals $\frac{16}{19.93/\sqrt{14}} = 3.00$.

The region of acceptance for a two-sided test (hence $H_1 : \mu_{Y-X} \neq 0$) on a 5 % level, based on t(13), is :]-2.16,+2.16] (cf. Table A.2). Hence we reject the null hypothesis and conclude, on a 5 % level, that both systems have a different retrieval time.

This test was used, for example, in Rousseau (1988a) to compare a two-year and a four-year impact factor of mathematics journals. (For the notion of a journal's impact factor, the reader is referred to Chapter IV.5).

I.3.4.6. Third test on the mean : measurements on different samples

The preceding test cannot be used when X and Y are independent (e.g. distributions of different populations). Consider, for instance, the case of two booksellers, and suppose that we want to test whether they have the same delivery time for books. It would be very uneconomical to order the same books at both booksellers. So we will order different books, distributed at random to bookseller A and bookseller B. To investigate whether the mean delivery times differ significantly, we have to develop a new test.

Generally speaking, we are in a situation in which there are two populations A and B. A sample of size N_1 is drawn from A and a sample of size N_2 is drawn from B. Let $x_1,\ldots,x_{N_1}$ be the observed values for the first sample and $y_1,\ldots,y_{N_2}$ the observed values for the second one. From Subsection I.2.3.2 and the theorem in Section I.3.1 we know that the function $\bar{Y} - \bar{X}$ is a random variable with the following characteristics :

$$E(\bar{Y} - \bar{X}) = \mu_2 - \mu_1 ,$$

$$Var(\bar{Y} - \bar{X}) = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2} ,$$

where μ_1 and σ_1^2 are the population mean and variance of population A, and μ_2 and σ_2^2 are the population mean and variance of population B. In testing for $H_0 : \mu_2 - \mu_1 = 0$, we can apply the first test of the population mean (see Subsection I.3.4.1). In this way, for N_1 and N_2 which are large enough (i.e. N_1 and $N_2 \geq 30$) :

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim N(0;1) \qquad [I.3.14]$$

if σ_1 and σ_2 are known.

If σ_1 and/or σ_2 are not known, we use

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} \sim N(0;1) \quad , \qquad [I.3.15]$$

with

$$S_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} (X_i - \bar{X})^2 \ ,$$

$$S_2^2 = \frac{1}{N_2 - 1} \sum_{j=1}^{N_2} (Y_j - \bar{Y})^2 \ .$$

If, however, N_1 or N_2 is small, and if moreover the two populations are normally distributed, we use

$$\frac{\bar{Y} - \bar{X}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}} \sim N(0;1) \quad , \qquad [I.3.16]$$

if σ_1 and σ_2 are known. If not, and if σ_1 and σ_2 are equal, we use

$$\frac{\bar{Y} - \bar{X}}{S_p \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \sim t(N_1 + N_2 - 2) \ , \qquad [I.3.17]$$

with

$$S_p^2 = \frac{1}{N_1 + N_2 - 2} \left((N_1 - 1)S_1^2 + (N_2 - 1)S_2^2\right) \ .$$

As in the previous cases we advise to take samples of size at least thirty. Returning now to booksellers A and B, we found that they delivered 100 books each. Bookseller A had a mean delivery time of 95.9 days with a standard

deviation equal to 97.39, and bookseller B had a mean delivery time of 85.7 days and a standard deviation of 114.98.

Using [I.3.15] yields :

$$z = \frac{95.9 - 85.7}{\sqrt{\frac{9485 + 13220}{100}}} = 0.677 \quad .$$

A two-sided test on a 10 % level has an acceptance region of]-1.645,+1.645[(see Table A.1). Hence we accept H_0, and we cannot conclude on the basis of these data that bookseller B is a better (i.e. faster) bookseller than A.

A similar example can be obtained by comparing delivery times for books from different countries. This third test on the mean, in the form given by [I.3.17], was used, for example, by Hurt (1980) to show that two studies on highly cited old papers yielded no significant difference on the number of citations.

If sample sizes are small ($N < 30$), population variances are unknown and there is no reason to believe they are equal, we have a so-called Behrens-Fischer problem. Welch (1947) derived a series solution to obtain critical values in this case. This solution was further manipulated by Aspin (1948) and tabulated to two decimals in the form of tables. These tables can be found in Pearson and Hartley (1966). The use of these tables to solve the Behrens-Fischer problem has since become known as the Welch-Aspin method.

An approximate solution to the Behrens-Fischer problem was recently proposed by Aucamp (1986). It consists of a simple z-test with a variance correction factor. It works well for a significance level of $\alpha \geq 0.05$ and sample sizes which are not extremely small (say not smaller than 10).

If H_0 is $\mu_1 = \mu_2$, then Aucamp uses the correction factor

$$F = \sqrt{1 + \frac{2C^2}{N_1 - 1} + \frac{2(1-C)^2}{N_2 - 1}} \qquad [I.3.18]$$

where $C = \dfrac{\frac{S_1^2}{N_1}}{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}$, and considers

$$Z = \frac{\bar{Y} - \bar{X}}{F\sqrt{\frac{S_1^2}{N_1} + \frac{S_1^2}{N_2}}} \sim N(0;1) \quad . \qquad [I.3.19]$$

This test was used in Rousseau (1988a).

I.3.5. Chi-square tests

I.3.5.1. Test on the variance of a normal distribution

To test the hypothesis that a normally distributed population has variance σ^2 (on the basis of a sample of size N) one considers the random variable

$$\chi^2 = \frac{(N-1)\ S^2}{\sigma^2} \qquad [I.3.20]$$

which has a chi-square distribution with (N-1) degrees of freedom. This test is not very often used in informetrics.

I.3.5.2. χ-square test for goodness of fit (χ = chi)

In this situation (also called the one-sample case) chi-square is used to test how well an observed set of frequencies produced by a sample investigation fits a theoretical frequency distribution.

We illustrate the method by testing whether the arrivals at a circulation desk in a library follow a Poisson distribution. Suppose that we have observed the number of arrivals in sixty consecutive one-minute intervals and found the data given in Table I.3.2.

Table I.3.2. Arrivals at a circulation desk

k	O(k)	E(k)
0	4	3.3
1	12	9.6
2	12	13.9
3	14	13.4
4	6	9.7
5	6	5.6
6	4	2.7
7	1	1.1
8	0	0.4
9	1	0.1
10 or more	0	0.2

k : number of arrivals in a 1-minute interval
O(k) : observed number of cases
E(k) : expected number of cases (see text)

We next propose a Poisson distribution to describe the observed frequencies and estimate the parameter λ (cf. Subsection I.2.4.2) from the data, i.e. we take $\lambda = \bar{x} = 2.87$. Note that the observed variance equals 3.54, which is not equal to $\bar{x}$, but does not differ too much either. As a null hypothesis for the observed data we suppose that it is a Poisson distribution with a parameter of $\lambda = 2.9$. The alternative hypothesis is simply that H_0 is not true (for whatever reason). Hence, under the null hypothesis, we expect the following frequencies

$$E(k) = 60\ P(X=k) = 60\ \frac{e^{-2.9}\ (2.9)^k}{k!}\ ,\quad k = 0,1,2,\ldots$$

This yields the third column of Table I.3.2.

Before we can perform the χ^2-test, there is another matter to be taken into account : there should not be many categories for which the expected frequency is small. What is meant in this context by 'many' and 'small' is a matter of dispute amongst statisticians, but it is safe to adopt the rule that no expected frequency should be less than five. If this rule has been broken, one is allowed to combine categories until the offending expected frequencies have been suitably increased. Applying this to Table I.3.2 results in Table I.3.3.

Table I.3.3. Circulation desk data with contracted classes

i	k	O_i	E_i	$(O_i-E_i)^2/E_i$
1	0-1	16	12.9	0.745
2	2	12	13.9	0.260
3	3	14	13.4	0.027
4	4	6	9.7	1.411
5	5 or more	12	10.1	0.357
				2.800

i : contracted classes
k : number of arrivals in a 1-minute interval
O_i : observed number of cases
E_i : expected number of cases
$(O_i-E_i)^2/E_i$: terms in the calculation of the χ^2-value.

The *chi-square statistic* is calculated as follows :

$$\chi^2 = \sum_i \frac{(O_i-E_i)^2}{E_i}\ , \qquad [I.3.21]$$

where summation is done over all categories of the contracted table. In our example the χ^2-value is 2.8. This value should be compared with the critical value of the $\chi^2(n-m-1)$ distribution, where n is the number of categories in the contracted table and m is the number of estimated parameters (we omit the proof). Here we have estimated one parameter, namely λ, and as $n = 5$ we find the following region of acceptance for a test on the 0.05 level : $[0,7.81[$ (see Table A.3). Since $2.8 \in [0,7.81[$, we accept the hypothesis that in this situation arrival data are Poisson distributed with a mean of 2.9.

Note that in order to apply this test, the data must be frequencies, i.e. the number of discrete objects occurring in different categories. The results of the chi-square test will usually be false when applied to data involving proportions or percentages. Also, the categories must be mutually exclusive, so that one individual cannot possibly be counted in more than one category.

When results for continuous distributions, such as χ^2, are applied to discrete situations (such as e.g. in the above example of a discrete Poisson distribution), certain corrections for continuity can be made. In this case, it consists of using

$$x^2 = \sum_i \frac{(|O_i - E_i| - 0.5)^2}{E_i} \quad . \qquad [I.3.22]$$

This modification is known as 'Yates' correction'.

A chi-square test for the goodness of fit has been applied in numerous cases : see e.g. Simonton (1986a,b), Allison (1980), Cohen (1980).

I.3.5.3. Tests of independence in contingency tables

A *contingency table* is a multiple classification. Items under study are classified according to two criteria, one having m categories and the other n categories, giving an (m,n) matrix, called a contingency table. This $m \times n$ distinct classifications are called cells. Cell frequencies are denoted by O_{ij} and $\Sigma O_{ij} = N$.

If different categories are mutually exclusive, the probability that an item belongs to the k^{th} category, according to the first criterion, is $\sum_{j=1}^{n} O_{kj}/N$ and the probability that it belongs to the ℓ^{th}, according to the second criterion, is $\sum_{i=1}^{m} O_{i\ell}/N$. Bearing in mind the fact that two events A and B are independent if $P(A \cap B) = P(A)P(B)$ [I.2.5], there is independence between the two criteria if for every k and ℓ : P(item belongs to cell (k,ℓ)) = P(item

belongs to row k).P(item belongs to column ℓ), or

$$\frac{\sum_{j=1}^{n} O_{kj}}{N} \cdot \frac{\sum_{i=1}^{m} O_{i\ell}}{N} = \frac{O_{k\ell}}{N}$$

or

$$\frac{\sum_{j=1}^{n} O_{kj} \cdot \sum_{i=1}^{m} O_{i\ell}}{N} = O_{k\ell} \cdot$$

Cell frequencies given by this formula are expected values $E_{k\ell}$ under the null hypothesis of independence. These are compared with observed frequencies $O_{k\ell}$. Then

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad [I.3.23]$$

is χ^2-distributed with (m-1)(n-1) degrees of freedom. If the expected frequencies can be computed only by estimating h population parameters, we have a χ^2-distribution with (m-1)(n-1) - h degrees of freedom. We omit the proofs.

In this case as well, if expected cell frequencies are too small (<5), we have to combine categories. Yates' correction can also be applied here, in particular for small (e.g. 2×2) contingency tables. It is recommended by Kuntz and Mayo (1979, p.372-378), to apply Yates' correction for every 2×2 table. These authors also give a direct discrete test for such contingency tables.

An example. Search keys are more efficient tools for searching automated library catalogues than complete author and title information. For the OCLC-network Kilgour and co-workers developed simple search keys, which eliminate problems associated with authors' first and middle names or when the user's knowledge of the author and title are incomplete (cf. Kilgour (1968), Kilgour et al. (1970), Long and Kilgour (1971), Kilgour et al. (1971), Long and Kilgour (1972)). These search keys consist of the first few letters of the author's last name concatenated with the first few letters of the first non-article word of the title. For example, the 3-3 search key for Gerard Salton and Michael J. McGill's 'Introduction to modern information retrieval' is SALINT.

The chief disadvantage of such algorithms is that one search key may correspond to several books. Still, this lack of precision is not really

harmful as long as the number of false drops is low.

One way to minimise the problem of false drops would be to use longer search keys since a longer search key will correspond to fewer books. But, as reported by Kilgour (1968), a longer search key decreases the chances of the desired book being found, even if it is in the file.

Another possible way to decrease the number of false drops would be to divide the file into subject areas. We report here on an investigation done by Kjell (1974) to determine whether Kilgour's algorithm works equally well in different subject files. Using the 3-3 key described above on 4148 MARC records resulted in Table I.3.4. The distinction between science and technology books and art and literature books was based on Dewey's classification. Table I.3.4 must be read as follows : the number 76 in the column headed 'double' means that 76 keys were members of matching pairs, i.e. that there were 38 pairs.

Table I.3.4. Kjell's observed values of matching 3-3 keys

	single	double	>2	row sums
sci/tech	1958	76	33	2067
art/lit	2032	46	3	2081
column sums	3990	122	36	4148

Under the null hypothesis of independence, i.e. the 3-3 key works equally well in sci/tech as in art/lit, we have the following table (Table I.3.5) of expected values. Here E_{11} is calculated as explained above : $\frac{(2067)(3990)}{4148} =$ 1988.3 and similarly for the other entries. Note that, by construction, the column and row sums are the same as for the contingency table of the observed values.

Table I.3.5. Contingency table : expected values associated with Table I.3.4

	single	double	>2	row sums
sci/tech	1988.3	60.8	17.9	2067
art/lit	2001.7	61.2	18.1	2081
column sums	3990	122.0	36.0	4148

Computing [I.3.23] yields : $\frac{(1958-1988.3)^2}{1988.3} + \frac{(76-60.8)^2}{60.8} + \frac{(33-17.9)^2}{17.9} + \frac{(2032-2001.7)^2}{2001.7} + \frac{(46-61.2)^2}{61.2} + \frac{(3-18.1)^2}{18.1} = 33.8$.

The critical value for a test on the 1 % level is 9.21 ($\chi^2(2)$) (see Table A.3). Consequently, the null hypothesis of independence is rejected.

To get an idea of which cells are responsible for a high χ^2-value, it is interesting to construct a table of $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ -values. In this example this produces Table I.3.6, showing that the difference in the occurrence of multiple keys (> 2) is the cause of the high χ^2-value.

Table I.3.6. χ^2-values of Table I.3.4 versus Table I.3.5

	single	double	> 2
sci/tech	0.46	3.8	12.7
art/lit	0.46	3.8	12.6

Other applications of the χ^2-test for independence can be found, for example, in Kretschmer and Vlachý (1986), Tagliacozzo (1977), Bonzi (1982), Jamieson et al. (1986), Whitley (1969).

I.3.6. The Kolmogorov-Smirnov test (Kolmogorov (1933, 1941), Smirnov (1939))

This test is a goodness of fit test used to compare an observed frequency distribution with a theoretical one. To be able to apply it, distributions have to be converted into cumulative probability distributions. This implies that data must at least be ordinal; it cannot be applied on nominal data. On the other hand, there is no minimal requirement for cell occupancies. The null hypothesis for this test is that sample data have been drawn from a specified theoretical distribution. The *Kolmogorov-Smirnov statistic*, denoted by D, is simply the maximum absolute difference between the theoretical and the observed cumulative probability distributions (denoted respectively by S_N and F). The degrees of freedom for the Kolmogorov-Smirnov goodness of fit test are the number of items in the observed frequency distribution (not the number of cells!). If the calculated value of D is greater than the tabulated critical value at the specified significance level, the null hypothesis is rejected. A table for the K-S test can be found in the appendix (Table A.4).

Using the data of Table I.3.2 again, we do a K-S test on the 5 % level. The null hypothesis is that data are sampled from a Poisson distribution with a parameter $\lambda = 2.9$. Original data and cumulative probability distributions are given in Table I.3.7.

Table I.3.7. Table for the Kolmogorov-Smirnov test on Poisson data of Table I.3.2

k	O(k)	$S_N(k)$	$\sum_{i \le k} O(i)$	$\sum_{i \le k} E(i)$	F(k)	$\lvert F(k) - S_N(k) \rvert$
0	4	0.067	4	3.3	0.055	0.012
1	12	0.267	16	12.9	0.215	0.052
2	12	0.467	28	26.8	0.447	0.020
3	14	0.700	42	40.2	0.670	0.030
4	6	0.800	48	49.9	0.832	0.032
5	6	0.900	54	55.5	0.925	0.025
6	4	0.967	58	58.2	0.970	0.003
7	1	0.983	59	59.3	0.988	0.005
8	0	0.983	59	59.7	0.995	0.012
9	1	1.000	60	59.8	0.997	0.003
10 or more	0	1.000	60	60.0	1.000	0.000
sum	60					

Hence : $D = 0.052$.

The critical value for a test on the 5 % level with 60 degrees of freedom is $1.36/\sqrt{60} = 0.176$ (see Table A.4). We accept the null hypothesis that data are Poisson distributed with a parameter $\lambda = 2.9$.

We stress the fact that the hypothesised cumulative distribution F must be specified in advance; when parameters are unknown and must be estimated from the data, standard tables of the K-S test are, in principle, not valid.

If the theoretical distribution is continuous, its cumulative probability distribution is also continuous. (The test has, in fact, been conceived for a continuous distribution, but it can also be used for a discrete one; in this case the test is somewhat conservative.) Thus, at every jump of the observed relative cumulative distribution function S, there are two differences between F and S. The appropriate procedure to follow is to calculate.

$\lim_{y \to x,\ y<} |F(y) - S(y)|$ and $\lim_{z \to x,\ z>} |F(z) - S(z)|$ for every jump point x, and to set D equal to the maximum over all these differences.

The Kolmogorov-Smirnov test is the best testing procedure for tests on informetric distributions (see Part IV and Pao (1985), Pao (1986), Nicholls (1986)). We end this section on K-S by mentioning that there are extensions

of the classical form we have given here (see e.g. Durbin (1975), Edgeman et al. (1988), Guilbaud (1988)).

I.3.7. Some other nonparametric tests

So-called parametric tests make assumptions about the distribution of values in the population from which samples are taken. Nonparametric, or distribution-free methods, do not involve such assumptions. Tests such as those on the mean of an observed distribution are parametric tests; the Kolmogorov-Smirnov test is not. It is generally argued that a parametric test, used in a situation in which its assumptions are justified, is more powerful than an equivalent nonparametric method.

Often, however, parametric tests cannot be applied, as there is no a priori information on the underlying population distribution. In these cases it is usually necessary to simplify the original stochastic variable. This means that one uses only ranks. For this reason nonparametric statistics are often termed '*rank order statistics*'. These methods are particularly suited for ordinal data.

I.3.7.1. Mann-Whitney U-test (Mann and Whitney (1947))

The *Mann-Whitney U-test* is a test on the difference between two samples with respect to their position on an ordinal scale. This test can be considered as a nonparametric analogue of tests on the differences of means (see Section I.3.4).

Suppose we are interested in the question whether sociologists at university A are more productive than sociologists at university B, or whether observed differences in output can be attributed to chance fluctuations. Therefore we consider their publication lists over the last ten years. Results are given in Table I.3.8. Note that seven sociologists work at university A and eleven at university B.

Table I.3.8. Number of publications of sociologists at university A and at university B

first row : affiliation
second row : number of publications
third row : rank (from lowest producer to highest)

B	B	B	A	B	A	B	A	B	B	B	A	B	B	A	A	B	A
7	8	11	12	14	15	17	19	20	26	32	40	49	57	61	76	94	102
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

The test is derived from the following line of reasoning. If the publication output of the sociologists at these universities differs strongly, the lower numbers of publications will mainly be found for sociologists of one university and the higher numbers for the others. In the most extreme case the lowest ranks will all be assigned to one group and the highest ranks to the other group. If the first group has m members and the second one n members and if the members of the second group all publish more than those of the first, the sum of the ranks of the second group, denoted generally by T_2, will be maximal. This maximal sum is equal to $nm + n(n+1)/2$. Indeed, in this extreme case, the members of the second group occupy ranks m+1 up to m+n. The sum of these ranks is the sum of the first m+n natural numbers minus the sum of the first m natural numbers. This is :

$$\frac{(m+n)(m+n+1)}{2} - \frac{m(m+1)}{2} = (m+n)(\frac{m+1}{2} + \frac{n}{2}) - m(\frac{m+1}{2}) = mn + \frac{n(n+1)}{2}$$

If the ranks of both groups are mixed, T_2 will be smaller than this maximum. This is the basic idea for considering the statistic U_2 calculated by

$$U_2 = mn + \frac{n(n+1)}{2} - T_2 \quad . \qquad [I.3.24]$$

U_2 is small when groups differ greatly and large when groups differ little. Of course, a symmetry argument shows that U_2 can be large when the ranks of the elements in the second group are the lowest. But in this case the roles of the first and the second group are interchanged. Consider

$$U_1 = mn + \frac{m(m+1)}{2} - T_1 \qquad [I.3.25]$$

where T_1 is the sum of the ranks of the elements in the first group, then use $U = \min (U_1, U_2)$ (since tables for the Mann-Whitney test are based on the smallest of these two U_i's). We now have that U is small if and only if groups differ greatly and large if and only if groups differ little. We are interested in high values of U since the null hypothesis is that both groups do not differ. However, it is not necessary to compute both U_1 and U_2, as they are related through the formula $U_1 + U_2 = mn$. Indeed :

$$U_1 = mn + \frac{m(m+1)}{2} - T_1 \quad \text{and} \quad U_2 = mn + \frac{n(n+1)}{2} - T_2 \quad .$$

Now, $T_1 + T_2$ is equal to the sum of the first (m+n) natural numbers, so that

$$U_1 + U_2 = 2\,mn + \frac{m(m+1)}{2} + \frac{n(n+1)}{2} - \frac{(m+n+1)(m+n)}{2}$$

$$= mn$$

Applying this procedure to the data in Table I.3.8 yields $T_1 = 79$ and $T_2 = 92$, and thus $U_1 = 26$ and $U_2 = 51$. So we will use the value 26. For a test on the 5 % level the acceptance region for the null hypothesis is $]19,+\infty[$. This means that we accept that the publication output of the sociologists of both universities does not differ significantly.

If m and n are both larger than 20, U is approximately normal with a mean of mn/2 and a variance of nm(n+m+1)/12. Standardising and applying a continuity correction produces

$$Z = \frac{(U + 0.5) - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim N(0;1) \quad . \qquad [I.3.26]$$

This test has been used by, for example, Smart and Bayer (1986) to study differences in the citation rates of single and multiple authored articles.

I.3.7.2. The Wald-Wolfowitz runs test (Wald and Wolfowitz (1940))

This test has the same purpose as the Mann-Whitney test : to determine whether two groups of data differ with respect to their position on an ordinal scale. However, a different procedure is used to accomplish this.

We consider the same example as in Subsection I.3.7.1. Again, sociologists are ordered according to their 10-year production. The basic idea here is that if both groups differ strongly, members of one group will follow one another closely. If, on the other hand, both groups are similar, members of both groups will alternate. The number of *runs* (groups of consecutive items of the same kind) is used as the test statistic (denoted by R).

For the example of Subsection I.3.7.1 this is :

B B B A B A B A B B B A B B A A B A

The number of runs is R = 12. Table A.6 (appendix) lists the critical values on the 5 % level. Under the null hypothesis that there is no difference between

groups, R-values that are too small lead to the rejection of H_0. In our case we accept H_0 if $R > 5$. So, we accept the null hypothesis that the publication output of the sociologists of both universities does not differ significantly.

In this case as well, if m and n are larger, i.e. $m+n > 20$, we use the table for the standard normal distribution (Table A.1). Indeed, using the cell occupancy theory (cf. Section I.2.5 and Fig.I.2.2), one can prove that when m and n are large, R is (approximately) normally distributed with a mean $\mu_R = \frac{2mn}{m+n} + 1$ and a variance of $\sigma_R^2 = \frac{2mn(2mn-n-m)}{(n+m)^2 (n+m-1)}$.
Hence

$$Z = \frac{(R + 0.5) - \mu_R}{\sigma_R} \sim N(0;1) \quad , \qquad [I.3.27]$$

where we have also used a continuity correction. The test is then a one-sided test to the left : only values which are too small i.e. negative and large in absolute value, lead to a rejection of the null hypothesis.

I.3.8. Regression and correlation

To study the relation between two quantitative variables, researchers use scatter diagrams and the notions of covariance and correlation coefficient. If the correlation coefficient is close to 1 (in absolute value) it makes sense to fit a linear model.

I.3.8.1. Covariance

Consider a number of observations (x_1,y_1), (x_2,y_2), ..., (x_N,y_N); the first expression used to measure the relation between the stochastic variables X and Y is the (sample) *covariance* defined as :

$$S_{X,Y} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \quad . \qquad [I.3.28]$$

The covariance can be considered as a generalised variance. If $S_{X,Y} \neq 0$, there is some relation between X and Y. A positive covariance means that if X has a large value, Y also has a large value and similarly for small values; a negative covariance indicates an opposite relation between X and Y. If X and Y are independent, their covariance is zero.

Note that the covariance measures a linear relation : it is possible for stochastic variables to have a perfect - nonlinear - relation, but for their covariance to be zero.

The covariance is a good measure of observations measured on an absolute or a difference scale. If observations are measured on a ratio or an interval scale, the covariance depends on the units chosen. Of course, the notion of covariance has no meaning for nominal or ordinal data. Nevertheless, we will consider some nonparametric tests for the latter cases.

The covariance we have defined in [I.3.28] is the so-called sample covariance; for the stochastic variables X and Y themselves the *covariance* is defined as :

$$\mathrm{Cov}(X,Y) = E((X-\mu_X)(Y-\mu_Y)) = E(XY) - \mu_X\mu_Y \ , \qquad \text{[I.3.29]}$$

which is meaningful whenever X and Y have finite variances.

I.3.8.2. The product-moment correlation coefficient or Pearson's correlation coefficient

A measure of association which also makes sense for variables measured on a ratio or an interval scale is the *product-moment correlation coefficient* (often abbreviated to : *correlation coefficient*). The sample correlation coefficient is

$$R_{X,Y} = \frac{S_{X,Y}}{S_X S_Y} \qquad \text{(or simply R)}$$

$$= \frac{N \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{\sqrt{\left(N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2\right)\left(N \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)^2\right)}} \ , \qquad \text{[I.3.30]}$$

where S_X and S_Y are the sample standard deviations of X and Y (see [I.3.2]).

In general, for stochastic variables X and Y their *correlation coefficient* $\rho_{X,Y}$ (also denoted by $\rho(X,Y)$ or simply ρ) is defined as :

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \text{[I.3.31]}$$

Obviously $\rho(a_1X + b_1, a_2Y + b_2) = \rho(X,Y)$; $a_1,a_2,b_1,b_2 \in \mathbb{R}$; $a_1,a_2 > 0$. From Subsection I.3.8.1 we already know that if X and Y are independent, their covariance, and hence also their correlation coefficient, is zero.

The correlation coefficient is by no means a general measure of dependence between X and Y. However, $\rho(X,Y)$ is connected to the linear

dependence of X and Y. This is stated in the following theorem (Feller (1968, p.236)) (without proof) :

Theorem. We always have $|\rho(X,Y)| \leq 1$; *furthermore,* $|\rho(X,Y)| = 1$ *if and only if constants a and b exist such that* $Y = aX + b$.

In an interesting tutorial paper Rodgers and Nicewander (1988) sketch the history of the product-moment correlation coefficient (showing, among other things, that it is seemingly more appropriate to call it the 'Galton-Pearson coefficient') and discuss thirteen different ways to look at and interpret the correlation coefficient.

I.3.8.3. Scatterplots

Scatterplots of data such as in Fig.I.3.3 illustrate the connection between Pearson's correlation coefficient, S_X, S_Y and the straight-line behaviour.

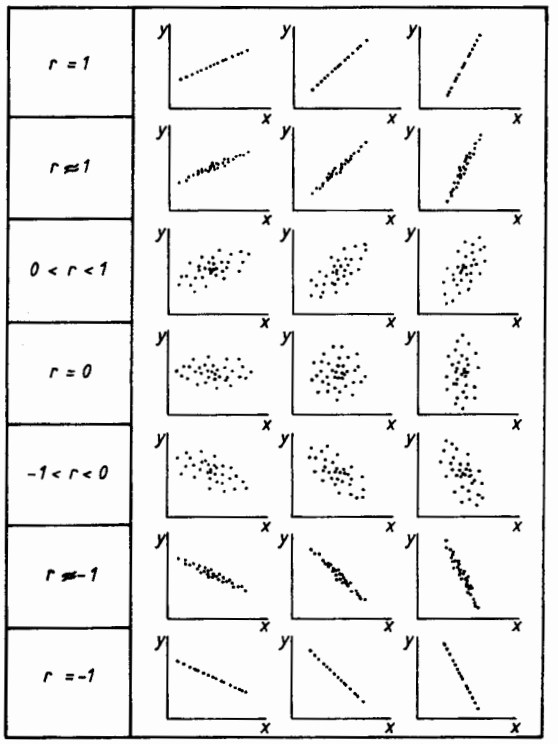

Fig.I.3.3 Correlation coefficients of different scatterplots

I.3.8.4. Linear regression

The correlation coefficient is a measure of the strength of the linear relation between two variables. When there is a perfect linear relationship between X and Y, $\rho(X,Y)$ is equal to 1 or to -1. However, knowing the correlation coefficient does not characterise the form of the linear relation. Still, we will need this exact form to be able to do predictions.

There are several methods and criteria which can be used to fit a straight line to a set of data. A line which minimises the sum of the squares of the distances from the observed points to the line, measured parallel to the vertical axis, is known as the least-squares line. This is the most commonly used best-fit line in informetric studies. *Linear regression* is the name conventionally assigned to a technique for calculating the equation of a least-squares line.

The equation of a straight line is given by (cf. Fig.I.3.4) :

$$y = a + bx \quad , \qquad [I.3.32]$$

where the constants a and b are said to be the intercept and the slope of the straight line. The intercept and the slope of the least-squares line of the set of data points (x_1,y_1), (x_2,y_2), ..., (x_N,y_N) are determined by expressing that the function

$$f(a,b) = \sum_{i=1}^{N} (y_i - (a + bx_i))^2 \qquad [I.3.33]$$

must be minimal. In this way the straight line $y = a + bx$ minimises the sum of squares of the distances from the observed points to the line, measured parallel to the vertical axis.

A result from the calculus of functions of several variables requires in a necessary way that

$$\frac{\partial f}{\partial a} = 0$$

and

$$\frac{\partial f}{\partial b} = 0 \ .$$

This produces respectively

$$-2 \sum_{i=1}^{N} (y_i - (a + bx_i)) = 0$$

and

$$2 \sum_{i=1}^{N} (y_i - (a + bx_i))\, x_i = 0 \ .$$

Consequently, we obtain the following system of linear equations :

$$\begin{cases} \sum_{i=1}^{N} y_i = Na + b \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i y_i = a \sum_{i=1}^{N} x_i + b \sum_{i=1}^{N} x_i^2 \ . \end{cases}$$

Solving for a and b yields

$$b = \frac{N \sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} y_i\right)}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{S_{X,Y}}{S_X^2} \qquad [I.3.34]$$

$$a = \bar{y} - b\bar{x} \ . \qquad [I.3.35]$$

An example. In his book 'Library Effectiveness : A Systems Approach' p.90, Morse (1968), considers the following table (Table I.3.9).

Table I.3.9. Mean second-year circulation N(m) for those books that in their first year had a circulation m

m	0	1	2	3	4	5	6	7	8	9	10
N(m)	0.4	0.7	1.2	1.3	2.2	2.4	2.5	3.7	3.8	4.5	5.1
Estimated	0.42	0.82	1.22	1.62	2.02	2.42	2.82	3.22	3.62	4.02	5.62

Applying equations [I.3.34] and [I.3.35] on these data results in :

$$b = 0.400 \quad \text{and} \quad a = 0.418 \ ,$$

with $\Sigma x_i = 58$, $\Sigma y_i = 278$, $\Sigma x_i y_i = 205.9$, $\Sigma x_i^2 = 454$ and $\Sigma y_i^2 = 95.02$.
Thus, the equation of the least-squares line is, using [I.3.32] :

$$y = 0.418 + 0.4\, x$$

Moreover, the Pearson correlation coefficient R is 0.979. This result is illustrated in Fig.I.3.4.

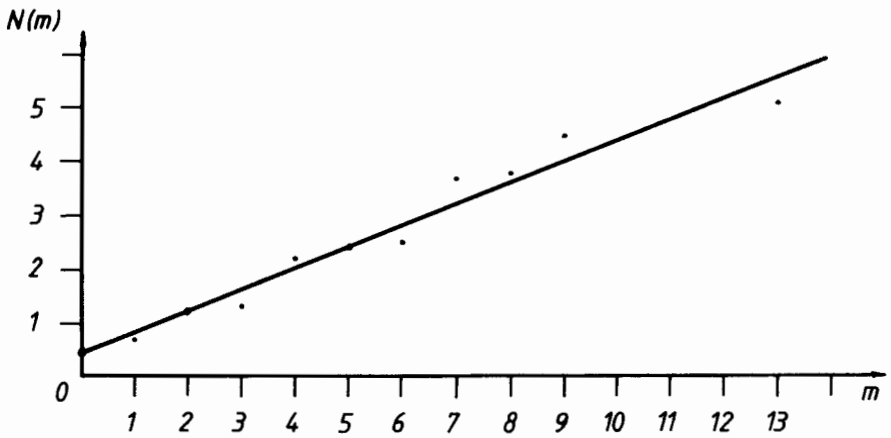

Fig.I.3.4 Linear regression line for Morse's circulation data (Table I.3.9)

If we let y_{est} denote the estimated value of y for a given x, as obtained from the regression curve of y on x, then one can show that

$$R^2 = 1 - \frac{\sum_i (y_i - y_{i,est})^2}{\sum_i (y_i - \bar{y})^2} \quad . \qquad [I.3.36]$$

Now, one also has :

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - y_{i,est})^2 + \sum_i (y_{i,est} - \bar{y})^2 \quad , \qquad [I.3.37]$$

where the left-hand side is called the '*total variation*', the first sum on the right is called the '*unexplained variation*', and the second sum is called the '*explained variation*'. This terminology has arisen because the deviations $y_i - y_{i,est}$ behave in an unpredictable manner, while the deviations $y_{i,est} - \bar{y}$ are explained by the least-square regression and so tend to follow a definite pattern.

Substituting [I.3.37] in [I.3.36] yields :

$$R^2 = \frac{\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - y_{i,est})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (y_{i,est} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

$$= \frac{\text{explained variation}}{\text{total variation}} \qquad [I.3.38]$$

Thus R^2 can be interpreted as the fraction of the total variation which is explained by the least-squares regression line. It is often referred to as the 'coefficient of determination'.

Regression relies on a lot of assumptions. If these assumptions are not met, inferences made from a regression are invalid - at least from a theoretical point of view - although the regression equation may still be of value in describing the relationship between two variables. For cases where x = t(time) the regression line indicates a trend, making (careful) predictions possible. This situation provides a simple example of time series analysis.

I.3.8.5. Pearson's product moment correlation coefficient as a measure of fit

A regression line can be calculated for any scatterplot, but what we actually need is a measure of how well the regression line fits the data and a statistical test on this measure.

As a measure to indicate a linear relationship, Pearson's product-moment correlation coefficient is used. The test statistic (for N which is large enough, i.e. N > 10) is

$$R\sqrt{\frac{N-2}{1-R^2}}\ , \qquad [I.3.39]$$

which can be shown to be (approximately) t-distributed with N-2 degrees of freedom. For N which is small (N < 10) one uses Table A.7 of critical values.

The null hypothesis is $H_0 : R = 0$,

$H_1 : R \neq 0$ (for a two-sided test).

In the previous example we found R = 0.979, so that $t(9) = 0.979\sqrt{\frac{9}{1-(0.979)^2}} = 14.4$, leading to the rejection of the hypothesis that there is no linear correlation (on any reasonable level). One is led to the same conclusion when using tables of critical correlation coefficients.

We note that when N is large (as is often the case in sociometric studies using questionnaires), H_0 is already rejected for rather small values of R.

For example, when R = 0.3, N = 125, $t(123) = R\sqrt{\frac{123}{1-R^2}} = 3.49$. For a two-sided test on the 1 % level the acceptance region for the hypothesis R = 0 is]-2.61,+2.61[(see Table A.2), leading to the rejection of H_0 in a very significant way.

Other test statistics.

1) To test the hypothesis that the regression coefficient b is equal to a specified value β, one uses the stochastic variable

$$\frac{\beta - b}{s_{y,x}/s_x}\sqrt{N-2} \quad , \qquad \text{[I.3.40]}$$

which is t-distributed with N-2 degrees of freedom.
This can also be used to find confidence intervals for population regression coefficients.

2) To test the hypothesis that the regression line passes through the origin (H_0 : a = 0), we use the statistic

$$a\sqrt{\frac{N(N-2)\sum_i (\bar{x}_i - x)^2}{(\sum_i x_i^2)(\sum_i (y_i - y_{i,est})^2)}} \qquad \text{[I.3.41]}$$

which is t-distributed with N-2 degrees of freedom.

Pearson's correlation coefficient has been used in many informetric papers. We mention only a few of them : Etnyre and Kretlow (1975), Zusne (1976), Williams (1978), Tomer (1986), Nelson (1988).

I.3.8.6. Spearman rank correlation

The *Spearman rank correlation coefficient* is a nonparametric measure of the relationship between two sets of ranked data. It is widely applied as an alternative to Pearson's correlation coefficient with its restricting assumptions. The Spearman test is usually applied to ordinal data, but can also be applied to other data if they are converted to a ranked form.

The equation for the Spearman rank correlation coefficient is :

$$R_s = 1 - \frac{6\sum_i d_i^2}{N(N^2-1)} \quad , \qquad \text{[I.3.42]}$$

where R_s denotes the Spearman rank correlation coefficient, d_i is the difference in ranking for the i^{th} item and N is the number of items studied. Like the product-moment correlation coefficient, the Spearman coefficient can have

a value between -1, indicating perfect negative correlation between the two sets of rankings, and +1, indicating perfect positive correlation. A value of 0 indicates an absence of correlation.

In Rousseau (1988a) the following table of the impact factors of mathematical journals is presented (Table I.3.10). (The notion of an impact factor will be explained in Chapter III.5).

Table I.3.10. Ordering of mathematical journals according to their 2-year and 4-year impact factor (1985). Data based on the JCR (see Part III for more information on this data source)

A Journal, abbreviated as in the JCR
B Ordering according to the 2-year impact factor
C Ordering according to the 4-year impact factor
D Absolute value of difference in rank (d_i)

A	B	C	D
COMMUN ALGEBRA	1	3	2
P K NED AKAD A MATH	2	14	12
DISCRETE MATH	3	8.5	5.5
NAGOYA MATH J	4	12	8
MATH SCAND	5	8.5	3.5
B SCI MATH	6	7	1
J MATH SOC JPN	7.5	5	2.5
P AM MATH SOC	7.5	13	5.5
B SOC MATH FR	9	1	8
J NUMBER THEORY	10.5	6	3.5
Q J MATH	10.5	2	8.5
ANN SCI ECOLE NORM S	12	11	1
MATH USSR SB	13	15	2
CAN J MATH	14	10	4
STUD MATH	15	4	11

Note that an average rank is used for ties. As $\sum_i d_i^2 = 582.5$, the Spearman rank correlation coefficient of these data is :

$$R_s = 1 - \frac{(6)(582.5)}{(15)(224)} = -0.04 \ .$$

The test in this situation resembles the one used for Pearson's correlation coefficient. We take

H_0 : there is no correlation between both rankings, i.e. $R_s = 0$;

H_1 : $R_s \neq 0$.

For $N > 10$ we actually use the same test as for Pearson's coefficient : the stochastic variable

$$R_s \sqrt{\frac{N-2}{1-R_s^2}} \qquad [I.3.43]$$

is t-distributed with (N-2) degrees of freedom. For smaller values of N special tables of critical values must be consulted.

For the example of impact factors t(13) = -0.14. For a test on the 5 % level the acceptance region for H_0 is]-2.16,+2.16[. Therefore, we accept the null hypothesis that both rankings are uncorrelated.

The Spearman rank correlation coefficient is used, for example, by Tomer (1986), He and Pao (1986), Nelson (1988). Bear in mind that in the case in which non-ordinal measurements are converted to rankings, there is bound to be some loss of information (roughly 10 %). There is also no reason to expect that rank correlation should produce the same result as product-moment correlation when the two techniques are applied to a common data set. Rank correlation may well be a more reliable measure in many instances, since it does not depend on any, possibly unwarranted, assumptions about the frequency distribution of the variable. See in this respect also Brookes and Griffiths (1978).

I.3.8.7. Kendall's tau (cf Kendall (1970), Hájek (1969))

The stochastic variable τ (tau) has the same purpose as Spearman's rank correlation coefficient, namely to investigate the relationship between ordered data. The use of *Kendall's tau* as a testing procedure is as powerful as Spearman's R_s.

Peritz (1986) presents the following table (Table I.3.11).

Table I.3.11. Journals on demography and family

A Journal
B Number of times cited according to Journal Citation Reports (1983)
C Rank according to B
D Number of papers in Population Index 1984
E Rank according to D

A	B	C	D	E
1 Journal of Marriage and the Family	1793	1	20	10
2 Demography	548	2	47	1
3 Family Planning Perspectives	523	3	23	7.5
4 Population Studies	454	4	25	6
5 Studies in Family Planning	266	5	27	5
6 Journal of Biosocial Science	262	6	36	4
7 Social Biology	248	7	11	12.5
8 Population and Development Review	233	8	43	2
9 Population	209	9	39	3
10 International Migration Review	146	10	23	7.5
11 Population Bulletin	104	11	6	15
12 Journal of Family History	87	12	0	18
13 Population Index	49	13	4	16
14 International Journal of Sociology of the Family	30	14	0	18
15 International Migration	29	15	12	11
16 Demografia	22	16.5	11	12.5
17 Journal of Family Welfare	22	16.5	22	9
18 Population and Environment	15	18	8	14
19 Population Research and Policy Review	4	19	0	18

Kendall's tau is defined as :

$$\tau = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \operatorname{sgn}(R_i - R_j) \operatorname{sgn}(Q_i - Q_j) = \frac{2S}{N(N-1)} . \qquad [I.3.44]$$

In this case the R's denote rankings in the first list, Q's rankings in the second list, while sgn(x) is the sign of x and is +1 if x is positive and -1 if x is negative.

Minimum and maximum values for τ are -1 and +1. One can show that τ is approximately normal. If both lists are independent, $E(\tau) = 0$ and

$$\operatorname{Var}(\tau) = \frac{2(2N+5)}{9N(N-1)} . \qquad [I.3.45]$$

For Peritz' table we find the underlying table (Table I.3.12) :

Table I.3.12. Calculation of Kendall's tau for Peritz' data (based on Table I.3.11)

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	-	-	-	-	-	+	-	-	-	+	+	+	+	-	+	-	+	+
2		0	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
3			0	-	-	-	+	-	-	0	+	+	+	+	+	+	+	+	+
4				0	-	-	+	-	-	+	+	+	+	+	+	+	+	+	+
5					0	-	+	-	-	+	+	+	+	+	+	+	+	+	+
6						0	+	-	-	+	+	+	+	+	+	+	+	+	+
7							0	-	-	-	+	+	+	+	-	0	-	+	+
8								0	+	+	+	+	+	+	+	+	+	+	+
9									0	+	+	+	+	+	+	+	+	+	+
10										0	+	+	+	+	+	+	+	+	+
11											0	+	+	+	-	-	-	-	+
12												0	-	0	-	-	-	-	0
13													0	+	-	-	-	-	+
14														0	-	-	-	-	0
15															0	-	-	+	+
16																0	-	+	+
17																	0	+	+
18																		0	+
19																			0

This yields 119 + signs and 49 - signs, and hence $\tau = 0.392$; its variance is 0.0279. Then

$$z = \frac{0.392}{\sqrt{0.0279}} = 2.345 \ .$$

On the 5 % level the acceptance region for a one-sided test (one could also do a two-sided test) is $]-\infty, 1.645[$. Thus we reject the null hypothesis that both lists are independent.

Kendall's tau (or better, the related value S) offers some practical advantages above Spearman's rank order correlation coefficient. Indeed, when processing questionnaires it often happens that some respondents turn in their answers very late. If R_s has already been calculated at that moment and one wishes to include the late answers as well, one has to compute R_s all over again. This is not the case for S : if a late answer is included in the ranking, one only has to consider +1's and -1's (as described above) for couples (i,j), where i or j is the latecomer, and add this to the previous value of S. This is only a small part of the total work (and in fact, if this reply had not been so late, this would have been done already).

Kendall's tau is also used in Daniłowicz and Szarski (1981).

I.4. SAMPLING THEORY

In the above chapter we used samples to perform statistical tests. However, we did not tell how to draw a sample or how to find the minimum sample size necessary to estimate a population characteristic within specified confidence limits. These questions will be the main topic of this section. Special attention will be paid to typical problems occurring in library and information science. Before starting off the study of sampling we remind the reader of the (obvious) fact that a sample does not have to be taken if the population is not too large : just check the property under investigation on every element of the population.

I.4.1. Classical sampling disciplines

The most important pitfall when drawing samples is the introduction of *bias*, i.e. giving relatively smaller attention to relatively larger classes in the population and vice-versa. This may result in an unreliable estimate of the true (but unknown) population characteristic. For example, if one wants to find out how many times a year citizens visit the local public library, it would be wrong to hand out questionnaires only to persons entering the library, as people who never visit the library will certainly not be included in the sample!

I.4.1.1. Random sampling : the technique

In order to avoid bias in sampling one must make sure that every element in the population has an equal chance of being included. This type of sampling is called '*random sampling*'. Random sampling is what one should always try to achieve.

In practice one uses a table of so-called '*random numbers*' or (directly) the output of a computer random generator. In recent years, these random number generators have been built into personal computers and even into most programmable calculators.

For the readers' convenience we reproduce a small part of such a table here (Table I.4.1). A larger table can be found in the appendix (Table A.9).

Table I.4.1. Random numbers

...
72682
21443
... 01176 ...
80582
13177
21785
47458
40405
... 71209 ...
85561
...

Suppose we want to take a sample of 50 persons from a population of 5000. To do this, the persons have to be numbered from 1 to 5000. Then we start anywhere in the table (e.g. where you see the number 72682). Since the enumeration uses four digits, we select only numbers consisting of four digits. Here we choose 7268. But 7268 > 5000, so that there is no person corresponding to this number. This means that we have to make another choice. We are free to move in the table in any direction we want. Suppose we decide to go downwards. The next number is 2144, meaning that person 2144 becomes an element of our sample. The next number is 117, 8058 is rejected, and then we get 1317 and so on, until we have a sample of size fifty.

I.4.1.2. Random sampling : drawbacks and remarks on the method (cf. Bookstein (1974))

a) The main drawback of this method is that it is tedious, if only for the reason that every element in the population must be numbered. This is usually not a big problem in computer files, but it definitely is when sampling from card files or, more importantly, from book shelves. It is precisely for this reason that other sampling methods are considered, at the cost of introducing a bias. One tries to find sampling techniques that are easy to carry out, that can be done quickly and that approximate as closely as possible the results obtained by random sampling (see below).

b) In numbering the population elements one must ensure that no element is numbered twice (do not mix co-authors or author and subject files), or that no elements are left unnumbered (books that are on loan perhaps?).

c) The following procedure is not allowed. Suppose we sample from a population of size 850. If we encounter 000 or a number larger than 850, we do not toss this number out. Instead, we skip only the first digit and add the

next digit from the table of random numbers. Instead of 855 we will take, say, 556 (cf last line of Table I.4.1). This is not a correct procedure for it increases the probability of numbers between 510 and 850 and 1 to 9 being chosen (and hence decreases this probability for numbers between 10 and 509).

d) In random sampling we often have to throw out a lot of random numbers. Suppose, for instance, that the population consists of only 300 elements. We will still need 3-digit random numbers, which results in a 70 % loss of time and effort. This can be partially remedied by ignoring the numbers between 901 and 999 (10 % of the choices) and than dividing every other random number by 3 and using the first natural number greater than or equal to this quotient. No bias will occur.

e) Random numbers generated by most random number generators are not really random. Indeed, as any sequence of numbers produced by an algorithm must be deterministic, it eventually repeats earlier values. Therefore, numbers obtained in this way are called '*pseudo-random numbers*'. Mainframe computers have usually been provided with subroutines libraries for generating high-quality pseudo-random numbers, which can also be converted to samples from other, non-uniform, distributions. Personal computers and calculators, however, are often supplied with a random number generator which may be totally inadequate for serious scientific work (although perhaps good enough for some small tests in a library). For a technical account of random number generators, the reader is referred to Knuth (1981). For background on the use of random number generators for small computers and a note of warning, one may consult Ripley (1983) and Fullerton (1987).

I.4.1.3. Random permutations

We will discuss in this section a special random sampling technique. Suppose we wish to study the activity in k sections of a large library. To avoid bias we visit these k sections every day in a different order. This means that we need *random permutations*, i.e. a random arrangement of the k areas. Although such permutations can be obtained from sampling in random number tables, dedicated tables of random permutations are much easier to work with (see e.g. Moses and Oakford (1963)).

Jain (1972) presents a plan for sampling in-library use in which the investigator goes through the book shelves in different areas of the library. The sequence in which these areas are surveyed is determined by using random permutations.

I.4.1.4. Systematic sampling

a) A technique used often to avoid the tedious task of random sampling is called '*systematic sampling*'. We will explain this sampling technique in terms of book shelves, although it can be applied generally.

We begin with any arbitrary book and then take whatever book lies 30 cm further along the shelf. The third book is again 30 cm further along the shelf, so that we are picking a book every 30 cm to compose the sample. Naturally, 30 cm is an arbitrary number, as one can take whatever distance one wants. This is also related with the sample size; (see further).

An obvious bias inherent in this method is that thick books have a bigger chance of being chosen than thin ones. Problems also occur at the end of the shelves or when books are leaning against each other.

b) A variant of the above method consists of changing length elements (30 cm) into counting (e.g. every 30 books) or time elements (e.g. every 30 minutes). So we take an arbitrary book and obtain the sample by taking every 30th book. Although we have eliminated the most obvious bias of the above method, the present method is taking much more time. Imagine a sample in the whole library : one would have to count every book! Still, it is faster than random sampling.

c) When the time variant (checking some library activity, say, every 30 minutes) is used, some bias may be introduced. When studying a school library it is not a good idea to sample every 30 minutes when classes take an hour (yielding short peaks of activity). Also, in the case of quality-checking of a printing job, it might be that a printing error appears every 30 copies. Checking response times of online services yields another example in which systematic sampling is not always unbiased. It might be that at fixed time periode average response times are larger than at other times.

An almost complete solution to these problems will be given in Section I.4.2.

I.4.1.5. Stratified random sampling

This type of sampling is not really a new method. One applies random sampling (or any other good sampling method) to different sections of the population, but the contributions of the sections to the sample are fixed in advance.

Suppose we investigate the use of a local public library by adults. This results in contingency table I.4.2.

Table I.4.2. Library users

		High school degree		
		yes	no	totals
library use	yes	200	10	210
	no	200	90	290
totals		400	100	500

This sample shows that 42 % of the adults are library users. We also see that 80 % of the persons in the sample have a high school degree. However, suppose we know from census data that in this particular village only 60 % of the adults have a high school degree. In view of this knowledge we can take the following actions :

1. We can modify the above results, so that we end up with 60 % high school degrees in the sample. This yields the following revised table (Table I.4.3).

Table I.4.3. Library users - revised table

		High school degree		
		yes	no	totals
library use	yes	150	20	170
	no	150	180	330
totals		300	200	500

This revised table results in only 34 % of library users, a more reliable result. There was an obvious overrepresentation in the sample of persons with a high school degree.

2. In case we have not yet performed the sampling, we can sample persons with and without a high school degree separately, in such a way that population proportions are respected.

A bias can occur, especially in small samples, and it is precisely in these cases that stratified random sampling can help. This technique is particularly useful in situations where the property under investigation is homogeneous within groups and heterogeneous between groups (Lied and Tolliver (1974)).

I.4.2. The Fussler sampling technique

In an attempt to combine speed and randomness, Fussler (1961) introduced

the following technique : use systematic sampling by length, but then select for the sample the k^{th} book (k stays fixed and is preferably small; one can even take k equal to one) behind the one located by systematic sampling. This method has the same advantages as sampling by length (one does not have to number hundreds of books, using large random tables), and as we will show, is at least as good and usually better.

The quality of the *Fussler sampling technique* has been investigated by Bookstein (1983), Rousseau (1988b) and Egghe (1988c). Since we consider this technique to be very important in informetrics and since it is easy to use, we will study this technique in greater detail.

I.4.2.1. The idealised situation consisting of two separate categories

A. Bookstein uses the model of sampling in a card file, in which in an (idealised) situation the cards can only have two possible thicknesses. Everything depends on the way the 'thin' and 'thick' cards are clustered. If we denote a thin card by t and a thick card by T, we can measure the degree of clustering by counting the number of groups of t's and T's. For instance, the clustering ttTTTTtTTtttttTTttTt has five groups of consecutive t's and four groups of consecutive T's, and therefore a total of nine runs (cf. Subsection I.3.7.2 or Section I.2.5). The distribution of these runs can be shown (see Subsection I.2.5.2 and especially Fig.I.2.2) to be approximately normal, as indicated in Fig.I.4.1.

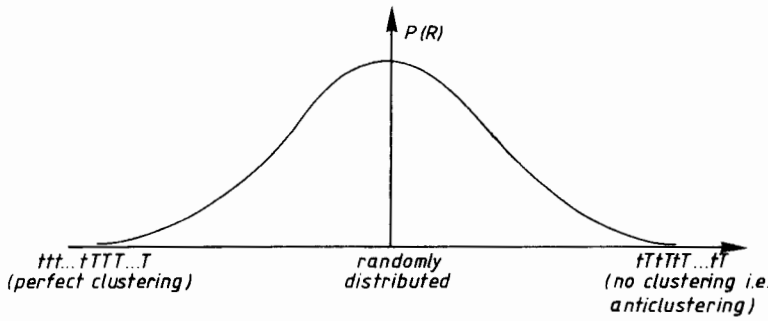

Fig.I.4.1 Probability distribution of runs of thick and thin cards

Bookstein (1983) studied only the left-hand side of the curve in Fig. I.4.1 and showed in this case that, no matter how the t's and T's are clustered, there is less bias with Fussler's sampling technique than with sampling by length. Indeed, if P_1 denotes the probability of picking a thin card at random, P_2 the probability of picking a thin card by using sampling by length and P_3 the same probability in the case of sampling by Fussler's procedure (using k = 1,2,3,...; say k = 1), Bookstein has shown that for clusters belonging to the left-hand side of Fig.I.4.1, we will always have

$$P_2 \leq P_3 \leq P_1 \quad .$$

Of course, the inequalities will be reversed for sampling thick cards. The Fussler technique is therefore always closer to the random sampling technique than sampling by length is.

The right-hand side of Fig.I.4.1, however, has an equal chance of occurring as the left-hand side. As will be shown further on, here it is even true that $P_2 < P_1 < P_3$. However, the following inequality holds for all types of clustering of thin and thick cards :

$$|P_1 - P_3| \leq P_1 - P_2 \quad , \qquad [I.4.1]$$

showing that Fussler's technique is never worse than sampling by length. In particular, for the most common cases (the middle part of Fig.I.4.1) $P_1 \approx P_3$. See Fig.I.4.2.

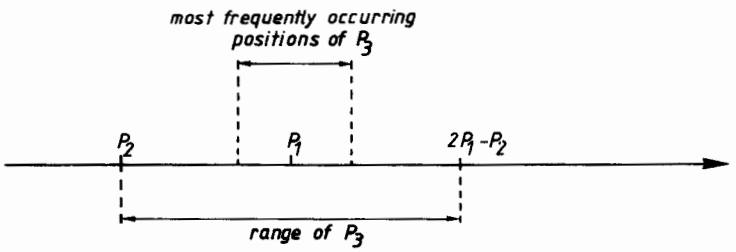

Fig.I.4.2 Illustration of the inequality $|P_1 - P_3| \leq P_1 - P_2$; showing that Fussler's technique is always better than sampling by length

Furthermore, Rousseau (1988b) uses the terminology of thin and thick books on a library shelf. Note that if we denote $1 - P_i$ by P_i' ($i = 1,2,3$), P_i' denotes the probability of sampling thick books. For these probabilities [I.4.1] directly yields :

$$|P_1' - P_3'| \leq P_2' - P_1' \quad . \qquad [I.4.2]$$

B. Proof of the inequality : $|P_1 - P_3| \leq P_1 - P_2$ (which includes also the proof of the case studied by Bookstein).

We will assume that the collection is made up of n_t thin books, having an average thickness w_t, and of n_f thick books, having an average thickness w_f. We define $r_1 = n_f/n_t$ (where we assume $n_t \neq 0$) and $r_2 = w_f/w_t$.

In the case of random sampling books are picked at random. The probability that the book will be a thin one is given by :

$$P_1 = \frac{n_t}{n_t + n_f} = \frac{1}{1 + r_1} \quad . \qquad [I.4.3]$$

The probability P_1 is what we are trying to achieve by using other techniques.

In the case of sampling by length each physical location has an equal chance of being selected. Thus the probability that a thin book will be chosen is given by :

$$P_2 = \frac{n_t w_t}{n_t w_t + n_f w_f} = \frac{1}{1 + r_1 r_2} \quad . \qquad [I.4.4]$$

As $r_2 \geq 1$, we note that $P_2 \leq P_1$.

Let P_3 be the probability that a thin book will be chosen by Fussler's method. Let P_t denote the probability that the book situated k books after a thin one will also be a thin one, and let P_f be the probability that the book situated k books after a thick one will be thin. Note that the probabilities P_t and P_f, being conditional probabilities do not add up to one unless the number of thick books equals the number of thin ones. Then

$$P_3 = P_2 P_t + (1 - P_2) P_f = \frac{P_t + r_1 r_2 P_f}{1 + r_1 r_2} \quad . \qquad [I.4.5]$$

Further, $n_t = n_t P_t + n_f P_f$ (neglecting end effects) or $P_f = (1 - P_t)/r_1$. Substituting this value for P_f in [I.4.5] yields :

$$P_3 = \frac{P_t (1 - r_2) + r_2}{1 + r_1 r_2} \quad . \qquad [I.4.6]$$

So, P_3 is not equal to P_1 but depends on P_t, which indicates the degree to which thin books are clustered together. As $0 \leq P_t \leq 1$ and $r_2 \geq 1$, we also see that $P_3 \geq P_2$. Bookstein (1981) also points out that if there is no clustering, i.e. $P_t = 1/(1+r_1)$, then $P_3 = P_1$ and Fussler's method is completely successful. On the other hand, if books of one type cluster together and $P_t = 1$, then $P_3 = P_2$. In that case nothing is gained with respect to sampling by length.

To show the inequality announced under B., we first prove that $1 - P_t \leq r_1$.

Let r_t be the number of thin books which are at a 'distance' k books from a thin book. Then $P_t = r_t/n_t$. Those thin books which are not at a 'distance' k books from a thin book are necessarily situated at a distance k books from a thick book (again neglecting end effects). So there are at least $n_t - r_t$ thick books. This yields : $n_t - r_t \leq n_f$ or $1 - P_t \leq r_1$.

Proof of the inequality showing that the Fussler method is at least as good as sampling by length.

We recall that $P_2 \leq P_1$ and that $P_2 \leq P_3$. If now $P_2 \leq P_1 \leq P_3$, then clearly $|P_1 - P_3| \leq |P_1 - P_2|$. If $P_2 \leq P_1 \leq P_3$, then

$$|P_1 - P_3| \leq |P_1 - P_2|$$

$$\Longleftrightarrow$$

$$\frac{(1-r_2)P_t + r_2}{1 + r_1 r_2} - \frac{1}{1+r_1} \leq \frac{1}{1+r_1} - \frac{1}{1+r_1 r_2}$$

$$\Longleftrightarrow$$

$$((1-r_2)P_t + r_2)(1+r_1) - (1+r_1 r_2) \leq (1+r_1 r_2) - (1+r_1)$$

$$\Longleftrightarrow$$

$$1 - P_t - P_t r_1 \leq r_1 \ ,$$

which is true according to the above remark. □

The probability P_3 can be expressed as a linear combination of the probabilities P_1 and P_2, as follows :

$$P_3 = (1-\Theta)P_2 + \Theta P_1 \ ,$$

where

$$\Theta = \frac{(1+r_1)(1-P_t)}{r_1} .$$

For a proof we refer the reader to Rousseau (1988b).

One final remark : problems with end effects can be eliminated by working modulo N, where N is the total number of elements in the population. In practical terms, this means considering the first book as the one following immediately after the last one.

I.4.2.2. Fussler sampling in book shelves : the case of a discrete distribution of thicknesses (Egghe (1988c))

The above-mentioned results do not deal with the realistic situation of more than two types of thicknesses on a book shelf. In fact, in reality we have a finite number of possible thicknesses of books, say (in increasing order) :

$$d_1 < d_2 < \ldots < d_n .$$

Denote by $P_1(d_j)$: the probability of picking a book with thickness d_j, using random sampling.
Denote by $P_2(d_j)$: the probability of picking a book with thickness d_j, using sampling by length.
Denote by $P_3(d_j)$: the probability of picking a book with thickness d_j, using the Fussler sampling technique (i.e. sampling by length as above and taking the k^{th} book after it, $k \in \{1,2,3,\ldots\}$ arbitrary but fixed).

The following theorem shows that the Fussler sampling procedure is always better than sampling by length, no matter how the books of different thicknesses are clustered on the shelf. Note that, in view of inequalities [I.4.1] and [I.4.2], we need the absolute value signs on both sides of the inequality. Based on implications of the theorem, we recommend the Fussler sampling technique for most practical cases.

Theorem (Egghe (1988c)). For every $j = 1,\ldots,n$ one has :

$$|P_1(d_j) - P_3(d_j)| \leq |P_1(d_j) - P_2(d_j)| . \qquad [I.4.7]$$

Proof. Determine any $j = 1,\ldots,n$. In the set $\{d_j,d_{j+1},\ldots,d_n\}$ the books with thickness d_j are thin and the books with thickness greater than d_j are thick. Hence inequality [I.4.1] yields (now using conditional expectations in $\{d_j,\ldots,d_n\}$) :

$$|P_1(d_j|\{d_j,\ldots,d_n\}) - P_3(d_j|\{d_j,\ldots,d_n\})|$$
$$\leq P_1(d_j|\{d_j,\ldots,d_n\}) - P_2(d_j|\{d_j,\ldots,d_n\}) .$$

Using the definition of conditional expectations, we thus find :

$$\left|\frac{P_1(d_j)}{P_1(\{d_1,\ldots,d_n\})} - \frac{P_3(d_j)}{P_3(\{d_j,\ldots,d_n\})}\right| \leq \frac{P_1(d_j)}{P_1(\{d_j,\ldots,d_n\})} - \frac{P_2(d_j)}{P_2(\{d_j,\ldots,d_n\})}$$

or

$$\left|\frac{P_1(d_j)}{\sum_{\ell=j}^{n} P_1(d_\ell)} - \frac{P_3(d_j)}{\sum_{\ell=j}^{n} P_3(d_\ell)}\right| \leq \frac{P_1(d_j)}{\sum_{\ell=j}^{n} P_1(d_\ell)} - \frac{P_2(d_j)}{\sum_{\ell=j}^{n} P_2(d_\ell)} . \qquad [I.4.8]$$

Likewise, books with thickness d_j are thick in the range of books with thickness $\{d_1,\ldots,d_j\}$. So, by using inequality [I.4.2] we now have :

$$|P_1(d_j|\{d_1,\ldots,d_j\}) - P_3(d_j|\{d_1,\ldots,d_j\})|$$
$$\leq P_2(d_j|\{d_1,\ldots,d_j\}) - P_1(d_j|\{d_1,\ldots,d_j\}) .$$

As above, we find :

$$\left|\frac{P_1(d_j)}{\sum_{\ell=1}^{j} P_1(d_\ell)} - \frac{P_3(d_j)}{\sum_{\ell=1}^{j} P_3(d_\ell)}\right| \leq \frac{P_2(d_j)}{\sum_{\ell=1}^{j} P_2(d_\ell)} - \frac{P_1(d_j)}{\sum_{\ell=1}^{j} P_1(d_\ell)} . \qquad [I.4.9]$$

To simplify further calculations, we adopt some new notations. We take

$$\alpha_i = \sum_{\ell=j}^{n} P_i(d_\ell) \qquad (i = 1,2,3)$$

and :

$$a_j = P_i(d_j) \qquad (i = 1,2,3)$$

(since j is fixed we do not mention the index j in a_i and α_i).
So, in this new notation formulae [I.4.8] and [I.4.9] become :

$$\left|\frac{a_1}{\alpha_1} - \frac{a_3}{\alpha_3}\right| \leq \frac{a_1}{\alpha_1} - \frac{a_2}{\alpha_2}$$

and

$$\left|\frac{a_1}{a_1 + 1 - \alpha_1} - \frac{a_3}{a_3 + 1 - \alpha_3}\right| \leq \frac{a_2}{a_2 + 1 - \alpha_2} - \frac{a_1}{a_1 + 1 - \alpha_1} \quad .$$

From these inequalities, it follows that :

$$|a_1\alpha_3 - a_3\alpha_1| \leq a_1\alpha_3 - a_2 \frac{\alpha_1\alpha_3}{\alpha_2}$$

and

$$|a_1(1-\alpha_3) - a_3(1-\alpha_1)| \leq a_2 \frac{(a_1 + 1 - \alpha_1)(a_3 + 1 - \alpha_3)}{a_2 + 1 - \alpha_2} - a_1(a_3 + 1 - \alpha_3) \ .$$

Hence, using a triangular inequality

$$|P_1(d_j) - P_3(d_j)|$$

$$= |a_1 - a_3|$$

$$\leq |a_1\alpha_3 - a_3\alpha_1| + |a_1(1-\alpha_3) - a_3(1-\alpha_1)|$$

$$\leq a_2 \left\{\frac{(a_1 + 1 - \alpha_1)(a_3 + 1 - \alpha_3)}{a_2 + 1 - \alpha_2} - \frac{\alpha_1\alpha_3}{\alpha_2}\right\} - a_1(a_3 + 1 - 2\,\alpha_3)$$

$$= P_2(d_j) \frac{(P_1(d_j) + 1 - \sum_{\ell=j}^{n} P_1(d_\ell))((P_3(d_j) + 1 - \sum_{\ell=j}^{n} P_3(d_\ell))}{P_2(d_j) + 1 - \sum_{\ell=j}^{n} P_2(d_\ell)}$$

$$- \frac{\sum_{\ell=j}^{n} P_1(d_\ell) \sum_{\ell=j}^{n} P_3(d_\ell)}{\sum_{\ell=j}^{n} P_2(d_\ell)} - P_1(d_j)\,[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)]. \quad [I.4.10]$$

We now accept a second order approximation :

$$P_2(d_j)\ P_1(d_{j'}) \approx P_2(d_j)\ P_2(d_{j'})$$

for all $j,j' = 1,\ldots,n$ (since $P_2(d_j)$ is small). Now inequality [I.4.10] becomes :

$$|P_1(d_j) - P_3(d_j)|$$

$$\leq P_2(d_j)\,[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)]$$

$$- P_1(d_j)\,[P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell)] \quad . \qquad [I.4.11]$$

We take

$$\alpha = P_3(d_j) + 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell) \; .$$

Then [I.4.11] reads :

$$|P_1(d_j) - P_3(d_j)| \leq \alpha(P_2(d_j) - P_1(d_j)) \; . \qquad I.4.12)$$

Now :

$$\alpha \begin{cases} = 1 - P_3(d_j) - 2 \sum_{\ell=j+1}^{n} P_3(d_\ell) & \text{if } j < n \\ = 1 - P_3(d_j) & \text{if } j = n \end{cases}$$

≤ 1 in all cases . [I.4.13]

Furthermore, since

$$1 = \sum_{\ell=1}^{n} P_3(d_\ell) \geq \sum_{\ell=j}^{n} P_3(d_\ell)$$

we have :

$$\alpha \geq 1 - 2 \sum_{\ell=j}^{n} P_3(d_\ell) \geq -1 \qquad [I.4.14]$$

in all cases.
From [I.4.13] and [I.4.14] it now follows that

$$|\alpha| \leq 1 \qquad [I.4.15]$$

in all cases. Inequalities [I.4.12] and [I.4.15] then imply :

$$|P_1(d_j) - P_3(d_j)| \leq |P_1(d_j) - P_2(d_j)|$$

for every $j = 1,\ldots,n$. □

Remarks :

1. From the definition of α we see that if d_j is small (the thinner books), $\alpha \approx -1$. For these books we have, using inequality [I.4.12] :

$$|P_1(d_j) - P_3(d_j)| \leq P_1(d_j) - P_2(d_j) .$$

If d_j is large (the thicker books), $\alpha \approx +1$ and hence, using [I.4.12] :

$$|P_1(d_j) - P_3(d_j)| \leq P_2(d_j) - P_1(d_j) .$$

2. The quantity $|P_1(d_j) - P_2(d_j)|$ is large for small or large d_j's. Obviously, for average values of d_j this difference is small. But no matter how large $|P_1(d_j) - P_2(d_j)|$ is, inequality [I.4.7] of the theorem is valid and it might be that, when books of thickness d_j are randomly distributed amongst the other books (which is likely to be the case in book shelves), $P_1(d_j) \approx P_3(d_j)$ even when $|P_1(d_j) - P_2(d_j)|$ is large. So the Fussler sampling technique's strongest impact is in eliminating the largest bias (for d_j small or d_j large) encountered when sampling by length.

3. The sign of $P_1(d_j) - P_3(d_j)$ depends on the degree of clustering of the books with thickness d_j.

I.4.2.3. Fussler sampling in the case of a continuous distribution function

The Fussler sampling technique can also be applied in the case of sampling from a continuous distribution (e.g. a continuous distribution of time periods). A typical case is the following. To find the distribution of checkout times in a library one might take a random sample from the population of all users (situation 1); one might also sample by time, e.g. every 10 minutes (situation 2). This method is biased towards the cases requiring a longer service time. In situation 3 one samples every 10 minutes, but the next borrower is taken. This is Fussler's technique with k = 1.

Let t_m denote the maximal checkout time and let $t_0, t_1 \in [0, t_m]$, $t_0 < t_1$. For i = 1,2,3 denote the probability of picking a borrower with a checkout time in the interval $[t_0, t_1]$ by $P_i[t_0, t_1]$, where the sampling method is as described in situation i above.

Theorem (Egghe (1988c)). For every $t_0, t_1 \in [0, t_m]$, $t_0 < t_1$, one has :

$$|P_1[t_0,t_1] - P_3[t_0,t_1]| \leq |P_1[t_0,t_1] - P_2[t_0,t_1]| . \qquad [I.4.16]$$

Proof. The proof follows the lines of the previous theorem. We observe in

this case that the times in $[t_0,t_1]$ are short with respect to the time interval $[t_0,t_m]$ and that the times in $[t_0,t_1]$ are long with respect to the time interval $[0,t_1]$. □

A similar inequality can be proved for density functions of continuous probabilities (Egghe (1988c)).

Another application of a continuous situation can be found by returning to the situation described in Bookstein (1983). Indeed, as mentioned by Buckland, Hindle and Walker (1975), different tensions in various parts of a card drawer, furthermore dependent on time, can result in a continuously changing 'real' thickness (including the air between the cards).

The general conclusion is that all these uncontrollable physical aspects do not bother us : the Fussler sampling procedure is as quick and simple as sampling by length (or time) but gives less bias, and in fact is reduced in most common cases to random sampling.

I.4.3. Overlap

I.4.3.1. Statement of the problem

We consider the collections of two libraries A and B and study the overlap between their book or journal collections. To establish the idea, we will consider the case of book titles. This situation is depicted schematically in the Venn diagram of Fig.I.4.3.

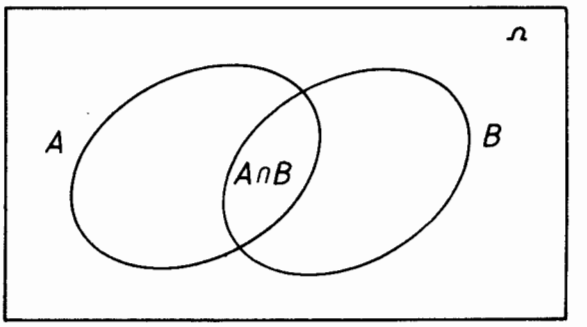

Fig.I.4.3 Venn diagram of book titles of libraries A and B. Ω denotes the set of all book titles

Studying *overlap* in book titles between libraries A and B means finding answers to at least one of the following questions :

- What is the set $A \cap B$?
- Determine $\frac{\#(A \cap B)}{\#B}$ and $\frac{\#(A \cap B)}{\#A}$, i.e. the proportion of books in A that are

also in B and vice versa.

Note that, using the notation of probability theory (Subsection I.2.1.2), $\frac{\#(A \cap B)}{\#B} = \frac{P(A \cap B)}{P(B)} = P(A|B)$, and similarly $\frac{\#(A \cap B)}{\#A} = P(B|A)$. Of course also it is interesting to know $P(\bar{A}|B)$ and $P(\bar{B}|A)$ as well, where $\bar{A} = B \setminus A$ and $\bar{B} = A \setminus B$, but these numbers follow from the preceding ones from the equation

$$P(A|B) + P(\bar{A}|B) = 1 \ .$$

When drafting a union catalogue, the study of overlap between several libraries becomes important. If n libraries $A_1, A_2, \ldots, A_n$ want to join forces, the following questions will have to be answered :

- How many titles are held by exactly 1, exactly 2, ..., exactly n libraries?
- Given that a title is in library A_i, what is the probability that it will be in none, one, two, ... of the others.

I.4.3.2. The importance of knowing overlap

Whether the overlap between collections is large or small, it is always important to find out what its extent is. A few practical examples are given below.

Collaboration between two (or more) libraries with respect to automation is much more economical when library collections have a big overlap.

When a union catalogue is produced, there are two conflicting viewpoints concerning overlap. If the aim is to cover (as much as possible) all libraries of a certain region (country, state), a big overlap is desirable for economical reasons. In this case, the union catalogue will not be too large, with respect to the catalogues of the most important libraries, so that printing costs are reduced. On the other hand, if the aim is to cover as many titles as possible, then a small overlap is preferred. Many smaller libraries having an overlap of, for example, more than 90 % with the larger libraries might then be excluded from the union catalogue, thus saving a lot of holding indicators.

The lower the overlap between two online files is, the more important it is to search both. Conversely, if the overlap between two online files is big, and if money is a problem, a savings of about 50 % can be reached by only searching one file (and still obtaining, say, 80 % of the total recall that one would have obtained when searching both files). In a recent study on forensic medicine (Snow and Ifshin (1984)) the overlap among MEDLINE and EMBASE on a number of sample questions was 30 % overall. However, in conjunction with the analysis of uniqueness (references found in only one of the databases studied, the others being BIOSIS, SCISEARCH and CASEARCH), it was found that

if MEDLINE or EMBASE were omitted from the search on forensic medicine topics, nearly one-third of the total recall would be lost.

The relation between overlap and IL (Interlibrary Lending) is rather complicated. Suppose that library B is used by library A for interlending purposes. If the overlap between A and B is small, this seems favourable for A : it has access to a lot of material that it does not own itself. However, if A is special library or a small scientific research library, highly specialised in one topic, it is important to find a larger library that has much more, on this specialised topic as well. In this case the overlap of the second library with respect to the first may well be nearly 100 %!

I.4.3.3. Some practical considerations

I.4.3.2.1. General aspects

Buckland et al. (1975) discuss several pitfalls of sampling in catalogues or external lists (e.g. national bibliographies) for measuring overlap. Summarising, we can say that the main problems are the inconsistency between the classification rules of different libraries and the lack of sufficient external lists. Therefore, these authors recommend adopting a direct statistical approach : pick a random sample from A (a library say, or an online file) and check this sample against the holdings of B (a different library or online file). The fraction of the sample from A which is also in B is then taken as an estimate $\bar{x}$ for the overall proportion of A which is in B (i.e. $P(B|A)$). The actual number of items held in common is then - approximately - found by multiplying by the total number of items in A.

I.4.3.3.2. Overlap and the binomial distribution

It can be assumed that the number of items from a sample in A, also found in B, is a binomial random variable. Indeed, every item in A has a probability p of also being in B (so we consider the experiment of picking one item in A and verifying whether it actually belongs to the intersection of A and B). This is a Bernoulli experiment with parameter p, cf. Subsection I.2.4.1). For this Bernoulli experiment the average μ is p and its variance is p(1-p). For N Bernoulli trials (sample of size N), the sampling mean $\bar{X}$ is normally distributed with parameters p and $\sigma^2/N = p(1-p)/N$ (where N is large). Hence $\bar{X} \sim N(p, p(1-p)/N)$.

In actual sampling, the variance of $\bar{X}$ is unknown, so that we will use $\bar{x}(1-\bar{x})/(N-1)$ (cf. Subsections I.3.1 , I.3.4.3 and I.3.4.4). In this way a 95 % confidence interval is given by

$$[\bar{x} - 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{N-1}} \; , \; \bar{x} + 1.96\sqrt{\frac{\bar{x}(1-\bar{x})}{N-1}}] \quad , \qquad [I.4.17]$$

where $\bar{x}$ is the observed fraction in the sample. As noted earlier, this interval for $P(B|A)$ also produces an interval for $P(\bar{B}|A) = 1 - P(B|A)$.

I.4.3.3.3. The case of several libraries

Suppose we have a group of libraries. How much of the material is duplicated, triplicated and so on? The obvious way to proceed is to take a stratified sample (according to library size) and then check this sample against the holdings of each library. This would yield an estimate on the number of items held by 1,2,3,...,n libraries. This procedure, however, introduces a bias. Let p be the probability of including a particular book in the sample. This p is constant for every library since we use a stratified sample. Then 1-p is the probability of not drawing this book in one particular library. As we draw an independent sample in every library, the probability that a particular title will not be in the sample when this title belongs to the holdings of k libraries is $(1-p)^k$. Hence, this title has a probability of $1 - (1-p)^k$ of being included in the sample. For example, when p = 0.01 and k = 5, this title has a probability of 0.049 (instead of 0.01) of being in the sample (see also Table I.4.4).

Table I.4.4. Probabilities that titles belonging to k libraries will belong to a stratified sample if p = 0.01 (see the text for the meaning of p)

k	$1 - (1-p)^k$
1	0.010
2	0.020
3	0.030
4	0.039
5	0.049
6	0.059
7	0.068
8	0.077
9	0.086
10	0.096

To correct this bias, it is sufficient to multiply the number of books found in k libraries by $1/(1 - (1-p)^k)$ and, if one wishes, to normalise the numbers.

I.4.4. Sample size

In Subsections I.3.4.4 and I.4.3.3.2 we briefly discussed how to construct a confidence interval around estimates made from samples. Obtaining from this the minimum sample size necessary to estimate, for example, the population mean within specified confidence limits,is a fairly straightforward calculation.

I.4.4.1. Tests on the mean

In Subsection I.3.4.4 we found the following 95 % confidence interval for the population mean μ (large sample, i.e. $N \geq 30$; σ known) :

$$[\bar{x} - 1.96 \frac{\sigma}{\sqrt{N}} , \quad \bar{x} + 1.96 \frac{\sigma}{\sqrt{N}}] \quad .$$

Suppose now that we specify in advance that the length of this interval must be at most L. This yields the following inequality :

$$2 \, (1.96 \frac{\sigma}{\sqrt{N}}) \leq L$$

or

$$3.92 \, \sigma \leq \sqrt{N} \, L$$

or

$$N \geq \frac{(3.92)^2 \, \sigma^2}{L^2} \quad .$$

If $\sigma = 36$, $L = 10$, this equation results in $N \geq 199.15$, indicating that we need a sample size of 200 or larger.

In many cases, however, it is more natural to specify a *relative error* on the mean, expressing the maximum length of a confidence interval as a fraction of $\bar{x}$, say $\beta\bar{x}$, where β is usually 0.1 or 0.2.

We then have to solve the following inequality :

$$3.92 \frac{\sigma}{\sqrt{N}} \leq \beta\bar{x}$$

or

$$N \geq \frac{(3.92)^2 \, \sigma^2}{\beta^2 \bar{x}^2} \quad . \qquad [I.4.18]$$

This leads, however, to a kind of circular argument : the sample size N needed to find $\bar{x}$ is given as a function of $\bar{x}$!

The way to deal with this problem is to simply draw two samples. First, we draw a provisional, small sample. This gives us a rough estimate for $\bar{x}$ (and for s^2 if the variance is unknown). This first estimate is used to determine N from [I.4.18]. We include the provisional sample in the final larger sample so as not to waste time or energy in drawing the first sample. This type of sampling tactic is called '*two stage sampling*'. Note that we have explained how to determine the sample size in one particular case. We trust that the reader will be able to apply the above reasoning to other cases, based on the other formulae in Subsection I.3.4.1.

I.4.4.2. Tests on fractions

A 95 % confidence interval for fractions ($\bar{x}$) is given by

$$[\bar{x} - 1.96 \frac{\bar{x}(1-\bar{x})}{\sqrt{N-1}}, \bar{x} + 1.96 \frac{\bar{x}(1-\bar{x})}{\sqrt{N-1}}]$$

(cf [I.4.17]).
Hence a confidence interval of length $\beta\bar{x}$ requires a sample size N at least equal to :

$$\frac{(3.92)^2 \bar{x}(1-\bar{x})}{\beta^2\bar{x}^2} + 1 = \frac{(3.92)^2 (1-\bar{x})}{\beta^2\bar{x}} + 1 \qquad [I.4.19]$$

Again, two-stage sampling is necessary.

An example. Suppose we wish to have a 95 % confidence interval of length $\bar{x}/10$ for the fraction $\bar{x}$ of overlap between two libraries. A provisional sample of size 100 yields an overlap of 60 % ($\bar{x} = 0.6$). We then need a sample of size $n = \frac{(3.92)^2 (0.4)}{(0.1)^2(0.6)} + 1 = 1026$.

This formula has also been applied to estimate the number of lost books in a large library (Miller and Sorum (1977)); see also Goldstein and Sedransk (1977). For more general information on sampling the reader is advised to consult, for example Cochran (1963).