

STATISTICAL METHODS
in **MEDICINE** *and* **PUBLIC HEALTH**

MASTER *in* PUBLIC HEALTH METHODOLOGY

**Medical Statistics Lab.
Public Health School
Faculty of Medicine
ULB**

Willy SERNICLAES

September 2003

TABLE OF CONTENTS

1. Basic Concepts	p.1
<ul style="list-style-type: none">• data and theories• sample and population• variation	
2. Sample description	p.7
<ul style="list-style-type: none">• variable types• mode, median, mean• pie chart, bar chart, histogram, cumulative polygon• variance, SD, skewness	
3. Probability	p.27
<ul style="list-style-type: none">• Probability• Poisson distribution• Binomial distribution• Normal distribution• ROC curves	
4. Confidence Intervals	p.51
<ul style="list-style-type: none">• Sampling distribution• Confidence interval for a mean• Confidence interval for a count• Confidence interval for a proportion	
5. Conformity tests for a single sample	p.61
<ul style="list-style-type: none">• Null hypothesis, Significance test• False positives and Confidence• t-test for a mean• False negatives and Power• Chi-square test for a count• Chi-square test for a proportion	

6. Univariate test significance tests for two or several samples	p.71
<ul style="list-style-type: none"> • t-test and ANOVA for two means • t-test and ANOVA for several means • contrasts for means • chi-square test for two proportions • chi-square test for two proportions • chi-square test for several proportions • non-Parametric tests 	
7. Univariate regression	p.85
<ul style="list-style-type: none"> • taxonomy of bivariate relationships • Linear regression • Non-parametric tests • Logistic regression • contrasts for proportions 	
8. Sample design and size	p.107
<ul style="list-style-type: none"> • Designs • Specification of sample size 	
9. Generalized Linear Model	p.139
<ul style="list-style-type: none"> • Generalized Linear model • Contrasts • Stratification • Bias (confounder) & Interaction (effect modification) • Multivariate model • Guidelines for the choice of a test • Strategies for model building • TABLES 	
10. Multiple factor ANOVA	p.159
<ul style="list-style-type: none"> • generalized factorial ANOVA • analysis of covariance • repeated measurements 	
11. Multiple Linear regression	p.169
12. Multiple Logistic regression	p.175

Practicals

p. 187

Solutions

p. 216

SPSS Examples

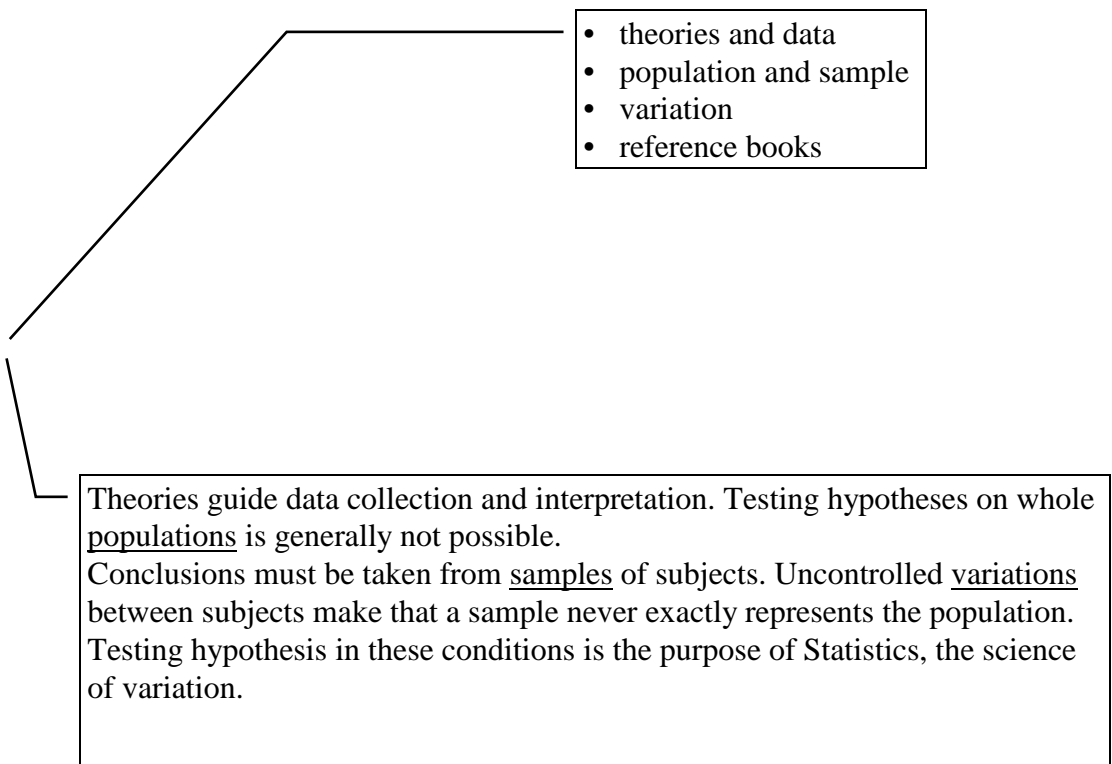
After p. 298

MODULE PLANNING

<u>MODULE</u>	<u>CONTENT</u>	<u>EXERCISES</u>
# 1	Chpt.1 + Chpt.2 [variable types; functions]	Exs. 1 to 3
# 2	Chpt.2 [central tendency parameters for categorical & ordinal variables, polygon]	Ex. 4
# 3	Chpt.2 [central tendency parameters for quantitative variables; histogram]	Exs. 5 & 6
# 4	Chpt.2 [dispersion parameters for quantitative & binary variables]	Exs. 7 & 8
# 5	Chpt.3 [probability, definition and rules]	Exs. 9 to 11
# 6	Chpt.3 [Binomial & Poisson distributions]	Exs. 12 & 13
# 7	Chpt.3 [Normal distribution]	Exs. 14 & 15
# 8	<i>Summary exercises on Probability</i>	Exs.* to ****
# 9	<i>Summary exercises on Probability</i>	
# 10	<i>Summary exercises on Probability</i>	
# 11	Chpt.3 [ROC curves]	Ex. 16
# 12	Chpt.4 [sampling distribution & confidence interval for a mean]	Ex. 17
# 13	Chpt.4 [confidence interval for a mean with Student's distribution]	Ex. 18
# 14	Chpt.4 [confidence interval for a count & for a proportion]	Exs. 19 & 20
# 15	<i>Summary exercises on Probability Distributions and Confidence Intervals</i>	Exs. 21 to 25
# 16	<i>Summary exercises on Probability Distributions and Confidence Intervals</i>	Exs. 26 to 29
# 17	<i>Starting Practicals with SPSS</i>	Ex. 30
# 18	<i>Starting Practicals with EPI-INFO</i>	Ex. 31
# 19	Chpt.5 [significance test; conformity test for mean, count & proportion]	Ex. 32
# 20	Chpt.6 [t-test & ANOVA for 2 means]	Exs. 33 & 34
# 21	Chpt.10** [Paired t-test]	Ex. 54-(1)
# 22	Chpt.6 [ANOVA & K-Wallis test for several means]	Ex. 35
# 23	Chpt.6 [contrasts between means]	Ex. 36
# 24	Chpt.6 [Chi-square test & Fisher exact test for proportions; contrasts]	Exs. 37 to 39
# 25	Chpt.12** [Mc Nemar test]	Ex. 59
# 26	Chpt.7 [taxonomy of bivariate relationships; linear regression; correlation; test]	Exs. 40 to 42
# 27	Chpt.7 [logistic regression with quantitative predictor]	Exs. 43 & 44
# 28	Chpt.7 [logistic regression with categorical predictor; contrast & Odds Ratio]	Ex. 45
# 29	Chpt.7 [contrasts for several proportions]	Ex. 46
# 30	REVISION	
# 31	<i>Summary exercises on univariate statistical tests</i>	Ex. 47
# 32	<i>Summary exercises on univariate statistical tests</i>	Ex. 48
# 33	<i>Summary exercises on univariate statistical tests</i>	
# 34	Chpt.8 [sampling methods; SRS, syst. sampling, stratified sampling, cluster sampling]	Ex. 49
# 35	Chpt.8 [specification of sample size]	Ex. 50

# 36	Chpt.9 [Generalized Linear Model]-Chpt.10 [multifactorial ANOVA]	Ex. 51
# 37	Chpt.10 [multifactorial ANOVA; Analysis of Covariance]	Exs. 52 & 53
# 38	Chpt.10 [Repeated Measures ANOVA]	Ex.54-(2)
# 39	Chpt.11 [Multiple Linear Regression]	Exs. 55 & 56
# 40	Chpt.12 [Multilogistic Regression]	Ex. 57
# 41	Chpt.12 [Multilogistic Regression; model building]	Exs. 58 & 59
# 42	<i>Summary exercises on multivariate statistical tests</i>	Ex. 60 & 61
# 43	<i>Summary exercises on multivariate statistical tests</i>	Ex. 61
# 44	Chpt.13 [Survival Analysis]	Exs. 62 & 63
# 45	Chpt.13 [Survival Analysis]	Exs. 64 & 65

Chapter 1. Basic Concepts

- 
- theories and data
 - population and sample
 - variation
 - reference books

Theories guide data collection and interpretation. Testing hypotheses on whole populations is generally not possible. Conclusions must be taken from samples of subjects. Uncontrolled variations between subjects make that a sample never exactly represents the population. Testing hypothesis in these conditions is the purpose of Statistics, the science of variation.

• Data and theories

Science needs data but data alone are by no means sufficient. Interpretation should be foresighted before collecting data, by choosing a model, or a **theory**. In its simplest form, a theory is simply the list of factors which might affect the variable which will be measured. These factors should then be controlled in the study. This will have the advantage of increasing the performances of statistical tests.

Models do not only give an advantage in statistical testing. They also guide research. In Popper's words "Theories are nets: only the one who throws, will fish" (Popper¹). This does not mean that the causal factor is within the model. Causal factors have however a better chance to be captured when research is guided by theory. Simply because the latter gives a framework for systematic collections of data. This is illustrated by the following example.

Exempla : Signing theory (Paracelse, Swiss alchemist and physician, XVIth century). According to Schwartz², this theory states that: "...God, being sorry for having created diseases, should have given to man the plants allowing to combat them, by affecting a recognition sign to each." (p.41, my translation).

This is illustrated in Fig.1 (Schwartz, p.42).

Paracelse, the man on the picture, said that the lungwort (left)with its white stains evoking the color of broncho-pulmonar diseases expectorations was a good remedy; that the nut (above) which imitates the brain hemispheres was good for this organ; that ginseng roots which look like thighs were aphrodisiacs; that colchicin (left), a remedy for gout, is extracted from colchic of which the bulb has the same form as a big toe.(Schwartz, p.42).

Schwartz goes on by stating that "This model makes us smile today, but, chance helping, signing theory led to an important discovery: as the willows were growing with their feet in water, they should contain remedies against fever and rheumatism; one looked into their bark and found salicylic acid, from which aspirin is derived" (p.41).

The issue of signing theory illustrates quite well the practical interest of models. They give a framework for progressing, although their truth is only provisional. According to Popper, science goes not by trying to verify theories but rather by trying to **falsify** theories. To falsify, or **reject**, hypotheses is also at the core of statistical methodology.

• Sample, population and variation

Imagine that a new drug is supposed to cure every subject affected by a given disease. If the drug can be prescribed to every patient the theory will be rejected as soon as a patient is not cured. This is a situation where we do not need statistics to test an hypothesis. However, for different reasons (population size, money, detectability) we can generally not access whole populations. Often you only have a subset of the population or "sample" for testing hypotheses. Now the

¹ Popper, K. (1959) *The Logic of Scientific Discovery*. London: Hutchinson

² Schwartz, D. (1996) "Les modèles en biologie et en médecine" *Pour la Science* 227, 38-45.

problem is that a sample never represents exactly the population. Even if we take care of controlling a set of factors which might affect the outcome of the disease (age, sex, ...), not all the possible determinants can be controlled because some of them are simply unknown. The fact that uncontrolled factors vary from subject to subject makes that the determinants will never be equivalent in two different samples. The consequence of uncontrolled, or intrinsic, variation is that the “cure everybody” hypothesis might be false even if everybody is cured in a sample. It is there that we need statistical inference. Statistics is the science of variation (Fisher³).

Statistical inference allows to quantify the risk taken whenever a false hypothesis is not rejected, as in the above example with the “cure everybody” example. As we shall see this risk is called the “Type II” error (symbol β) and it is quantified with a probability. It is also possible to reject a true hypothesis, as we will see later with other examples. This is the “Type I” error (α or p) which is also quantified with a probability.

• Reference books

Basic statistics:

Statistics in Medicine. T.Colton (1974) Boston: Little Brown Cy.

Essentials of medical statistics. Kirkwood,B. (1998) . Oxford: Blackwell

Nonparametric Statistics for the behavioral sciences. S.Siegel & N.J.Castellan (1988) New York: MCGraw-Hill.

Adequacy of sample size in health studies. D.W.Lemeshow, D.W.Hosmer, J.Klar & S.W.Lwanga (1990) New York: J.Wiley.

Advanced statistics:

Statistical Methods for Medical Investigations. B.S.Everitt (1988) New York: Oxford Univ. Press / London: E. Arnold.

Statistical Methods in Medical Research. F.Armitage (1971) Blackwell Scientific Pub.

Statistical methods for rates and proportions. J.L.Fleiss (1981) New York: J.Wiley.

Statistics. W.Hays (1988). New York: Holt, Rinehart & Wilson.

Logistic Regression. D.G. Kleinbaum. (1994) New York: Springer.

Applied Logistic Regression. D.W. Hosmer & S. Lemeshow. (1989) New York: J.Wiley.

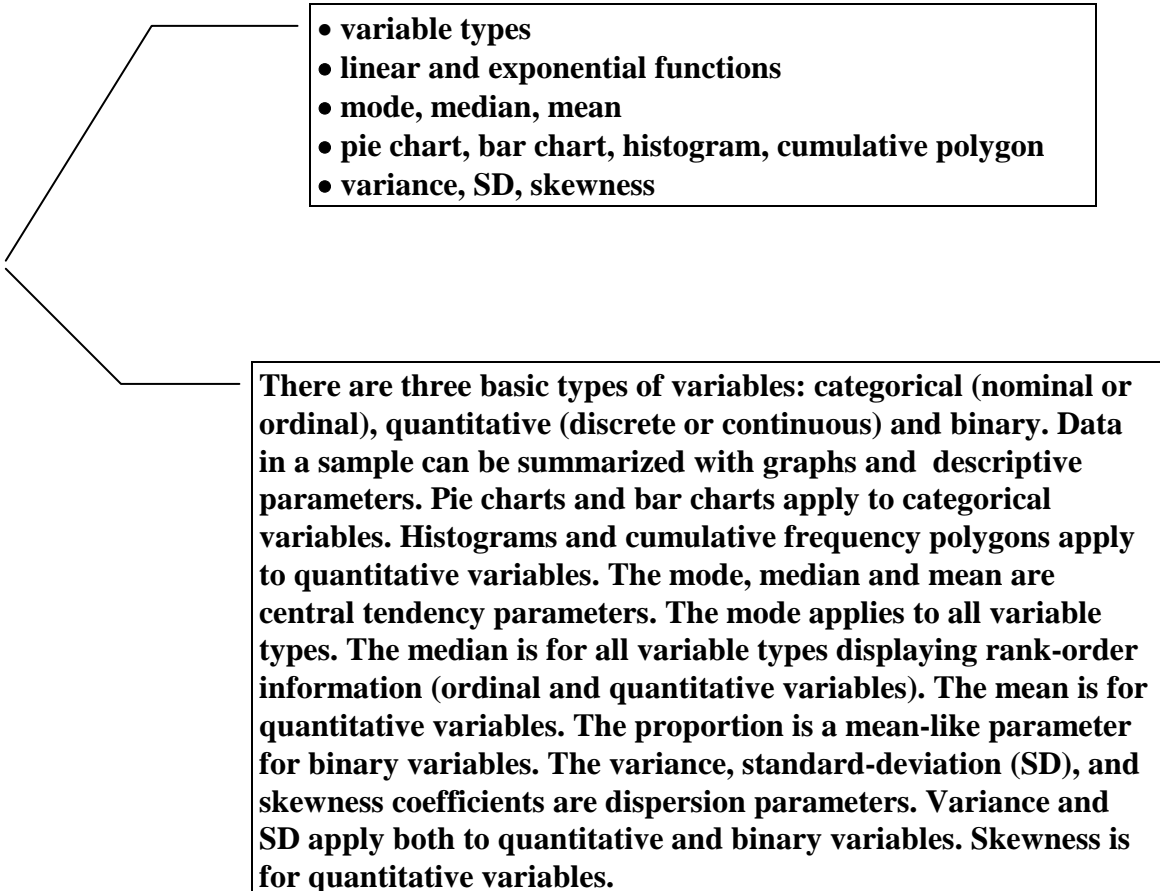
³ Fisher, R.A.(1958). *Statistical methods for research workers*. New York: Hafner.



3. LA THÉORIE DES SIGNATURES FUT UN MODÈLE : selon l'alchimiste et médecin suisse Paracelse (*au centre*), la pulmonaire (*à gauche*), avec ses feuilles aux taches blanches évoquant la couleur des expectorations des maladies broncho-pulmonaires, en constituait un remède de cholx ; la nolx (*en haut à gauche*) était

bonne pour le cerveau, dont elle imite les deux hémisphères ; les racines de ginseng (*en haut à droite*), en formes de cuisses, étaient des aphrodisiaques ; la colchicine, médicamenteusement de la goutte, est extraite du colchique, dont le bulbe a précisément la forme d'un gros orteil (*à droite*).

Chapter 2. Sample description

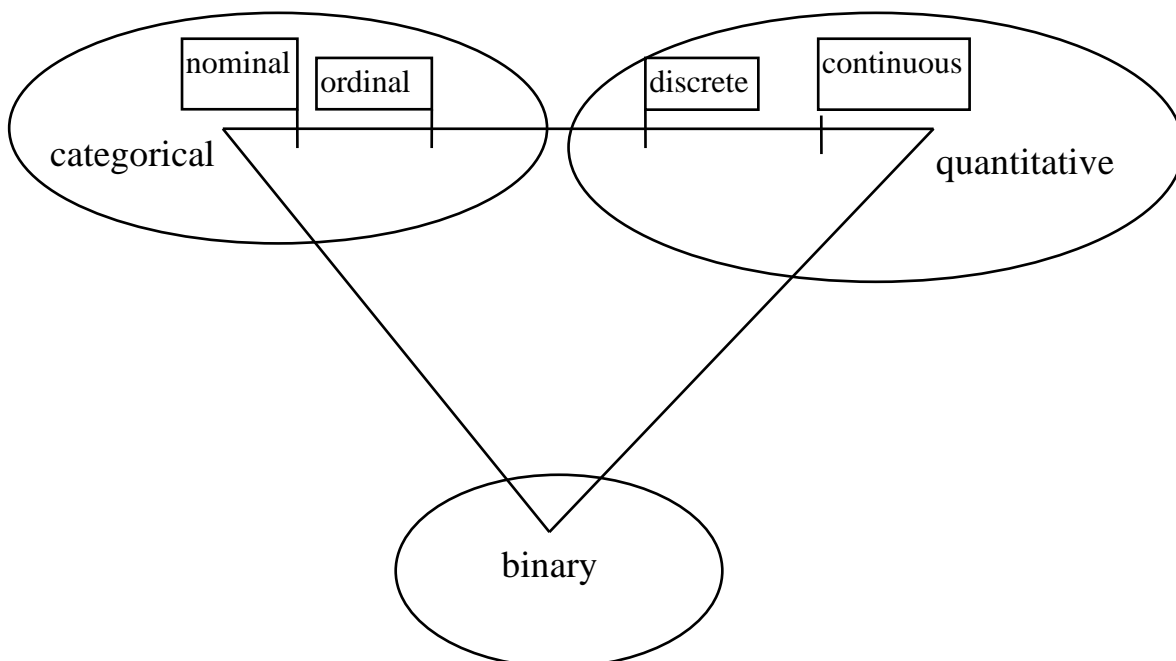
- 
- **variable types**
 - **linear and exponential functions**
 - **mode, median, mean**
 - **pie chart, bar chart, histogram, cumulative polygon**
 - **variance, SD, skewness**

There are three basic types of variables: categorical (nominal or ordinal), quantitative (discrete or continuous) and binary. Data in a sample can be summarized with graphs and descriptive parameters. Pie charts and bar charts apply to categorical variables. Histograms and cumulative frequency polygons apply to quantitative variables. The mode, median and mean are central tendency parameters. The mode applies to all variable types. The median is for all variable types displaying rank-order information (ordinal and quantitative variables). The mean is for quantitative variables. The proportion is a mean-like parameter for binary variables. The variance, standard-deviation (SD), and skewness coefficients are dispersion parameters. Variance and SD apply both to quantitative and binary variables. Skewness is for quantitative variables.

• Variable types

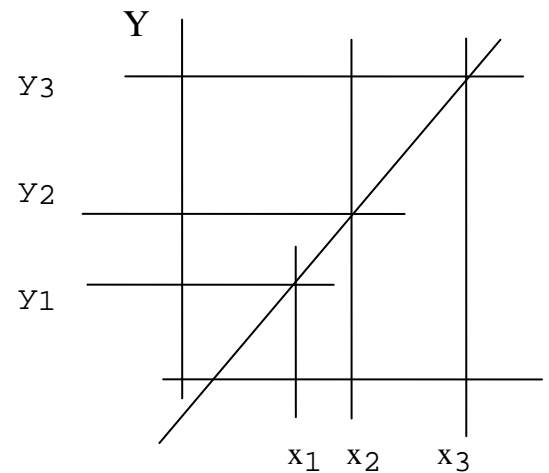
Data are used of capturing different aspects of the individual such as presence versus absence of hearing impairment, kind of impairment, degree of impairment, number of impaired per district, or hearing loss in decibels (dB). These are all different types of variables¹. Variables such as presence versus absence of impairment, vaccinated or not, male or female, can only take two different values. These are **binary** (or “dummy”) variables. Variables such as kind of impairment, bloodgroup, method of delivery, can take several values, each corresponding to a different category. As the categories displayed by these variables cannot be ranked in a definite order, these are **nominal** variables. Variables such as degree of impairment, age group, health-related quality of life (HRQOL),are also categorical. But, unlike nominal variables, they can be ranked in a specific order. These are **ordinal** variables. Variables such as number of impaired per district, number of childbirth per day, number of trypanosomes per blood sample, are quantitative in nature, although not continuous. These are **discrete** variables or “counts”. Finally, variables such as hearing loss in dB, systolic pressure, weight at birth,... are quantitative and **continuous**.

Figure 2 gives variable types differences represented in a triangle. Basic distinctions are between categorical variables (either nominal or ordinal), quantitative variables (either discrete or continuous) and binary variables. As we shall see the latter share both categorical and quantitative properties.

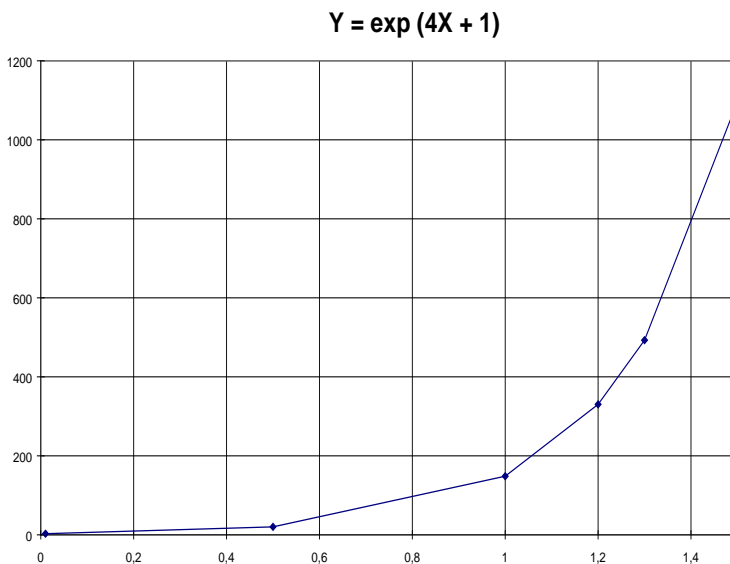


¹ Stevens, S.S.(1946) On the theory of scales of measurement. *Science* 103, 677-680.

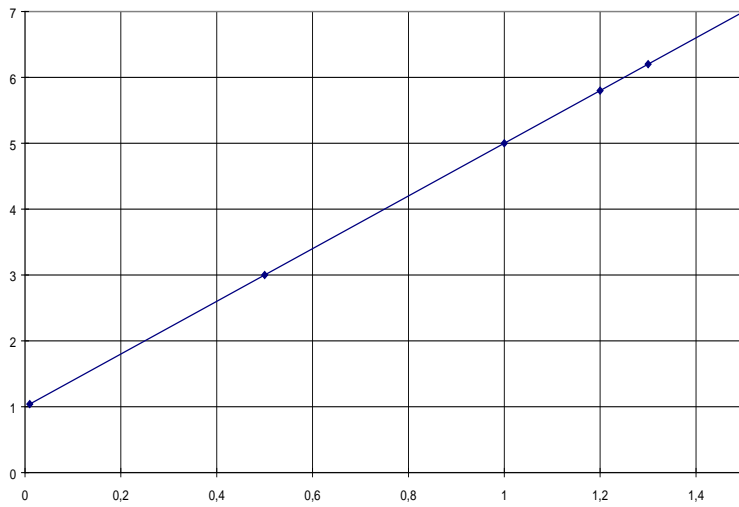
- Linear and exponential functions



One quantitative variable (Y) is a **linear function** of another quantitative variable (X) if any difference in Y divided by the related difference in X is constant. Further, any ratio between two differences in Y is equal to the corresponding ratio in X. The **exponential function** is one of the many possible nonlinear relationship between two quantitative variables. This function is very useful for describing statistical distributions. The exponential function can be **linearized** by taking its **natural logarithm** (LN Y).



$$\text{LN}(Y) = 4X + 1$$



Linear function:

$$y = a + bx$$

b = slope = increase of y for 1 unit

increase of x

a = intercept = value of y when x=0

Exponential functions $y = e^x$ where $e \cong 2.72$

$$y = e^{(a + bx)}$$

Linearization of exponential functions

$$\text{LN}(y) = a + bx$$

Properties of linear functions

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_2}{x_3 - x_2} = \frac{y_3 - y_1}{x_3 - x_1} = b$$

$$\frac{y_2 - y_1}{y_3 - y_2} = \frac{x_2 - x_1}{x_3 - x_2}$$

$$\frac{y_3 - y_1}{y_3 - y_2} = \frac{x_3 - x_1}{x_3 - x_2}$$

$$\frac{y_2 - y_1}{y_3 - y_1} = \frac{x_2 - x_1}{x_3 - x_1}$$

Example of application of linear function: The consumption of drugs in a hospital amounts 500 000 B.F. for 100 patients, 560 000 B.F. for 120 patients and 800 000 B.F. for 200 patients. On these grounds, calculate the consumption for 160 patients ?

N patients	Cost
100	500 kF
120	560 kF
200	800 kF
160	Unknown Cost(U.C.)

Two examples of solutions (among many others)

$$\frac{U.C. - 560}{160 - 120} = \frac{800 - 500}{200 - 100} \quad U.C. = 560 + 300 \cdot 40/100 = 680$$

$$\frac{U.C. - 500}{560 - 500} = \frac{160 - 100}{120 - 100} \quad U.C. = 500 + 60 \cdot 60/20 = 680$$

What about the equation ?

$$b = \frac{800 - 500}{200 - 100} = 3$$

$$\frac{a - 800}{0 - 200} = \frac{500 - 800}{100 - 200} \quad a = 800 + (-200)*(-300)/(-100) = 200$$

$$\text{Cost} = 200 + 3*(\text{nber of patients})$$

• Descriptive parameters

Data in a sample can be summarized by different central tendency and dispersion parameters. Available parameters depend on variable type.

• Central tendency parameters for categorical variables (non-binary): there is only 1 central tendency parameter, called the Mode.

The mode is obtained by comparing the number of data in the different categories. The number of data in a given category is called its **frequency**. The share-out of data in different categories is called a frequency **distribution**. Take the example of language-impairment. The frequencies of different kinds of language impairments are given in Tab.1, together with **relative frequencies** in %. Different graphical representations of category frequencies are possible. Relative frequencies of kinds of language impairment are represented in Fig.3 with an **apple-pie chart** and in Fig.4 with a **bar chart**. In general, some categories are more frequent than others. As can be seen, phonological impairments are most frequent. By definition, the most frequent category is the **mode**.

$$\text{relative } f = f / n$$

$$n = \text{sample size} = \sum_{i=1}^k (f_i)$$

where i represents a numerical *index* varying from 1 (first class) to k (last class)

various impairments	47	13%
stuttering	58	16%
hearing impairment	51	14%
phonological impairment	208	57%
total	364	100%

Table 1. Frequencies of different kinds of language-impairments (from: Woods, Fletcher & Hughes, 1986²; p.9).

² Woods, A., Fletcher, P., and Hughes, A. (1986) *Statistics in Language Studies*. Cambridge: Univ. Press.

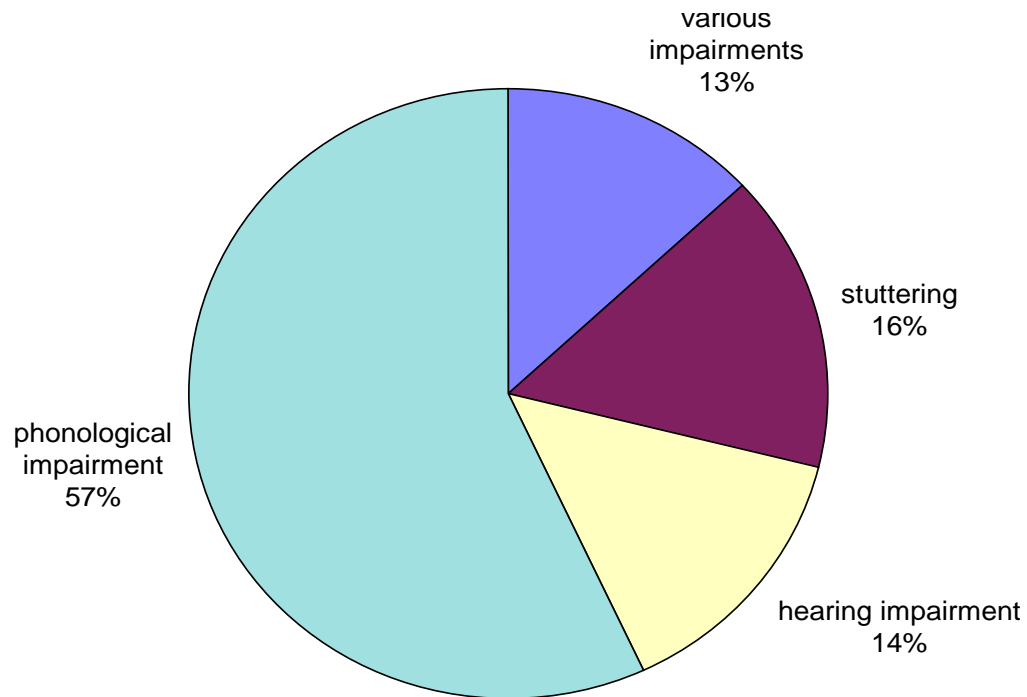


Fig.3 Apple pie chart

Relative frequencies of different kinds of language impairments in a sample of 364 male subjects with language impairments (Woods, Fletcher & Hughes, 1986; p.9).

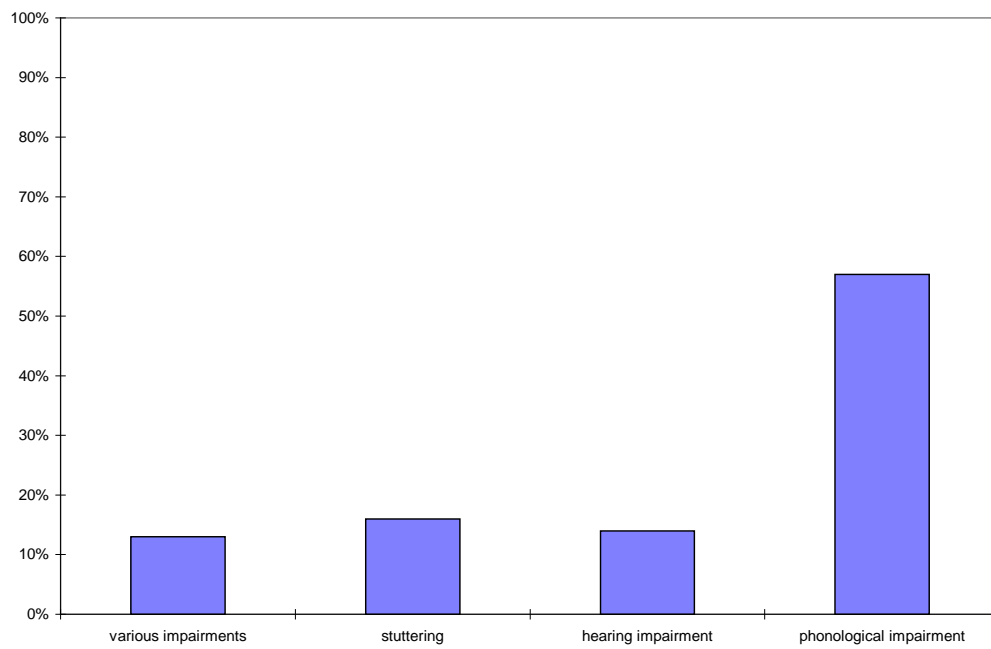


Fig.4 Bar chart

Relative frequencies of different kinds of language impairments in a sample of 364 male subjects with language impairments (Woods, Fletcher & Hughes, 1986; p.9).

• Central tendency parameters for ordinal variables: Mode and Median.

For ordinal variables, the mode can also be used for obtaining the central tendency. But there is another possible parameter based on rank-order. The **median** is value such that 50 % of the (other) values in the sample are lower and 50 % are higher.

The median is part of a family of parameters called **percentiles**.

PERCENTILE 50 = value corresponding to 50% of the cumulative frequencies (rank $n/2$).

Percentile 50 is slightly different from the median (rank $n/2$). For large samples, percentile 50 = median.

PERCENTILE 25 = is the value which leaves 25% of the observations in the sample below.

PERCENTILE 75 leaves 75% of the sample below etc...

Percentiles 25, 50 and 75 are also called respectively first, second and third **quartiles**.

The median and other percentiles are not easily seen in a histogram but are straightforward in a **polygon of cumulative frequencies** (see Fig.6) which relates the upper limit of each class and the sum of the frequencies of the preceding classes.

- Exact formula (for small samples):

$$\text{median} = \text{percentile 50} = \text{value with rank order} = (n + 1) / 2$$

- Approximate formula for percentiles (large samples).

P50 = rank order just below P50 +

$$\frac{50 \% - \text{cum.fr.}(\%) \text{ rank order just below P50}}{\text{cum.fr.}(\%) \text{ rank order just above P50} - \text{cum.fr.}(\%) \text{ rank order just below P50}}$$

Example with a small sample: recovery on a 1 to 7 scale in a sample of 5 patients is 5, 2, 7, 1, 6

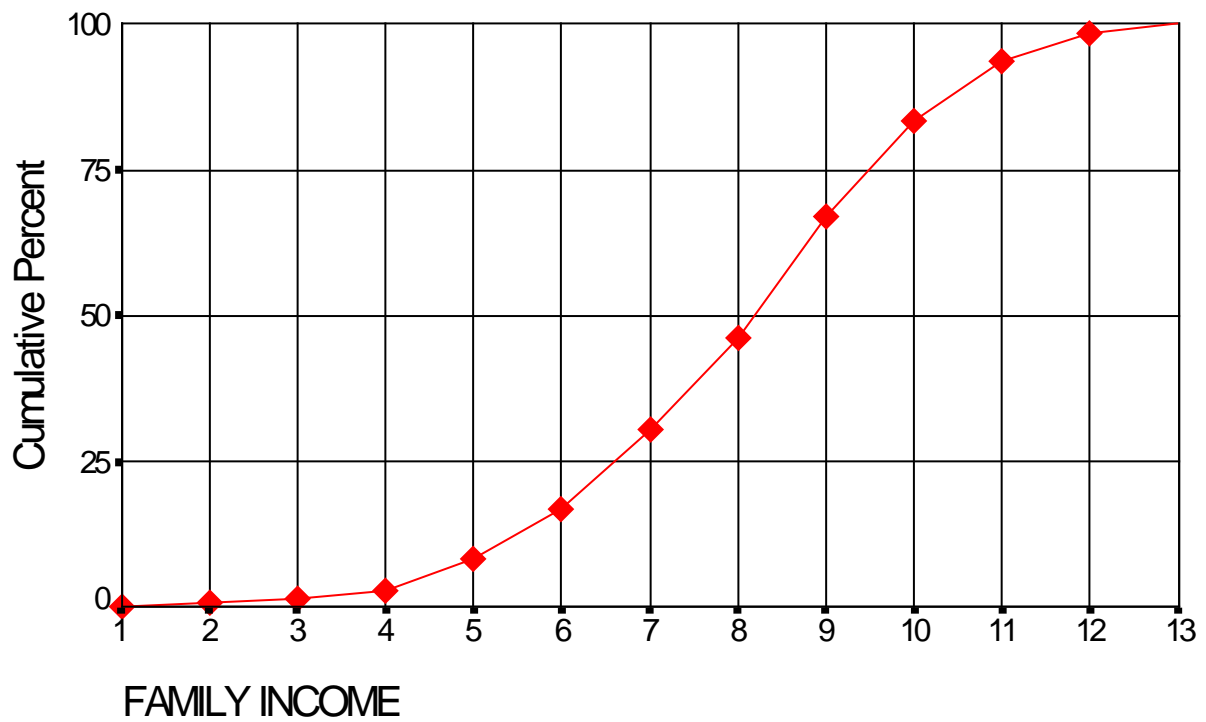
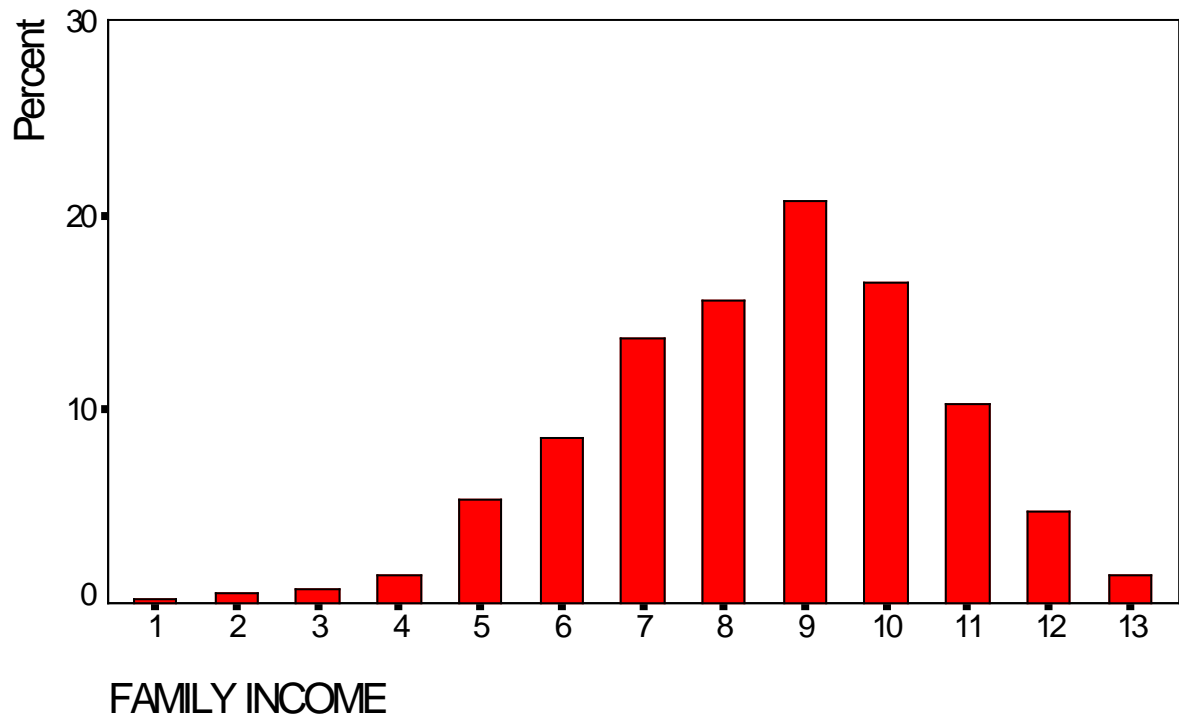
value rank order

1	1
2	2
5	3
6	4
7	5

Median is 5

Example: frequencies of family income in 13 categories (by courtesy of Prof. Lagasse)

Value	Frequ.	Percent	Valid Percent	Cum Percent	
1	2	,3	,3	,3	
2	4	,5	,6	,9	
3	5	,7	,7	1,6	
4	10	1,3	1,4	3,0	
5	37	4,9	5,4	8,4	
6	59	7,9	8,6	17,0	→ $P25 = 6 + (25-17.0)/(30.6-17.0) = 6.59$
7	94	12,5	13,6	30,6	
8	108	14,4	15,7	46,2	→ $P50 = 8 + (50-46.2)/(67.0-46.2) = 8.18$
9	143	19,1	20,7	67,0	→ $P75 = 9 + (75-67.0)/(83.5-67.0) = 9.48$
10	114	15,2	16,5	83,5	
11	71	9,5	10,3	93,8	
12	33	4,4	4,8	98,6	
13	10	1,3	1,4	100,0	
,	60	8,0	Missing		



• **Central tendency parameters for quantitative variables: Mode, Median and Mean.**

Besides mode and median, there is one more central tendency parameter available for quantitative data.

This parameter is the usual arithmetic **mean**, i.e. the sum of the values of the sample divided by the size (symbol n) of the sample. The mean is a meaningful parameter with quantitative data because the 4 arithmetic operations make sense, which is not the case with merely ordinal data. However, the median still has an interest for metric data because it is not affected by deviant (extreme) values. On the contrary the mean is affected by deviant values, especially in small samples.

$$\text{MEAN} = m = \left(\sum_{i=1}^n x_i \right) / n$$

n = sample size; i represents a numerical *index* varying from 1 (first data) to n (last data)

The **histogram** is the most usual graphical representation for quantitative data. The histogram is obtained by first grouping data into classes (see procedure below). The frequency of each class is then represented as a function of its central value (see Fig.5). When data are grouped into classes, the approximate value of the mean can be computed as follows.

$$m = \left(\sum_{c=1}^k f_c x_c \right) / n$$

x_c = central value of class; f_c =frequency of class; k = number of classes

The mode is easy to see in a histogram. For data grouped into classes the mode is the central value of the class with the highest frequency. The interest of the mode is that it gives indications on the homogeneity of the sample. The presence of 2 or several modes, or at least of 2 or several local modes (peaks in the distribution separated by valleys), indicates that several kinds of data have been mixed.

Example: rate of insulin in the umbilical vein for 30 subjects.
Data (in units per cm³)

subject index	units per cc
1	37
2	39
3	40
4	40
5	40
6	28
7	37
8	42
9	27
10	29
11	58
12	36
13	42
14	30
15	21
16	36
17	34
18	53
19	84
20	38
21	43
22	40
23	66
24	36
25	50
26	23
27	56
28	47
29	76
30	36

BUILDING AN HISTOGRAM & A CUMULATIVE FREQUENCY POLYGON

A simpler picture of the data can be obtained by grouping the data into classes. The following strategy can be used in this purpose:

1) calculate the RANGE of values (maximum -minimum = $84 - 21 = 63$).

2) choose the NUMBER OF CLASSES. Guideline values are between 10 and 20. (Take 10 with the small sample used here).

3) define the WIDTH of the class by taking a number just above the following ratio:

$$\text{RANGE} / \text{NUMBER OF CLASSES}$$

(In the example: $63 / 10 = 6.3$; take 7 as class width).

4) define the LIMITS of the classes in such a way as each observation falls into one and only one class. Simplest strategy is to put the limits of the classes between possible values. The CENTRAL VALUE of the class is midway between the limits (e.g. for the lowest class $(20.5 + 27.5)/2 = 24$).

(example: the lower limit of the lowest class is 20.5 which is just below the lowest value of the sample and next limits are 27.5, 34.5 etc...)

5) count the number of data (frequency) per class.

6) The HISTOGRAM is obtained with the classes indicated on the abscissa (by their central values-as in Fig.5 or by their limits) and their frequencies on the ordinate (by bars).

Relative frequencies (in %) can also be used.

7) With the exact formula, the mean = 42,13 units per cc

The mean calculated with the approximate formula for data grouped into classes is:

$$m \cong (3 \cdot 24 + 4 \cdot 31 + \dots + 1 \cdot 87) / 30 = 42.67 \text{ units per cc}$$

8) Fig.5 shows that insulin rate distribution is homogeneous because there is only a single mode located at 38 units per cc.

9) the POLYGON OF CUMULATED FREQUENCIES shown in Fig.6 is easily obtained from the histogram data.

10) the PERCENTILES calculated with the approximate formula for data in classes are:

$$\text{percentile } 25 = 34.5 + 7 \cdot (25 - 23.33) / (63.33 - 23.33) = 34.79$$

$$\text{percentile } 50 = 34.5 + 7 \cdot (50 - 23.33) / (63.33 - 23.33) = 39.17$$

$$\text{percentile } 75 = 41.5 + 7 \cdot (75 - 63.33) / (76.67 - 63.33) = 47.62$$

class	frequency	frequency	class	cumulativ	cumulativ
-------	-----------	-----------	-------	-----------	-----------

central value		in percent	upper limit	frequency	frequency in percent
24,00	3,00	10,00	27,50	3,00	10,00
31,00	4,00	13,33	34,50	7,00	23,33
38,00	12,00	40,00	41,50	19,00	63,33
45,00	4,00	13,33	48,50	23,00	76,67
52,00	2,00	6,67	55,50	25,00	83,33
59,00	2,00	6,67	62,50	27,00	90,00
66,00	1,00	3,00	69,50	28,00	93,33
73,00	1,00	3,00	76,50	29,00	96,67
80,00	,00	0,00	83,50	29,00	96,67
87,00	1,00	3,00	90,50	30,00	100,00

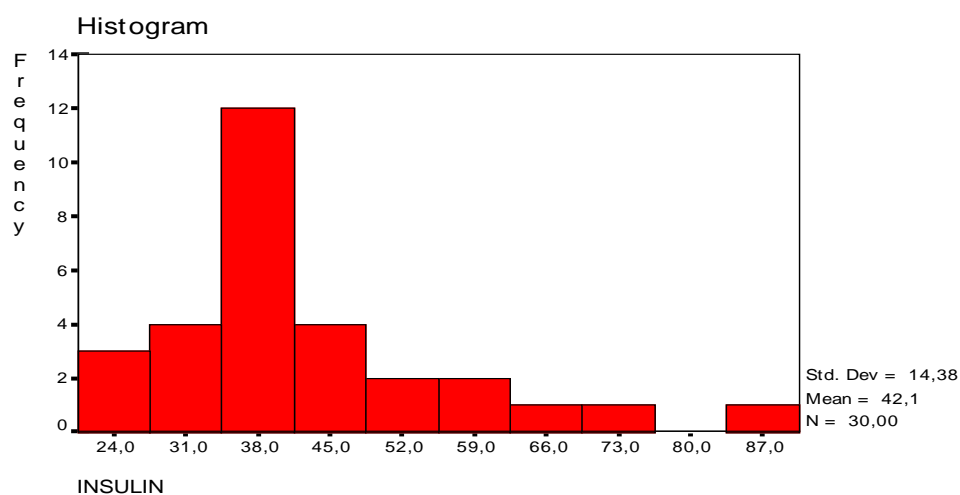


Fig.5 Histogram of insulin rate

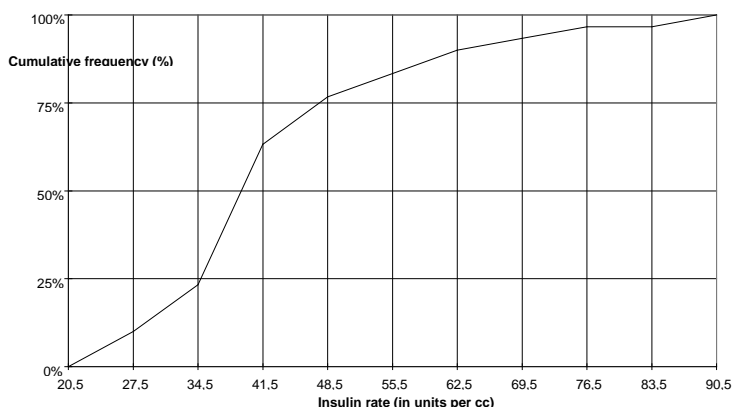


Fig.6 Polygon of cumulative frequencies for insulin rate

• **Dispersion parameters for quantitative variables: Variance, Standard-deviation (SD), Skewness coefficients.**

Dispersion is matter of degree and kind of variability.

The **variance** (Symbol s^2) is the sum of the squared differences between each value in the sample and the mean, divided by the number of degrees of freedom (DF). DF of variance in the sample is n (sample size). We will see later (Chapter 3) that DF for estimating variance in the population is equal to $n-1$.

The **standard-deviation** (or SD; Symbol s) is the square root of the variance. This gives a parameter measured in the same units as m . The **coefficient of variation** provides a index of dispersion which does not depend on the units of measurement.

Variance and SD are parameters of dispersion around the mean. To understand the variance and SD formulas it must be stressed that we cannot simply add the differences between sample values and their mean because the sum is always null. There are several ways to escape this problem. One solution is to take the mean of the absolute differences to obtain what we call the "mean deviation". This parameter is however seldom used because absolute values are not easily processed in the mathematical framework. The alternative consists in squaring the differences.

The **skewness** refers to the degree of asymmetry of the distribution. The distribution is symmetrical when there is the same number of data below and above the mean, i.e. when mean and median coincide. When the mean is lower than the median, the distribution is "skewed" to the left. When the mean is larger than the median, the distribution is generally (but not always) skewed to the right. **Fisher skewness coefficient** (g_1) is negative in case of left asymmetry, null in case of symmetry and positive in case of right asymmetry.

$$\text{Variance} = s^2 = \left(\sum_{i=1}^n (x_i - m)^2 \right) / n$$

$$\text{Standard-deviation (SD)} = s = \sqrt{\left(\sum_{i=1}^n (x_i - m)^2 \right) / n}$$

$$\text{Standard-deviation (SD)} = s \cong \sqrt{\left(\sum_{c=1}^k f_c (x_c - m)^2 \right) / n}$$

x_c = central value of class; f_c =frequency of class; k = number of classes

$$\text{Coefficient of variation} = s/m$$

Examples of variance, SD and coefficient of variation.

for the following sample of weights in kg: 6, 9, 10, 12, 15

$m=10.4$ kg

$$\text{variance} = S(x_i - m)^2 / n = (19.36 + 1.96 + 0.16 + 2.56 + 21.16) / 5 = 9.04 \text{ kg}^2$$

$$SD = \sqrt{S(x_i - m)^2 / n} = \sqrt{9.04} = 3.01 \text{ kg}$$

Notice that $S(x_i - m) / n = (-4.4 - 1.4 - 0.4 + 1.6 + 4.6) / 5 = 0$

$S(x_i - m) / n = \text{always } 0$

Comparison between samples with different variances.

Take the two following samples of age measurements:

first sample: 5, 7, 8, 10, 12, 14, 15

second sample: 1, 4, 7, 10, 14, 18, 20

For the first sample: $m=10.14$ years and $s^2=11.84$ years squared

For the second sample: $m=10.57$ years and $s^2=43.39$ years squared

Comparison between SD and coefficient of variation:

if 3108, 3245, 3302, 3104, 4002 are weights in grams

$$m= 3352.2 \quad s= 333.91 \quad s/m \cong .10$$

if the same weights are measured in kg: 3.108, 3.245, 3.302, 3.104, 4.002

$$m= 3.352 \quad s= 0.334 \quad s/m \cong .10$$

 SD for data in classes.

example: albumin data (see above)

With the exact formula, $SD = 14.38$ units per cc

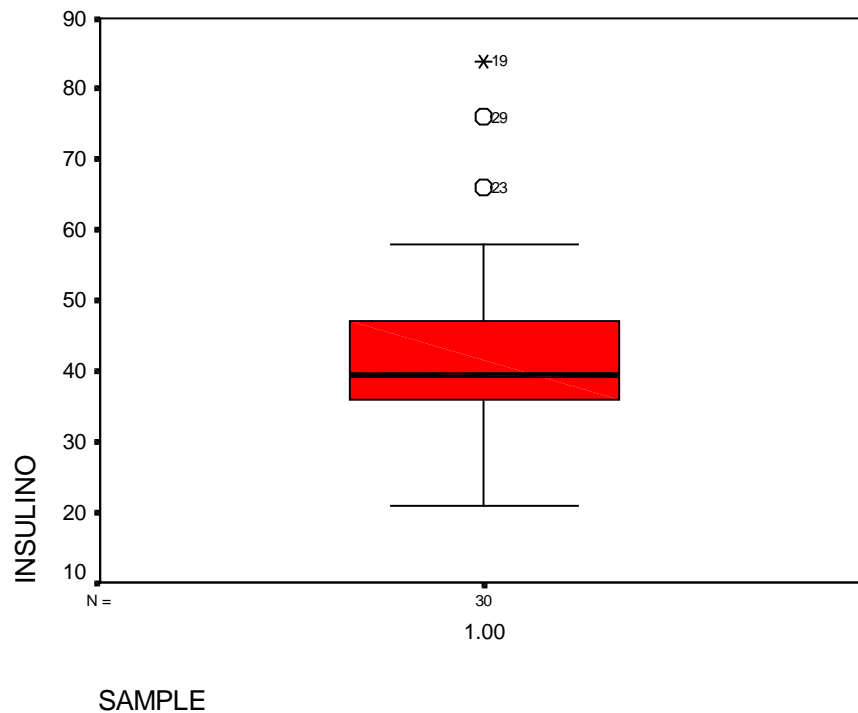
The SD calculated with the approximate formula for data grouped into classes is:

$$SD \cong \sqrt{(3 \cdot (24-42,67)^2 + 4 \cdot (31-42,67)^2 + \dots + 1 \cdot (87-42,67)^2) / 30} = 14.15 \text{ units per cc}$$

Example of skewed distribution

The distribution of insulin rate (in Fig.5) is skewed to the right, as it often happens for physiological or psychological variables. Accordingly, Fisher skewness coefficient is positive ($g_1= 1.946$). Right asymmetry can be removed by taking the logarithm of the variable.

● Boxplot graph: values inside the box are between P25 and P75 (50% of the distribution); small horizontal bars are the largest and smallest values which are not outliers; O points are values more than 1.5 boxlength below P25 or above P75; * points are values more than 3 boxlength below P25 or above P75.



• **Central tendency and dispersion parameters for binary variables: Proportions.**

The proportion of cases in one category (p) also gives the other (1-p). A proportion is equivalent to a mean value provided that the value 1 is assigned to one category and 0 to the other. The variance also has a meaning and is equal to $p*(1-p)$.

$$m = (n_1 * 1 + n_0 * 0) / n = n_1 / n = p$$

n_1 = number of values in category 1; n_0 = number of values in category 0

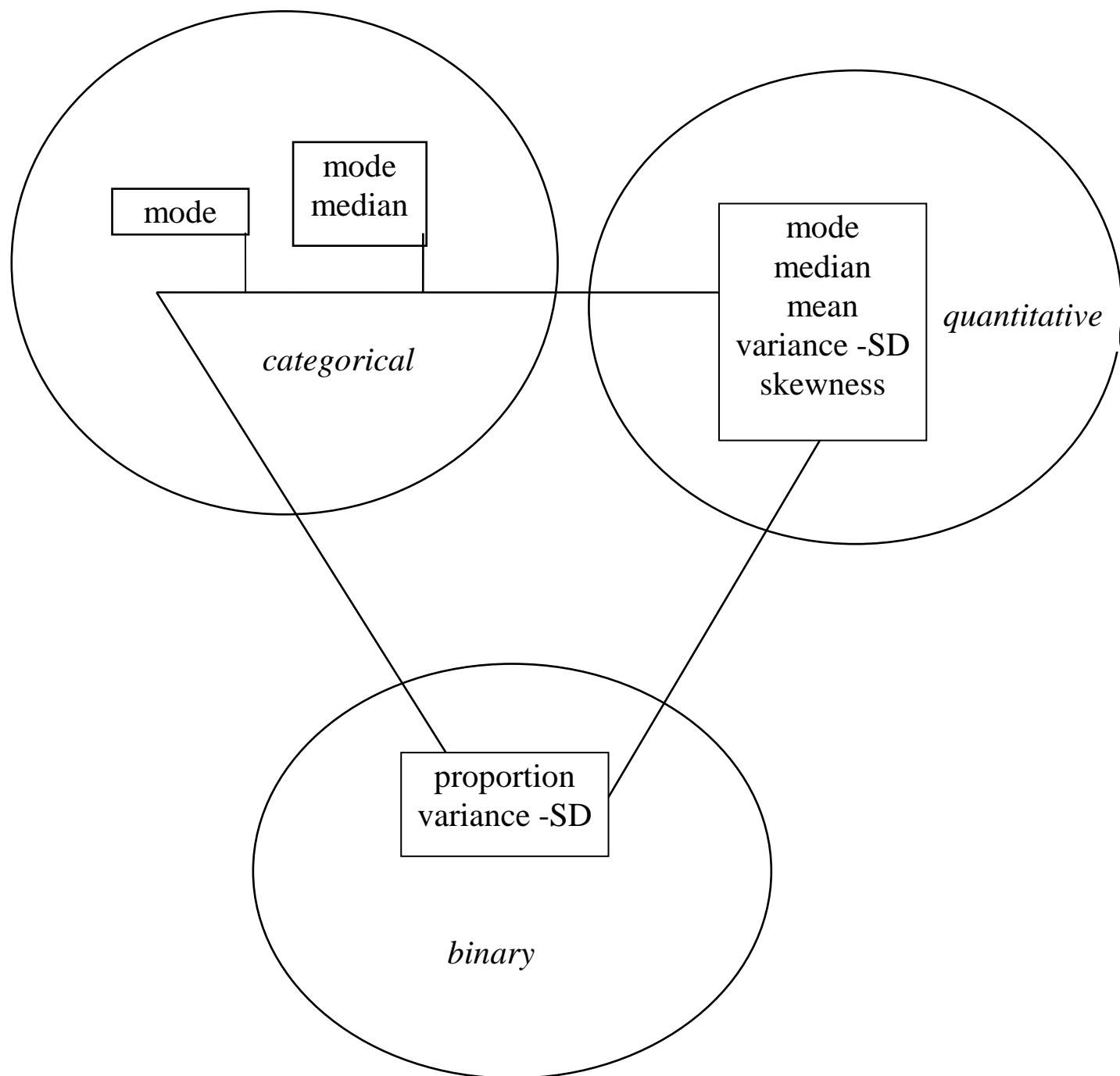
$$n = n_1 + n_0$$

$$s^2 = (n_0(0-p)^2 + n_1(1-p)^2) / n = p*(1-p)$$

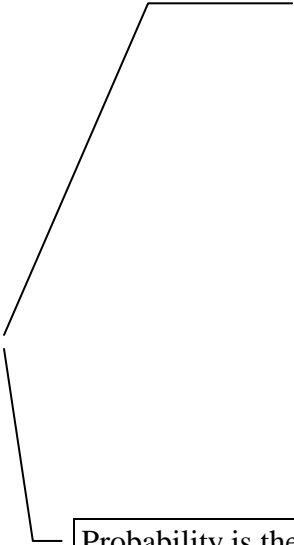
Examples: in a sample of 100 subjects, the variance of p is

p	variance	SD
.5	.25	.5
.1	.09	.3
.9	.09	.3

The variance is the highest for .5 and gets lower when the proportion gets closer to 0 or to 1. This makes sense. Sample heterogeneity is maximal when the two characters occur with the same frequency (50 %).



Chapter 3. Probability

- 
- Probability
 - Normal distribution
 - Poisson distribution
 - Binomial distribution
 - Categorization and ROC curves

Probability is the limite value of relative frequency in an ideal-infinite sample called the population. Frequency distributions are obtained with data, probability distributions are given by laws (formulas). For continuous variables, probability distribution follows the Normal law. For discrete probability distribution follows the Poisson law, or Normal law (expected frequency ≥ 5). For binary variables, probability distribution follows the Binomial law or the Normal law (for expected frequencies ≥ 5).

• Probability

- Empirical definition (Bernoulli, Kolmogorov). The **probability (P)** of an event is the limit value reached by its relative frequency when the size of the sample tends to the infinite. The exact probability cannot be given because the relative frequency of an infinite sample is not available. The relative frequency however allows to approximate the probability. And, the larger the sample, the better the approximation. The problem raised by this definition of probability is that the precision of the estimate cannot be specified without referring to the notion of probability: there is a danger of circularity in the empirical definition of probability.

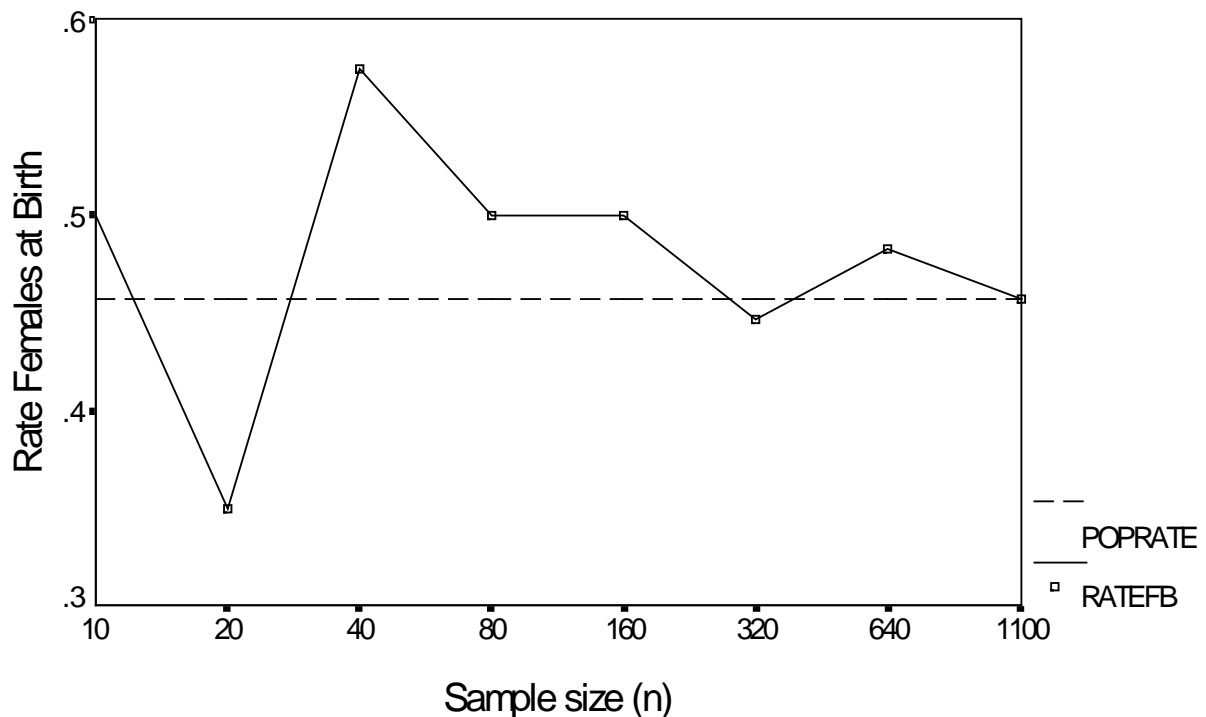


Fig.3.1 Empirical Probability - Rate of females at birth as a function of sample size in a finite population of 1100 babies.

- Subjective definition (Bayes), a **probability** is the quantification of the subjective "degree of belief" that a proposition is true. This definition allows to start from an initial **prior probability** which can thereafter be improved by empirical observations. Prior probability can be based on a theory, or a model.

Example of empirical probability: a screening test reveals that the number of patients affected by tuberculosis amounts 2700 in a sample of 105000 inhabitants, taken at random in a given district. The relative frequency is of $2700/105000 = 2.57\%$. This can be taken as an estimation of the probability of tuberculosis in the district. The precision of the estimate is provided by sampling theory (see below), which is also based on the notion of probability.

- Probability rules:

Limits

$$0 \leq P(e) \leq 1$$

Addition rule

$$P(e_1 \text{ OR } e_2) = P(e_1) + P(e_2) - P(e_1 \text{ AND } e_2)$$

Events are **exclusive** when they cannot occur together, then: $P(e_1 \text{ AND } e_2) = 0$.

Multiplication rule

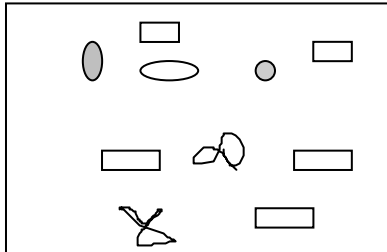
$$P(e_1 \text{ AND } e_2) = P(e_1) * P(e_2 / e_1) = P(e_2) * P(e_1 / e_2)$$

where $P(e_2 / e_1)$ is a **conditional probability**, namely the probability that e_2 occurs when e_1 is present.

Events are **independent** if probability of one event does not depend on the presence vs. absence of the other, then:

$$P(e_2 / e_1 \text{ present}) = P(e_2 / e_1 \text{ absent}) = P(e_2)$$

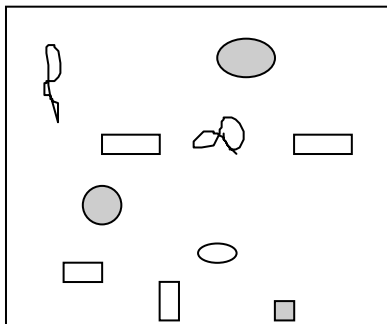
Example of **addition** with exclusive events: different types of cells



$$P(\text{grey or rectangle}) = P(\text{grey}) + P(\text{rectangle}) - P(\text{grey and rectangle})$$

$$= 0.2 + 0.5 - 0 = 0.7$$

Example of **addition** with non-exclusive events:



$$P(\text{grey or rectangle}) = P(\text{grey}) + P(\text{rectangle}) - P(\text{grey and rectangle})$$

$$= 0.3 + 0.5 - 0.1 = 0.7$$

Example of **multiplication** with non-independent (related) events:

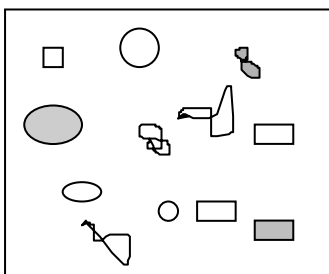
$$P(\text{grey and rectangle}) = P(\text{grey}) * p(\text{rectangle/grey}) =$$

$$= (3/10) * (1/3) = 1/10$$

$$P(\text{grey and rectangle}) = P(\text{rectangle}) * p(\text{grey/rectangle}) =$$

$$= (1/2) * (1/5) = 1/10$$

Example of **multiplication** with independent (non-related) events:



$$P(\text{grey and rectangle}) = P(\text{grey}) * p(\text{rectangle}) =$$

$$= 0.25 * (1/3) = 1/12$$

Examples of addition exclusive events: membership of a bloodgroup for the same individual because the belonging to a bloodgroup excludes the one to another bloodgroup.

if $P(\text{belonging to gr.O}) = .40$ and $P(\text{belonging to gr.A}) = .15$

then $P(\text{belonging to O or A}) = .40 + .15 = .55$

Examples of addition of non-exclusive events: toxic reactions in repeated administrations of a drug because the development of a toxic reaction after the first administration does not prevent a second one (adapted from Colton p.70).

Suppose the $P(\text{toxic reaction}) = .1$ and that the probability of developing 2 successive toxic reactions equals .06, then

$P(\text{toxic reaction either at first or at second administration}) = .1 + .1 - .06 = .14$

Examples of multiplication of independent events: if the probability of toxic reaction remains constant for repeated administration then $P(2 \text{ successive toxic reactions}) = .1 * .1 = .01$

Examples of multiplication of non-independent events: if probability of a second toxic reaction, given a previous one is larger, say .6 instead of .1, then $P(2 \text{ successive toxic reactions}) = .1 * .6 = .06$

Example of independent vs. dependent events in a 2 by 2 table. Consider a finite population of 950 subjects.

	D+	D-	
T+	30	50	80
T-	20	850	870
	50	900	950

D and T are **dependent** as shown by unequal conditional probabilities either in columns or in lines:

$$P(T+ / D+) > P(T+ / D-)$$

$$P(D+ / T+) > P(D+ / T-)$$

	D+	D-	
T+	5	90	95
T-	45	810	855
	50	900	950

D and T are **independent** because conditional probabilities are equal

$$P(D+ \text{ and } T+) = P(D+) * P(T+)$$

Just the same: cell frequency (D+ and T+) = line total * column total / grand total

$$\text{Example: } 95 * 50 / 950 = 5$$

• Bayes' Theorem:

$$P(T+ \text{ and } D+) = P(T+/D+)*P(D+) = P(D+/T+)*P(T+)$$

$$P(T+/D+) = P(D+/T+)*P(T+)/ P(D+)$$

We see that the probability to test positive when diseased is not the same as the probability to be diseased when testing positive. It is only when the probability of testing positively is *equal* to the prevalence that the two conditional probabilities are equal. Otherwise probability to test positive when diseased is larger than the probability to be diseased when testing positive if the probability of testing positively is *larger* than the prevalence. And the converse is true when the probability of testing positively is *smaller* than the prevalence.

Similarly:

$$P(T-/D-) = P(D-/T-)*P(T-)/ P(D-)$$

• Normal distribution

Most natural frequency distributions are bell-shaped. Central values are much more frequent than extreme values and there is a gradual frequency decrease from central to extreme values (see insulin distribution, Fig.5). The typical bell-shaped is symmetrical. Natural distributions are not always symmetrical but can be made symmetrical with appropriate variable transformations (such as log-transforms). Ideal bell-shaped distributions are given by the **Normal** probability formula. A frequency distribution is empirical by nature and is never perfectly Normal.

Normal formula

$$P(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

where **x is a continuous variable**

$$z = \frac{x - \mu}{\sigma} = \text{“Normal deviate”}$$

μ = mean

σ = SD

Why is the Normal distribution so common? Because the sum of a large number of variables follows a Normal distribution whatever the distributions of the variables, provided that the variables are independent.

The **central-limit theorem** says this in more precise terms:

Each sum of n independent random variables $X_1, X_2, X_3, \dots, X_n$ is an asymptotic Normal variable.

Asymptotic means that the distribution gets closer and closer to the Normal as sample size (n) gets larger and larger. Besides the independence requirement, the only other restriction to this theorem is that the variables must be of the same order of magnitude. Otherwise, if the numerical values taken by one of the variables are much larger than those taken by the others, its distribution will dominate the sum.

Example: comparison between two games with dice.

First game: with a single dice, each player chooses a figure from 1 to 6 and wins if the dice falls on the figure. In this game the 6 possible events are equiprobable, provided the dice is fair, and the probability distribution is rectangular (Fig.13).

Second game: with 2 dices, each player chooses a number corresponding to the sum of 2 figures between 1 and 6, that is a number between 2 and 12.

Which number would you choose ?

Choose the 7 because the outcomes are no more equiprobable and 7 is the most frequent combination (Fig.). The distribution has lost its rectangular look for a triangular shape. Further, the distribution becomes unimodal and symmetrical (the most frequent value corresponds to the mean (7)). These are also the two main features of the Normal distribution. However, the triangular distribution is still far away from the Normal one, which is bell shaped and provides a probability for each of the values taken by a continuous variable from minus infinite to plus infinite. For 3 independent variables, the distribution of the sum is already bell shaped, and for 5 it is almost indistinguishable by eye from the Normal.

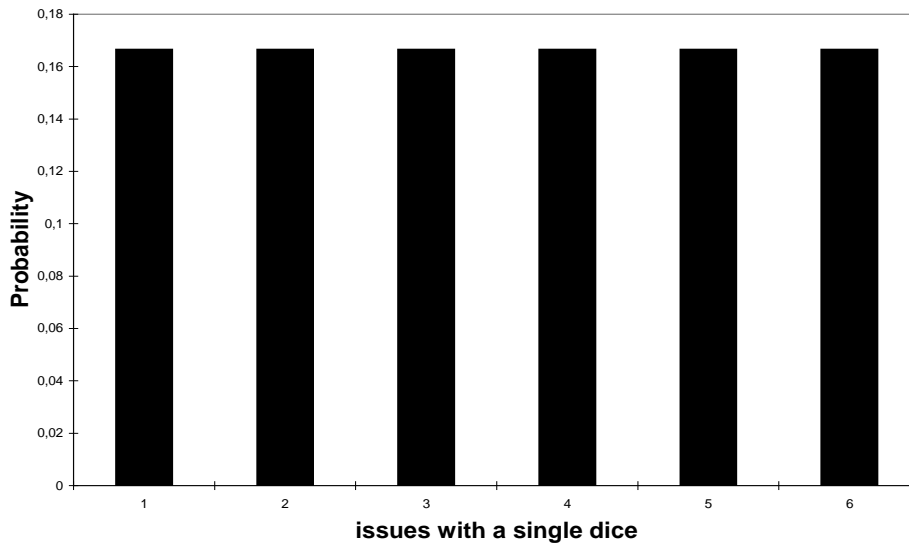


Fig.3.2 Illustration of the Central-Limit theorem : comparison between two games with dice.
First game: 1 dice, the 6 possible figures are equiprobable.

		DICE A					
		I	II	III	IV	V	VI
DICE B	I	2	3	4	5	6	7
	II	3	4	5	6	7	8
	III	4	5	6	7	8	9
	IV	5	6	7	8	9	10
	V	6	7	8	9	10	11
	VI	7	8	9	10	11	12

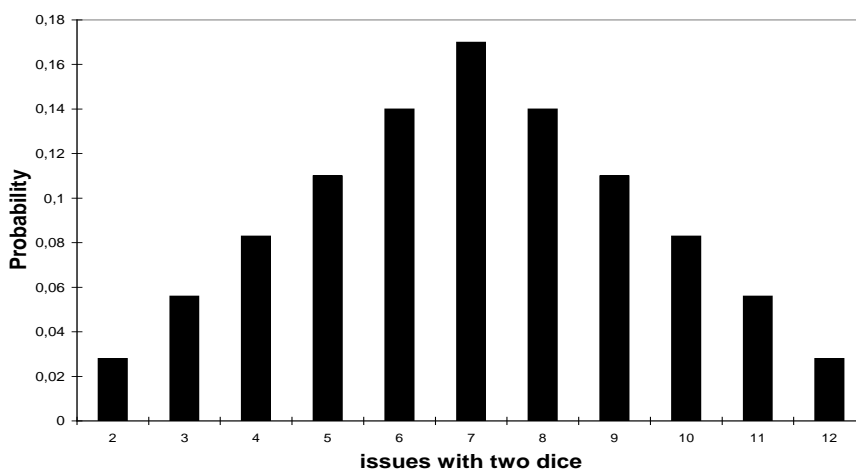


Fig.3.3 Illustration of the Central-Limit theorem : comparison between two games with dice.
Second game: 2 dices, there are $6 \times 6 = 36$ possible issues corresponding to the sum of the 2 dices. The table below gives the SUM of the 2 dices for each possible combination. As can be seen, the middle-range values are more frequent than the extreme values. The distribution is no more rectangular but triangular (as shown on the graph). With 3, 4, 5, ... dices, the distribution becomes progressively bell-shaped.

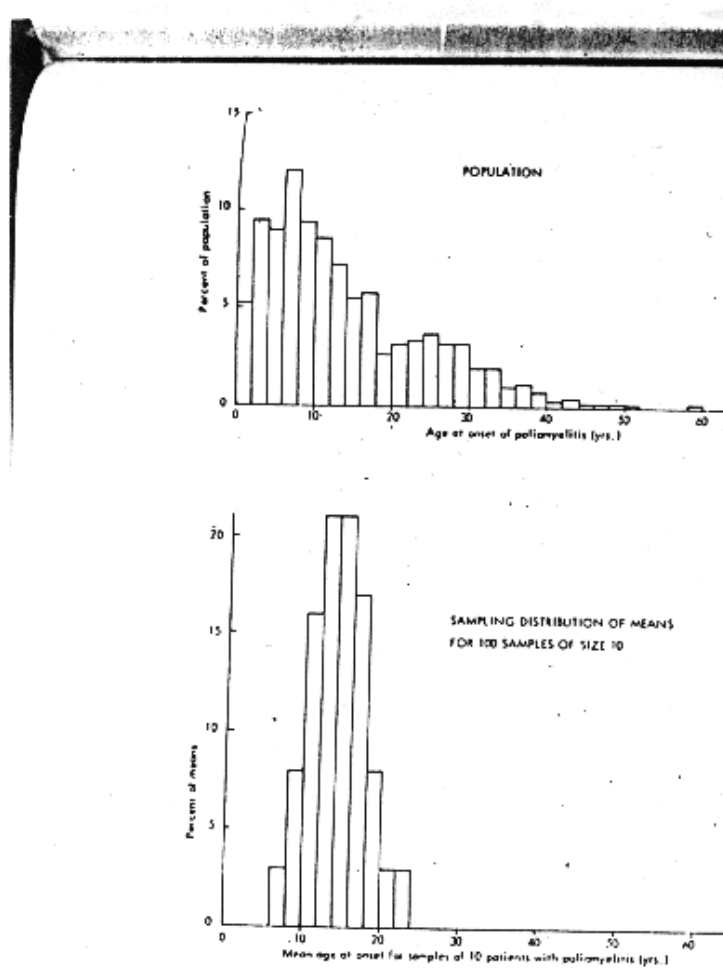


Figure 4.3
Percentage distribution by age of the reported cases of poliomyelitis in Massachusetts in 1949 (top) and empirical sampling distribution of means for 100 samples of size 10 (bottom).

Fig.3.3.1. An example showing how the distribution gets closer to Normal as the size of the sample increases (From Colton, Fig.44; see Ref. in Chapter1).

A large sum of random variables has a probability distribution close to the Normal one (see Fig.10). The Normal distribution is unimodal and symmetrical. As the Normal distribution is for a continuous variable, the sum of the probabilities between any two values is a **probability area**, which is represented on the graph by the area below the curve and between the two values. The total probability area is equal to 1 (values are exclusive and exhaustive events).

Remarkable values:

<u>Limits</u>	<u>Probability Areas</u>
between $\mu - \sigma$ and $\mu + \sigma$	about 2/3 (67 %)
between $\mu - 2\sigma$ and $\mu + 2\sigma$	about 95 %
between $\mu - 3\sigma$ and $\mu + 3\sigma$	about 99.5 %

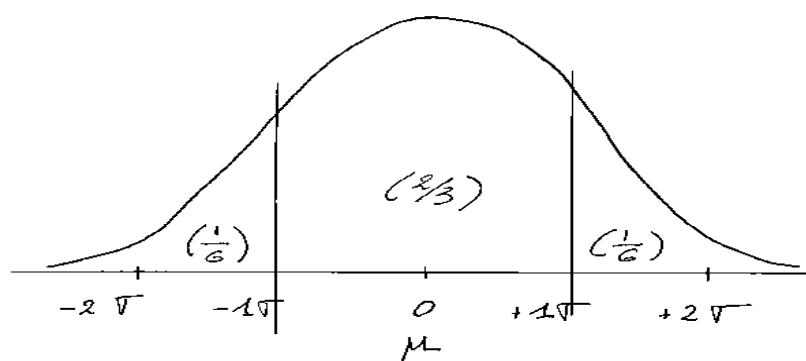
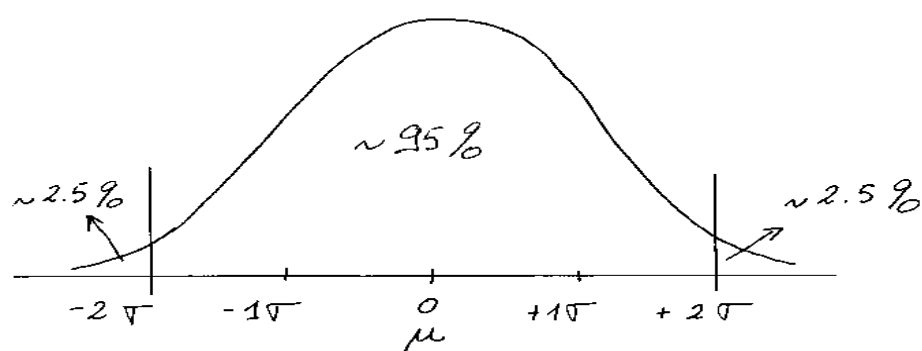
Any Normal distribution is completely specified by 2 parameters (μ and σ) and can be transformed into the **standard Normal distribution** of mean=0 and SD=1.

$$z = (x - \mu) / \sigma$$

If x is distributed $N(\mu, \sigma)$ then z is distributed $N(0, 1)$. The **normal-deviate z** gives the distance between the mean and any point of the distribution in standard deviation units. Example: if the weight is $N(3000g, 500g)$, a weight of 4000g corresponds to $z=2$ which indicates that it is 2 SD above the mean; a weight of 2250g corresponds to $z=-1.5$ which indicates that it is 1.5 SD below the mean, etc... A condensed table of Normal probability values is given below (for a full table see Kirkwood pp.206-207). It gives the probability area above z , for z values regularly spaced between 0 and +3. As the Normal distribution is symmetrical, the probability area below $-z$ is equal to the one above $+z$.

Examples:

above $z=1$	probability area = .1587
below $z=-1$	$p = .1587$
above $z=1.96$	$p = .025$
below $z=1.64$	$p = 1 - .0505 = .9495$
above $z=-1.64$	$p = 1 - .0505 = .9495$



z	p
0.0	0.5
0.84	0.2
1.0	0.16
1.64	0.05
1.96	0.025
2.33	0.01
2.58	0.005
3.09	0.001

Fig.3.4 Normal Probability Curve and Table (condensed)

• Poisson distribution (for discrete variables)

Poisson distribution applies to variables such as:

- the number of childbirth per day in a hospital
- the number of accidents per year at a crossroads
- the number of trypanosomes per blood sample
- the number of bacteria per volume of water.

Each of these variables is a number of events as a function of an extraneous factor such as time, space, volume, etc... The occurrence of an event is distributed according the Poisson law if the probability rate of occurrence is constant (over time, space; e.g. birthrate constant over days...). Constancy can be admitted if the 2 following conditions are fulfilled:

1) **PROPORTIONNALITY**: the number of events must be proportionnal to the extraneous factor taken as reference (for time: number of events per month = number per year/12 etc...). This implies that the number of events should not depend on the piece of reference (the moment of observation, portion of space ...). Non proportionnality arises from trends, especially for long periods of time (over decades). Cyclic trends, or "seasonal variations", can disturb proportionnality in the short run (months in the year with increase in childbirth).

2) **INDEPENDENCE**: the events must be independent (one delivery must not affect the occurrence of another, one accident must not give rise to another). This can be admitted provided that the time span is appropriate (a day rather than an hour for accidents at a crossroads).

The variance of the Poisson distribution is equal to the mean: $\sigma^2 = \mu$. The Poisson distribution is characterized by a right (positive) asymmetry but tends to be symmetrical as the mean increases. Further, Poisson distribution is approximatively Normal for mean values equal or larger than 5.

Poisson Formula (for $\mu > 0$)

$$p(x \text{ events}) = \mu^x / (e^\mu * x!)$$

$$\sigma^2 = \mu$$

where x = any positive integer

e = constant = basis on natural logarithms = about 2.72

μ = expected number of events = mean of the distribution

Normal approximation

For $\mu \geq 5$: Poisson \rightarrow Normal

Example of Poisson variable: if the expected number of childbirth in a given hospital is 2300 per year; then the expected number per day is about $2300/365 = 6.3$. Application of Poisson formula gives (A graphical representation is provided in Fig.3.5):

number of childbirth	probability
0	.0018
1	.0115
2	.0363
3	.0762
4	.1200
5	.1513
6	.1588
7	.1429
8	.1126
9	.0788
10	.0496
etc...

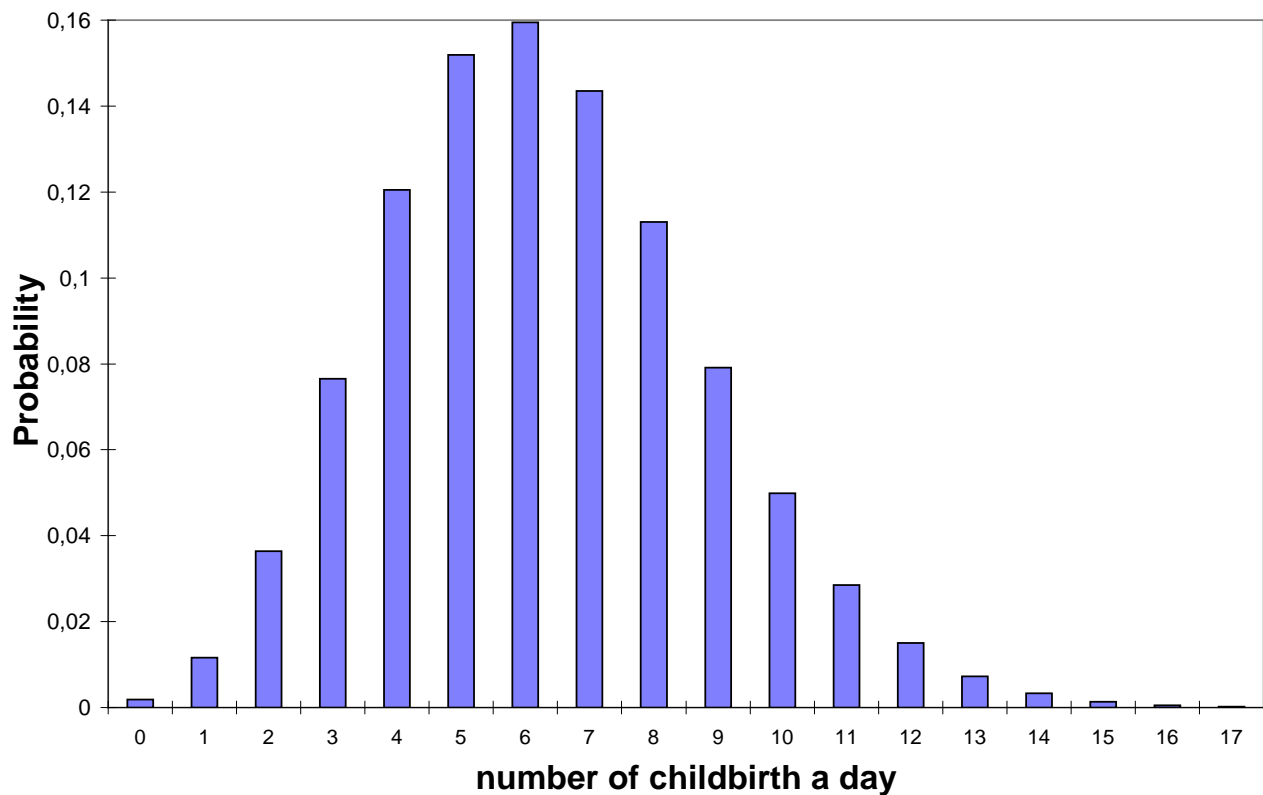


Fig.3.5 Poisson distribution

• **Binomial distribution (for proportions)**

Poisson and Binomial law both apply to number of events. However, a Binomial variable corresponds to the realisation of some event versus another in a sample of definite size (the size of the sample corresponds to the sum of the frequencies of the two events). For a Poisson variable, the size of the sample is not specified (the frequency of the alternative event is not specified). In fact, the Poisson law applies to events of which the frequency is very small by comparison with the alternatives (in the above examples: the number of women who do not deliver a given day in a hospital, the number of vehicles who get through the crossroads without accident ...).

Knowing the probability of an event (π) for a single item in a population, what is the probability that this event occurs a given number of times (x) in a random sample of n items? This is the sampling distribution problem for a proportion. The solution is provided by the **Binomial** law, or Bernoulli law. The Binomial distribution is symmetrical for $\pi = .5$, otherwise it is not symmetrical. Notice that three different proportion-like values are involved in the Binomial formula: π the probability of infections in the population, p the probability of samples with x events/ n , and x/n the proportion of events in a sample.

Binomial Formula

$$p(x \text{ events over } n) = C_n^x * (\pi)^x * (1 - \pi)^{n-x}$$

$$C_n^x = n! / x!(n-x)! \quad (\text{combination of } n \text{ events } x \text{ by } x)$$

π = prob. of the event in the population = mean of the Binomial distribution ($\mu = \pi$). Variance is $\sigma^2 = \pi * (1 - \pi) / n$

p = prob. of x times the event in a sample of size n

x = number of events per sample

Normal approximation for a Binomial distribution: if $n * \pi$ and $n * (1 - \pi)$ are both equal or larger than 5

Binomial → Normal

Example: If the probability of infection is .3 for the subjects which undergo a specific operation, what is the probability of having 0, 1, 2, 3, 4 infected subjects among the 4 operated each day in a hospital?

Partial answers can be obtained by the application of the multiplicative law. If we admit that the group of 4 daily operated subjects is a random sample taken from the population of all those who undergo the operation, then the individual probabilities of infection are independent and remain equal to .3, and :

$$p(\text{all infected}) = (.3)^4 = .008$$

$$p(\text{none infected}) = (.7)^4 = .24$$

The multiplicative law does not give the complete solution if we want the probability that some part of the group will be infected.

For 1 infection over 4:

$$p(\text{patient in the first bed infected and the 3 others not}) = .3 * (.7)^3$$

This probability must be multiplied by 4 because there are 4 patients in the sample and for each of them the infection risk amounts 30 %:

$$4 * .3 * (.7)^3 = .41$$

For 2 infections over 4:

$$p(\text{patients in the first 2 beds infected}) = (.3)^2 (.7)^2$$

This probability must be multiplied by the number of combinations of 4 elements 2 by 2 :

$$C^2_4 = 4! / 2! (4-2)! = 6$$

$$p(2 \text{ infected over } 4) = 6 * (.3)^2 (.7)^2 = .264$$

Similarly for 3 infections over 4:

$$p(3 \text{ infected over } 4) = C^3_4 * (.3)^3 (.7)^1 = .0756$$

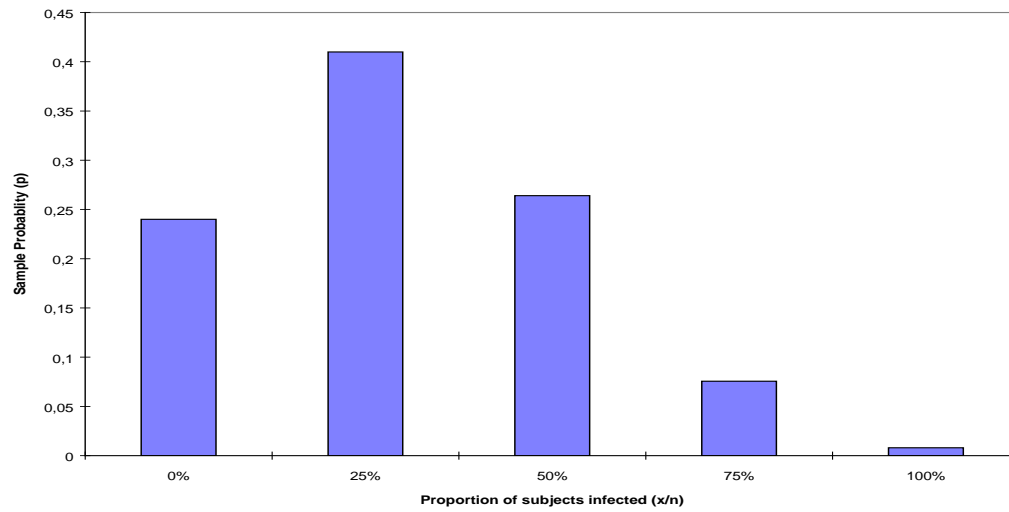
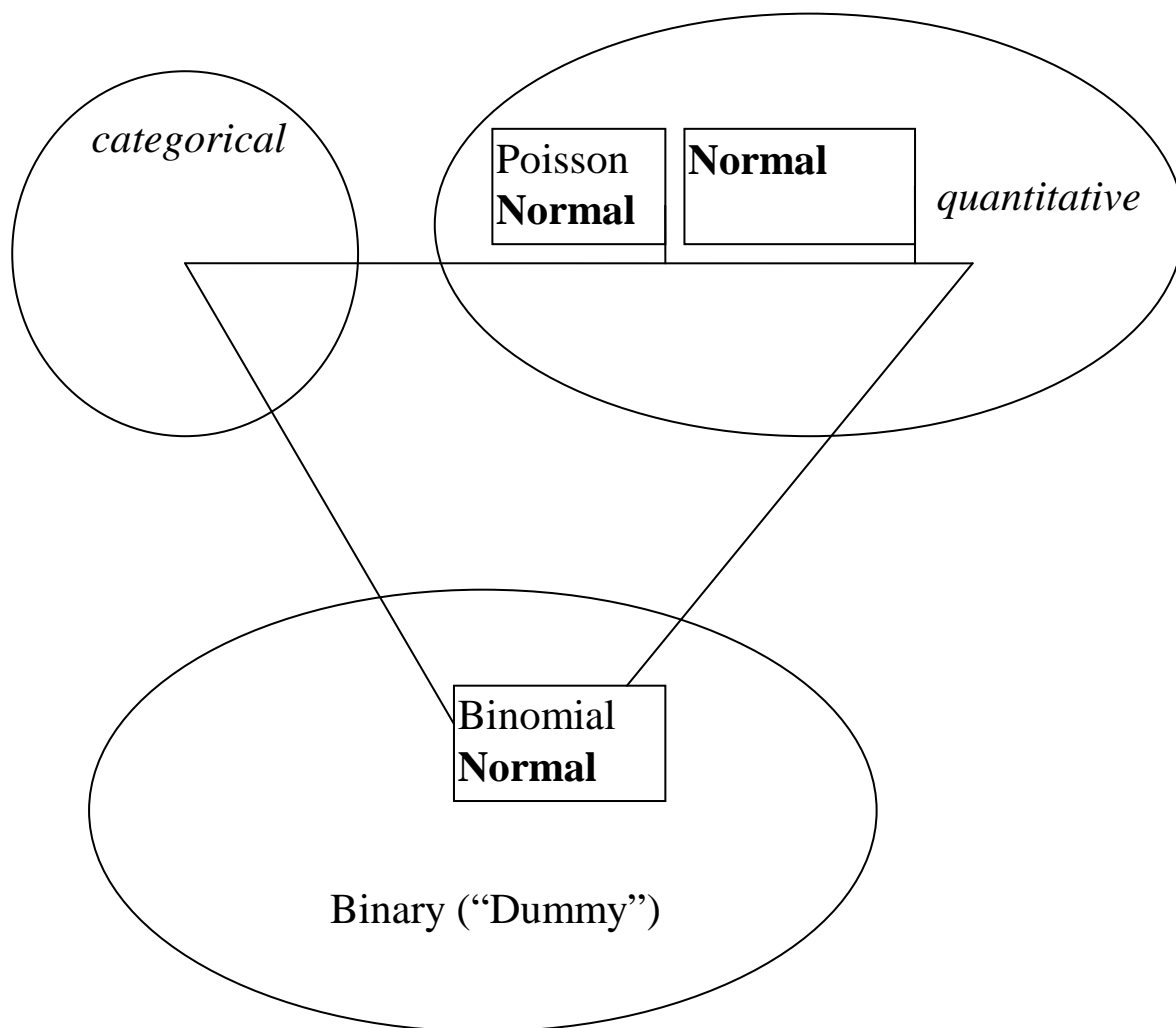


Fig.3.6 Binomial law.



• ROC curves for diagnostic tests

In applied statistics we often consider more than just one probability distribution. For example, analysis of diagnostic tests typically requires two different distributions, one for “normal” subjects and the other for “diseased” (see Fig. 3.8). As these distributions always exhibit some degree of overlap perfect classification is not possible. We cannot find some criteria (C) such that all normal subjects fall on one side and all diseased on the other side. In other words, diagnostic tests are always below 100 % specificity and sensitivity, and above 0% false positives and false negatives.

Notations and terminology

	D+	D-
T+	(P(D/d) or P(T+/D+)	(P(D/n) or P(T+/D-)
T-	(P(N/d) or P(T-/D+)	(P(N/n) or P(T-/D-)

	D+	D-
T+	sensitivity	false positive
T-	false negative	specificity

Different tests will differ in performance, the lesser the overlap between the distributions the better the performance. But for a given test, with a specific degree of overlap, specificity and sensitivity will depend on the location of the criteria. Moving the criteria along the test scale has opposite effects on specificity and sensitivity, with one coefficient increasing if the other is decreasing (compare Figs. a & b in 3.8).

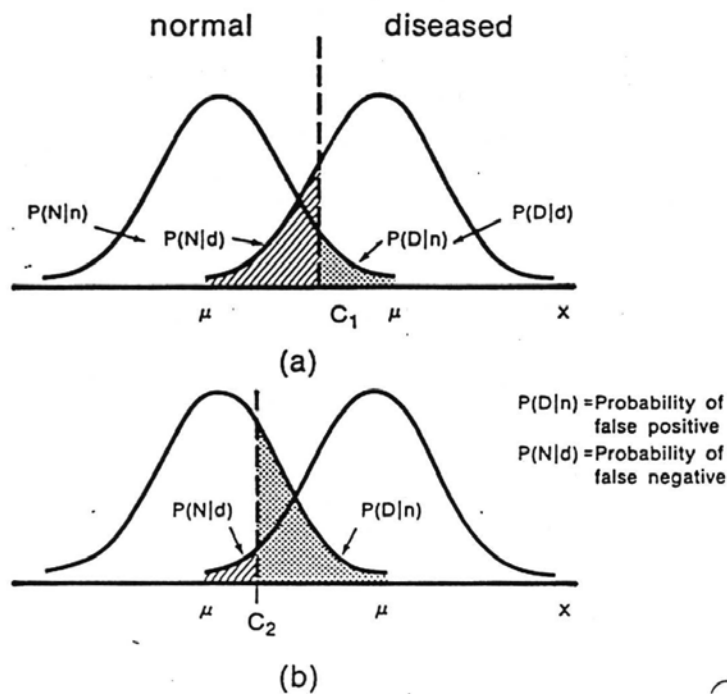
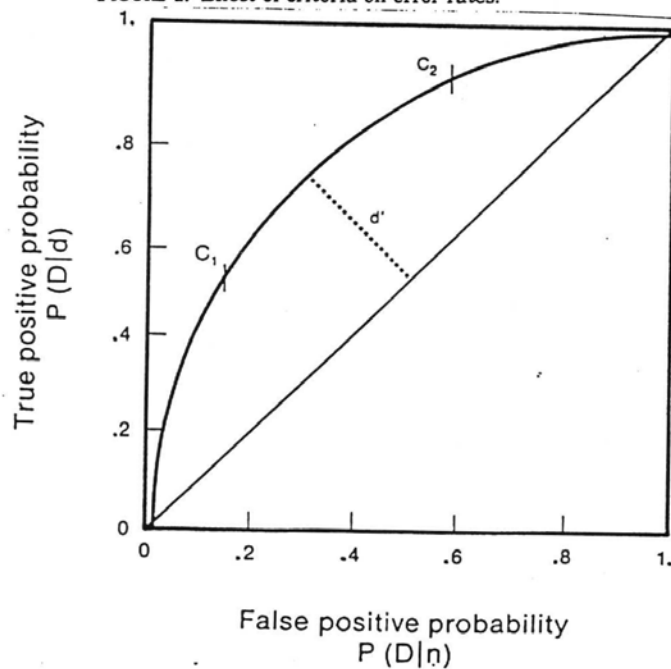


FIGURE 1. Effect of criteria on error rates.



from Descheich & Lee (1981)

Figure 3.8. ROC curve

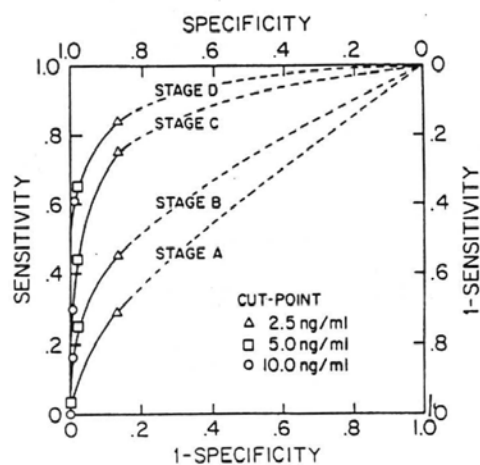
When distributions are Normal with equal variance (as in Fig. 3.8), moving the criteria has regular effects on sensitivity and specificity. These effects follow what is called a “**Receiver Operating Characteristic**” curve (ROC curve) in a two-dimensional diagram with sensitivity on the Y axis

and false positive rate (= 1 -specificity) on the X axis. (Fig.3.8, bottom). Any point along this curve corresponds to the sensitivity-specificity pattern for a given criteria location.

The diagonal in a ROC diagram corresponds to a test with sensitivity equal to false positives. This occurs when distributions of normal and diseased subjects completely overlap. The lesser the overlap the larger the distance between the ROC curve and the diagonal. Degree of non-overlap can also be measured by taking the difference between distribution means divided by their SD:

$$d' = \frac{m_d - m_n}{s}$$

The d' is similar to a z value. It gives the distance between means in number of SD (SD units). Comparison of ROC curves for different tests allow to rank their performances independently of criteria location (see Fig.3.9).



The sensitivity/specificity of a test varies with the stage of disease. ROC curve for carcinoembryonic antigen (CEA) as a diagnostic test for colorectal cancer according to stage of disease. (Redrawn from Fletcher RH: Carcinoembryonic antigen. *Ann Intern Med* 104:66-73, 1986.)

Figure 3.9. ROC curves for different tests.

The smooth aspect of the ROC curve is only obtained with ideally Normal and equal-variance distributions. In practice, ROC curves are less regular as shown by the curve in Fig.3.10., which relates sensitivity-specificity for different prediction of disease at birth with a combination of different parameters (skull perimeter, delivery mode, mother's age, mother's education, nber. living children).

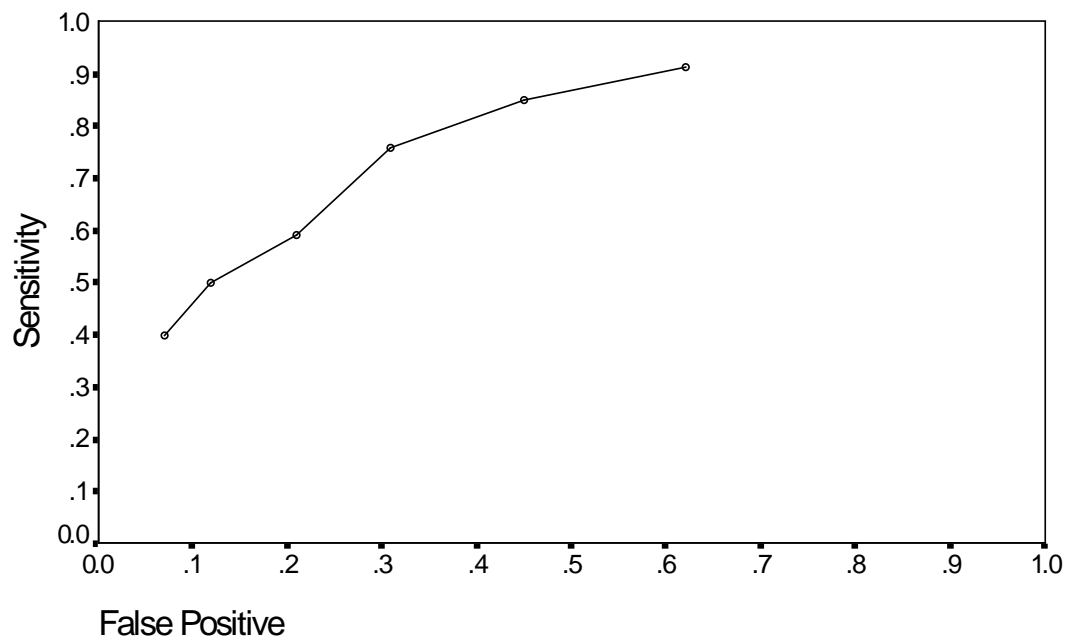
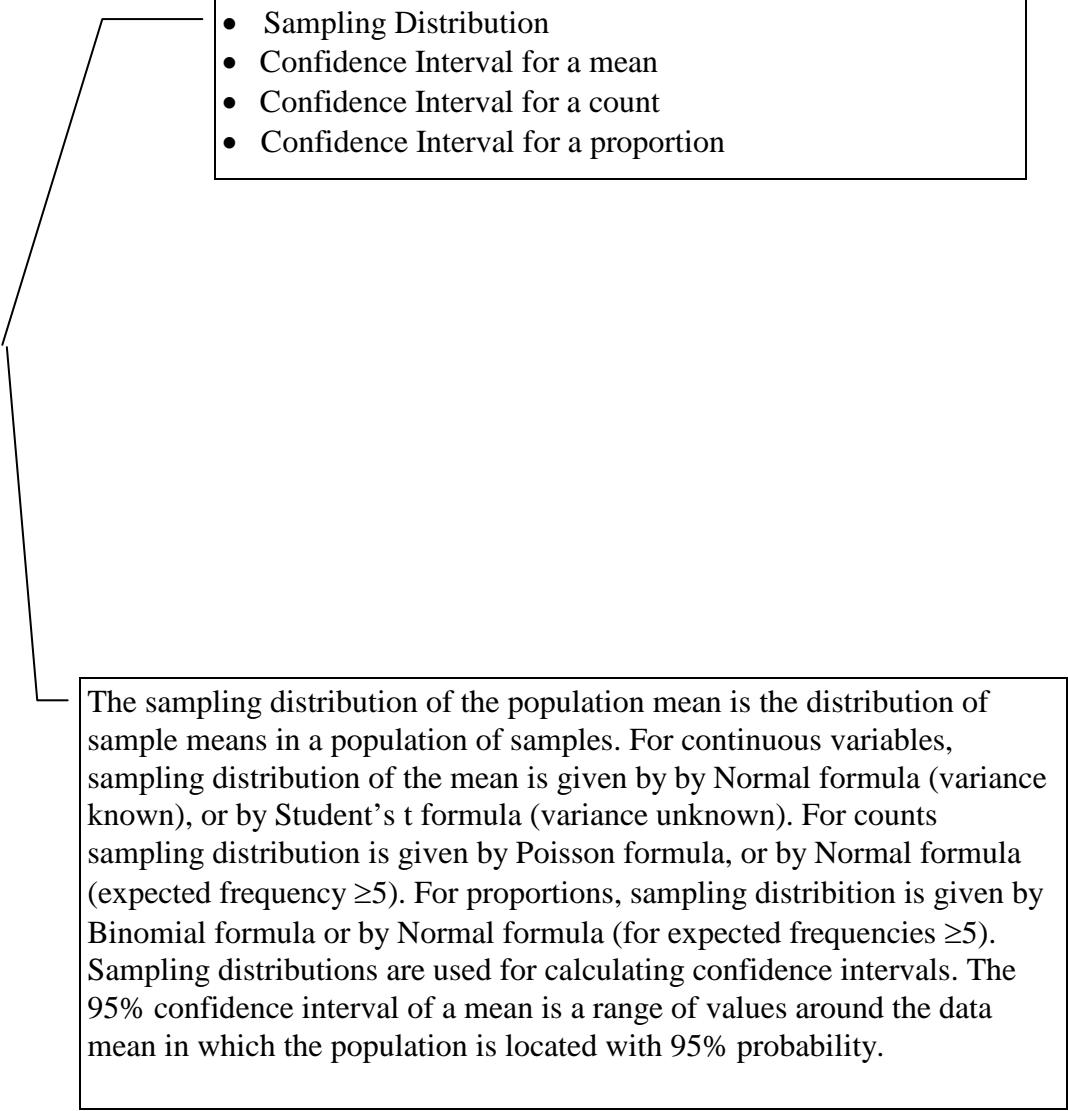


Fig.3.10. ROC curve for prediction of disease at birth with a combination of different parameters (calculated from data in mat7e97.sav file; data by courtesy of Prof. Hennart PHS-ULB).

Chapter 4. Confidence Intervals

- 
- Sampling Distribution
 - Confidence Interval for a mean
 - Confidence Interval for a count
 - Confidence Interval for a proportion

The sampling distribution of the population mean is the distribution of sample means in a population of samples. For continuous variables, sampling distribution of the mean is given by Normal formula (variance known), or by Student's t formula (variance unknown). For counts sampling distribution is given by Poisson formula, or by Normal formula (expected frequency ≥ 5). For proportions, sampling distribution is given by Binomial formula or by Normal formula (for expected frequencies ≥ 5). Sampling distributions are used for calculating confidence intervals. The 95% confidence interval of a mean is a range of values around the data mean in which the population is located with 95% probability.

• Sampling Distribution for Mean

There are several ways of extracting a sample from a population (see Chapter 5 for further information). The simplest way to proceed, for the purpose of parameter estimation, is to extract elements at random elements (i.e. each element must have the same probability to be extracted) and independently of each other (i.e. the probability that one element is extracted must not depend on the extraction of another element; counterexample: "snowball" sampling technique in toxicology). When these 2 conditions are fulfilled, the sample is said to be "**random**" and "**simple**". When samples are random and simple, each possible sample has the same chance of being selected from the population.

Suppose that we measure the mean values (m) of the random-simple samples of a given size (n) extracted from a given population. We can imagine a population of sample means and a corresponding **sampling distribution**. The sampling distribution of a mean is Normal even if the distribution of the population of individual items is not Normal, provided that sample size (n) is not too small. This is a consequence of the Central-Limit Theorem. Finally, the variance of the sampling distribution of the mean is n times smaller than the variance of the individual items in the population. The variance of the mean is n times smaller than the one of individual items. The SD of the mean is called the **standard error**.

The variance of a sum of n independent variables is equal to the sum of their variances

$$\sigma^2 \text{ of } (\sum x_i) = \sum \sigma_i^2 = n * \sigma^2$$

As each variable is extracted from the same population of variance

$$\sigma^2 (x_i) = \sigma^2$$

and as

$$\sigma^2 (x_{i/n}) = \sigma^2/n$$

then

$$\sigma^2 (\sum x_{i/n}) = n * \sigma^2/n$$

$$\sigma^2 \text{ of } m = \sigma^2/n$$

$$\text{Standard Error} = \text{SD}(\bar{m}) = \sigma/\sqrt{n}$$

• Confidence interval for a mean

Statistical inference theory can be used for assessing the generality of a parameter value. This is called parameter **estimation**. Parameter estimation leads to the specification of a “**confidence interval**”, which is the statistical equivalent of the precision interval in measurement theory (if the precision of weight measurements is in grams, then each weight is measured with a precision of ± 0.5 g). However, the confidence interval is not a deterministic concept. The true value is not necessarily within the interval, it only has some chance to be there.

The mean value in a sample can be considered as an estimate (called a **point estimate**) of the true mean, or population mean. The precision of the estimate is given by the limits of the confidence interval.

A **confidence interval** is an interval as small as possible around the sample mean and such that the population mean is contained in it for a given percentage of the samples. Thus the 95 % confidence interval contains the true mean for 95 % of the samples; the 99% confidence interval contains the true mean for 99% of the samples and so on. Let us simplify the problem by looking for a symmetric interval around \bar{m} . Then, all we have to find to specify is a single value, let us call it c , such that:

$$\bar{m} - c < \mu < \bar{m} + c \quad \text{for 95\% of the samples.}$$

As we have: $-\bar{c} < \mu - \bar{m} < +\bar{c}$, the problem is to find an interval centered on 0 which is the midpoint between $-\bar{c}$ and $+\bar{c}$...

and which contains the difference $\mu - \bar{m}$ for 95% of the samples.

Given that $\mu - \bar{m}$ is distributed normally (like \bar{m}) with zero mean ($\mu - \mu$) and σ/\sqrt{n} as standard deviation (like \bar{m}),

the limits of the interval are equal to $-\text{or } + 1.96 * \sigma/\sqrt{n}$ (see

Fig.4.1).

Indeed the limits would be $-\text{or } + 1.96$ for the standard Normal distribution because the probability of having a value either above 1.96 or below -1.96 is .05 (see Normal table in Chapter 9) and hence the probability of having a value between -1.96 and 1.96 is 95%. Given that $\mu - \bar{m}$ is also distributed normally with zero mean but with a SD = σ/\sqrt{n} , these values must be multiplied by σ/\sqrt{n} in order to obtain the corresponding limits.

Conclusion:

the values $m \pm 1.96 \frac{\sigma}{\sqrt{n}}$ are the 95% confidence limits for the population mean μ , that is μ is contained within these limits for 95% of the samples.

In the same way:

the 99% confidence limits are $m \pm 2.58 \frac{\sigma}{\sqrt{n}}$, where 2.58 corresponds to $p=.01$ in the bilateral table (A2).

Confidence Interval
Intervalle de Confiance

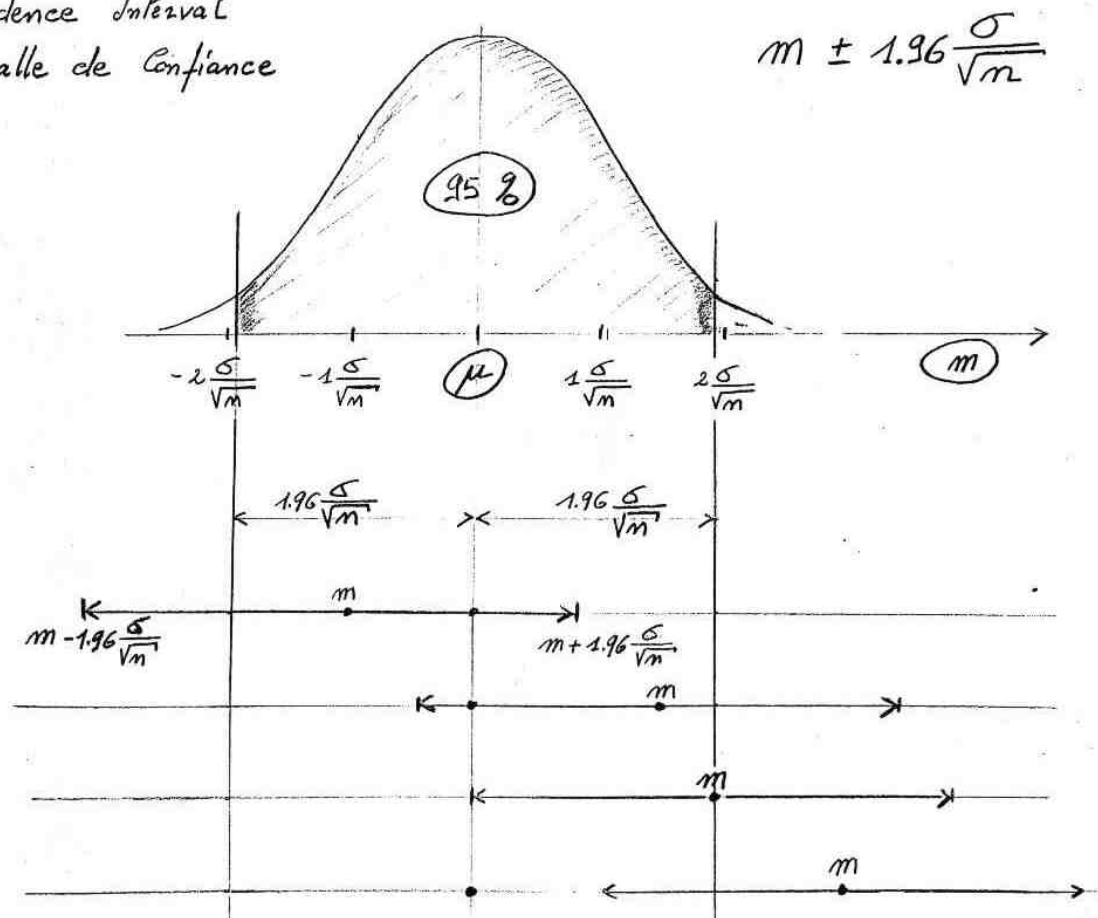


Fig.15 Confidence Interval

Fig.4.1 Confidence interval

example: from Colton p. 127/ survival time for drug-treated cancer patients.

$m = 46.9$ months

$n = 100$ subjects

$\sigma = 43.3$ months

the 95% confidence limits are: $46.9 \pm 1.96 \cdot 43.3 / \sqrt{100} =$

$$46.9 \pm 8.5 = (38.4 ; 55.4)$$

the 99% confidence limits are: $46.9 \pm 2.58 \cdot 43.3 / \sqrt{100} =$

$$46.9 \pm 11.17 = (35.7 ; 58.1)$$

• Confidence interval for a mean with Student's t distribution

What to do when the population SD (σ) is unknown ?

When the SD of the population is not known, the SD of the sampling distribution of the mean is not known exactly but can be estimated from the SD measured in the sample.

DF for estimating variance in the population is equal to $n-1$. Why n minus 1? Clearly if there is only 1 observation, there is no information about dispersion around the mean in the population. The number of degrees of freedom (DF) is zero. With $n=2$, there is 1 possible source of variability and $df=2-1=1$, etc...Thus: $df=n-1$.

$$\text{estimation of population Variance} = s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{(n-1)}$$

$$\text{estimation of standard error} = s / \sqrt{n}$$

In these conditions, $m - \mu / s / \sqrt{n}$ is not distributed normally because not only m but also s / \sqrt{n} fluctuates from sample to sample. The distribution is somewhat different from the Normal and is called the **Student's t distribution**. The t distribution depends on the number of degrees of freedom (df) of s , which are equal to the size of the sample minus 1 ($n-1$). The mean of any t distribution is 0 and its $SD = 1 + (2/(df-2))$ for $df > 2$, which indicates that the SD becomes closer and closer to 1 as the df get larger. Table A3 in Kirkwood gives the t values for different probabilities and different df s. The structure of the table is different to that of the Normal table: the values taken by the random variable (the t values) are inside the table, and not outside, on the periphery. Each line corresponds to a degree of freedom and each column to a probability. Notice that the table A3 provides both the unilateral and bilateral probabilities: the lower row gives the probability area in 2 tails, i.e. the one above the corresponding t value and the other below $-t$, whereas the first row of the table only gives the area in the tail above the t value. The t distribution is symmetrical, just like the z distribution. This makes that if we take the probability of having a value larger than a given t value in the table (e.g. for $df=30$ and $t=2.042$, $P=.025$ as indicated by "One sided P value"), this probability taken twice is the probability of having a t value larger than the tabulated value or lower than minus the tabulated value (in our example $P=.050$ as indicated in "Two sided P value"). Finally we see that the t values approach the z values when the df gets larger. With 30 df the first digit is generally identical. For 120 df the 2 first digits are generally identical. The values in the last row, with an infinite number of df , are the same as the z values.

The t table is used for the specification of the confidence interval when σ is unknown. The 95% confidence limits are:

$$m \pm t^{n-1} * s / \sqrt{n}$$

where t^{n-1} is the value in the Student's t table corresponding to n-1 df and to a bilateral P=.05.

example: mean CBF (Cerebral Blood Flow) is 98.44 with a SD of 3.066 in a sample of 13 subjects without cognitive neglect.

95% confidence limits are: $98.44 \pm 2.179 * 3.066 / \sqrt{13} =$
 $98.44 \pm 1.853 = (96.59 ; 100.29)$

99% confidence limits are: $98.44 \pm 3.055 * 3.066 / \sqrt{13} =$
 $98.44 \pm 2.598 = (95.84 ; 101.04)$

- **Confidence interval for a count**

If $E \geq 5$ Poisson \rightarrow Normal

95% Confidence Interval: $m \pm 1.96 * \sqrt{m/n}$

Example of Poisson variable: if the expected number of childbirth in a given hospital is 2300 per year; then the expected number per day is about $2300/365 = 6.3$. Application of Poisson formula gives (A graphical representation is provided in Fig.12)::

number of childbirth	probability
0	.0018
1	.0115
2	.0363
3	.0762
4	.1200
5	.1513
6	.1588
7	.1429
8	.1126
9	.0788
10	.0496
etc...

95% Confidence interval for year count

$$2300 \pm 1.96 * \sqrt{2300/1}$$

Limits are (2206; 2394)

precision is ± 94 childbirth a year

relative precision is $94/2300 = 4.1\%$

95% Confidence interval for day count

$$6.3 \pm 1.96 * \sqrt{6.3/365}$$

Limits are (6.04; 6.56)

precision is ± 0.26 childbirth a day

relative precision is $0.26/6.3 = 4.1\%$

Absolute precision is better for day vs. year, but **relative precision** is constant.

- **Confidence interval for a proportion**

Normal approximation for a Binomial distribution

if $n \cdot p$ and $n \cdot (1-p)$ are both equal or larger than 5

Binomial \rightarrow Normal

95% Confidence Interval: $p \pm 1.96 * \sqrt{p \cdot (1-p)/n}$

Confidence interval for a proportion: example in which Normal approximation is possible.

(from E. Truy, Vth Int. Cochlear Implant Conf., New York 1996; p.21): Obliteration of the cochlea was investigated with high resolution computer tomography (HRCT) in a sample of 101 candidates for cochlear implantation. Result showed partial or total obliteration of cochlea in 14 cases.

$$p = 14/101 = 13.86 \%$$

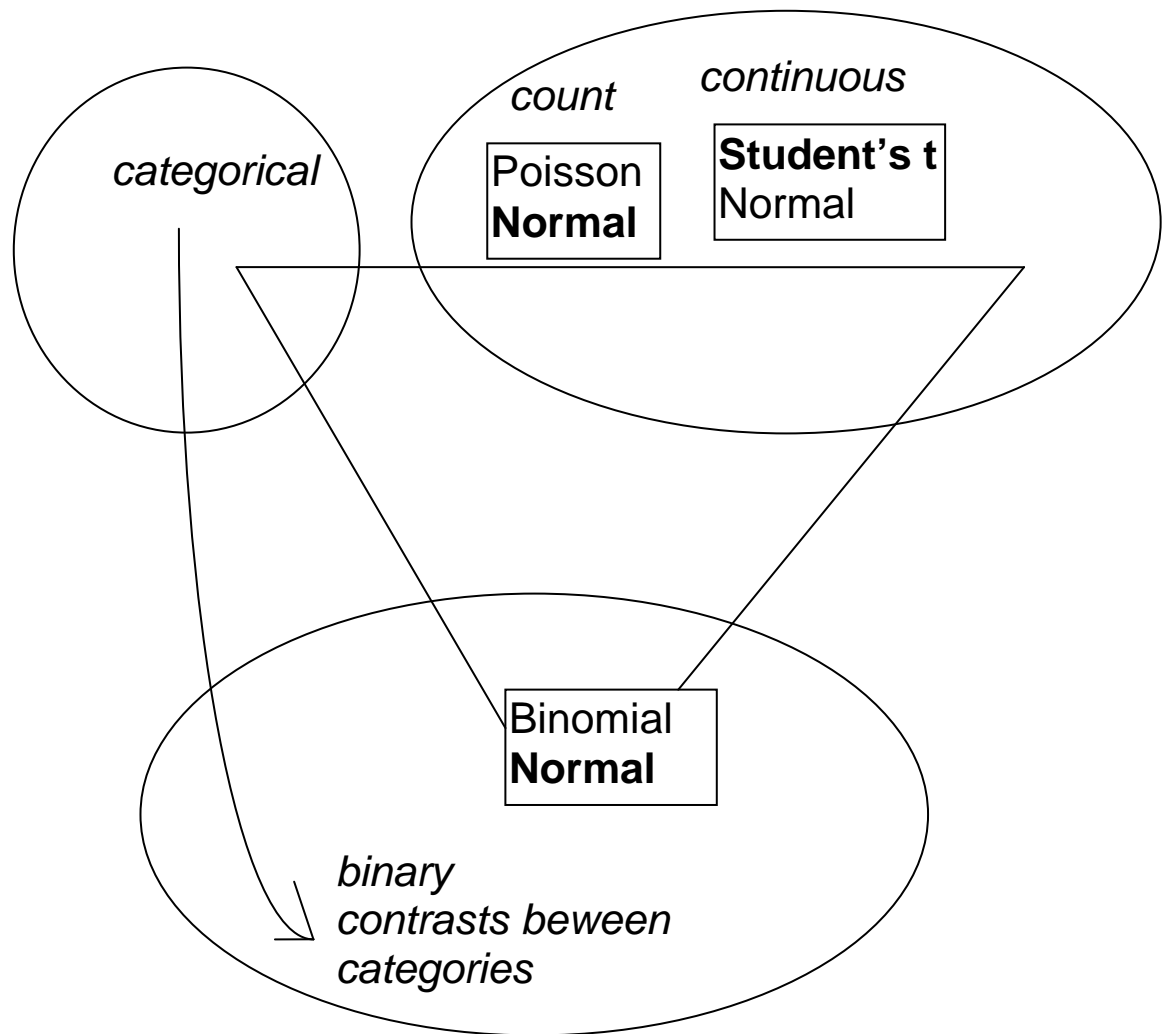
$$np = 14; n(1-p) = 101 - 14 = 87$$

np and $n(1-p)$ are both larger than 5, Normal approximation is possible.

$$\text{the 95\% Confidence Interval is: } 13.86 \pm 1.96 * \sqrt{13.86 * 86.14 / 101} = 13.86 \pm 6.74$$

Confidence Limits are: 7.12 ; 20.59

There is a 95% probability that the proportion of cochlea obliteration in the population of implantee candidates is in the 7 to 21% interval.



95% Confidence Interval:

continuous variable: $m \pm t(.95) * \sqrt{s^2/n}$

count (if $E \geq 5$): $m \pm z(.95) * \sqrt{m/n}$

proportion (if $E \geq 5$): $p \pm z(.95) * \sqrt{p*(1-p)/n}$

Chapter 5. Conformity tests for a single sample

- Null hypothesis, Significance test
- False positives and Confidence
- t-test for a mean
- False negatives and Power
- chi-square test for a count
- chi-square test for a proportion

● **Null Hypothesis and False positives (Type I Error)**

Purpose of statistical tests: What do sample measurements of contrasts, R and OR coefficients tell us about population values ?

The Null Hypothesis (H_0) says that the contrast is null in the population (that the difference between the sample value and zero is only due to chance).

The Null hypothesis is rejected if:

- the 95% Confidence Interval does not contain the H_0 value
- or just the same: the P value is lesser than 5% (.05).

When the H_0 is rejected we say that the test is significant and we give the P value (S at $p = .0\dots$).

When the H_0 is not rejected we say that the test is non-significant and we also give the P value (NS at $p = .0\dots$). The P value is the chance that the difference between m and H_0 value is due to random variation between samples.

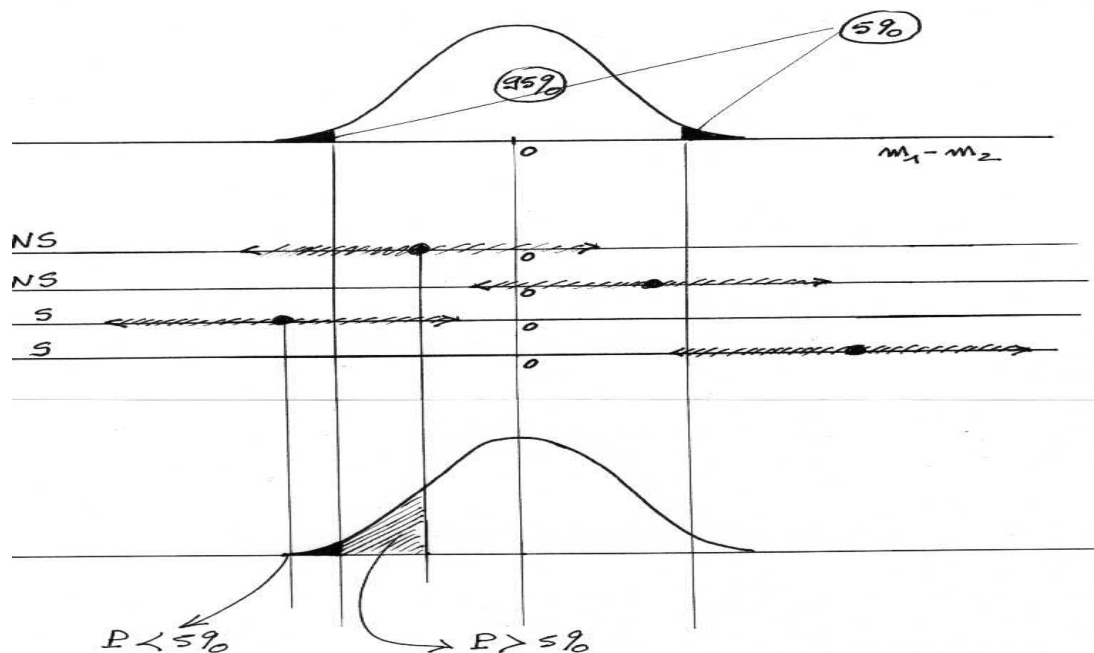


Figure 6.1 False positives (Type I error)

- **t-test for a mean**

Example: conformity of mean weight at birth with a population value.

Suppose that previous studies suggest that the mean weight at birth should be 3100 g. Is this compatible with the obtained mean of 3251 g in a sample of 41 babies, where $SD = 525$?

In other words:

- Is our sample extracted from the same population as before ?
- Is there a change in mean weight at birth ?
- Is the null hypothesis of 3100 g true ?

These questions lead to the following one in statistical terms:

What is the risk we take if we state that the difference between the sample mean and the our hypothesis on the population mean is NOT due to random fluctuations ?

Answer: if the true mean is really 3100, the risk is equal to the probability of obtaining a difference at least as large as $|3251 - 3100| = |151|$.

Calculation: given that population SD is unknown, the sampling distribution of m is Student's t with 40 df and

$$p(\text{difference} > |151|) =$$

$$p(t^{40}_{>\frac{|151|}{525/\sqrt{41}}}) =$$

$$p(t^{40}_{>|1.84|})$$

The table indicates that the probability is between .10 and .05, thus:

$p > .05$, the risk we take in concluding to a difference is higher than 5%.

There is more than 5 % chance that the difference between the mean weight at birth and the population value is due to random variation (sampling variation).

The issue of the test is not simply "no, the null hypothesis is not true" or "yes, the hypothesis is true"...

but between "reject the null hypothesis with a given risk" or "the hypothesis is compatible with a given risk".

How to conclude then ? BY CONVENTION, we decide NOT TO REJECT the population hypothesis if the risk is larger than 5%. We then say that the test is NON SIGNIFICANT (NS, $p > .05$).

Conversely, if the risk is lesser than (or equal to) 5% we decide to REJECT the population hypothesis. The test is SIGNIFICANT (S, $p \leq .05$).

The risk to reject the H_0 when in fact it is true is called "TYPE I ERROR" or α or P :

$$\alpha = p(RH_0/H_0 \text{ true}).$$

This risk is equivalent to the FALSE POSITIVE in diagnostic tests. However, whereas false positive probability in diagnostic test is the risk of deciding that a single subject is different from some norm when in fact he is not, type I error is the risk of deciding that the mean of a sample of subjects is different from some norm when in fact it is not.

• False negatives (Type II error) and Power

Type I error is the risk to reject a true hypothesis. This is the risk we take when the test is significant. But we also take a risk when the test is non significant. This is TYPE II

ERROR (or β) and it is defined as the probability of not rejecting the H_0 when in fact it is false:

$$\beta = p(\text{NRH}_0/\text{H}_0 \text{ false}).$$

This risk is equivalent to the a FALSE NEGATIVE in diagnosis tests. Again, whereas false negative probability in diagnostic test is the risk of deciding that a single subject is not different from some norm when in fact he is, type II error is the risk of deciding that the mean of a sample of subjects is not different from some norm when in fact it is.

An alternative hypothesis, labelled H_1 , must be specified for calculating the **Type II error**. The difference between H_0 and H_1 corresponds to the **precision** of the test. The higher the precision, the higher the **Type II error**.

There is a straightforward relationship between the **significance test** and the **confidence interval**. **Confidence** is the probability of not rejecting the H_0 when it is true with a given precision.

A 95% C.I. contains all the H_0 values which would not be rejected at $p=5\%$. This means for instance that a OR is not significant if its C.I. contains the value 1. A 99% C.I. contains all the H_0 values which would not be rejected at $p=1\%$. Etc...The confidence level $(1-\alpha)$ is the complement of the rejection level (α) .

The **Power** of a test $(1-\beta)$ is the complement of the type II error (β) . Power is the probability of rejecting the H_0 when it is false with a given precision.

	H_0 is true	H_0 is false
rejection of H_0 (significant test)	type I error type I error = α or p $p = P(\text{RH}_0/\text{H}_0 \text{ true})$	right issue Power = $(1-\beta)$ Power = $P(\text{RH}_0/\text{H}_0 \text{ false})$
non rejection of H_0 (non significant test)	right issue Confidence = $(1-\alpha)$ Confidence = $P(\text{NRH}_0/\text{H}_0 \text{ true})$	type II error type II error = β $\beta = P(\text{NRH}_0/\text{H}_0 \text{ false})$ $= P(\text{NRH}_0/\text{H}_1 \text{ true})$

Example of Type II error calculation . Return to the “cure everybody” drug (Chapter 1).

$H_0 : \pi = 1$

Data: 100% cured in a sample of 30

Sampling distribution: Binomial (calculations are simpler than with Normal approximation)

Alternative hypothesis

Type II error and Power

$H_1 \pi = .99$ $\beta = P(NR_{H_0} / H_1 \text{ true}) = P(p=100\% / \pi = .99) = .99^{30} \cong 0.74$; Power = 0.26

$H_1 \pi = .9$ $\beta = P(NR_{H_0} / H_1 \text{ true}) = P(p=100\% / \pi = .90) = .90^{30} \cong 0.04$; Power = 0.96

• Chi-square test for a count

The sampling distribution of a count is the **Poisson** distribution. Thus we should in principle make use of the Poisson distribution for testing an hypothesis on **a count**. However, with the Poisson distribution the procedure is rather tedious, especially for large samples. We can therefore use the **Normal** approximation, as we did for the confidence interval, provided that observed frequencies are equal or larger than 5 (Chapter 4). We can also use a **Pearson Chi-square** (symbol χ^2) test, which will give exactly the same results as the Normal test. The interest of the Chi-square is that it will be useful for further applications. Chi-square with only 1 DF (as for the present application) the same as a Normal variable squared (z^2)

Pearson Chi-square for counts

O is the observed count

Null hypothesis (H_0): E is the population count

Sampling distribution of

$$= \frac{O-E}{\sqrt{E}}$$

is a **Normal z** distribution . Remember that mean is the same as variance for a Poisson variable.

Sampling distribution of

$$= \frac{(O-E)^2}{E}$$

is a **Chi-square** distribution with **DF=1**

Example of conformity test for a count in which Chi-square (or Normal) approximation is possible.

Is a rate of 6.3 childbirth a day compatible with the 11 expected in the population ?
As 11 is larger than 5, Normal or Chi-square approximation can be used.

$$z = (6.3-11)/\sqrt{11} = -1.42$$

Table shows that test is NS ($p = .156$). Rate of childbirth in sample is not significantly lesser than in population.

$$\chi^2 = (6.3-11)^2/11 = 2.01$$

Table shows that test is NS ($p > .1$). Rate of childbirth in sample is not significantly lesser than in population.

• Chi-square test for a proportion

The sampling distribution of a proportion is the **Binomial** distribution. Thus we should in principle make use of the Binomial distribution for testing an hypothesis on a **proportion**. However, with the Binomial distribution the procedure is rather tedious, especially for large samples. We can therefore use the **Normal** approximation, as we did for the confidence interval, provided that observed frequencies are equal or larger than 5 (Chapter 4). We can also use a **Pearson Chi-square** (symbol χ^2) test, which will give exactly the same results as the Normal test. The interest of the Chi-square is that it will be useful for further applications.

Pearson Chi-square for proportions

p is the expected value in sample of size n ; thus np and $n(1-p)$ are the observed absolute frequencies O_i

Null hypothesis (H_0): P_0 is the population value; thus nP_0 and $n(1-P_0)$ are the expected absolute frequencies E_i

$$\text{Sampling distribution of} \quad = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

is a **Chi-square** distribution with **DF=1**

Example of conformity test for a proportion in which Chi-square (or Normal) approximation is possible.

(from E. Truy, Vth Int. Cochlear Implant Conf., New York 1996; p.21): Obliteration of the cochlea was investigated with high resolution computer tomography (HRCT) in a sample of 101 candidates for cochlear implantation. Result showed partial or total obliteration of cochlea in 14 cases. Is this compatible with an expected prevalence of 25% obliteration in deaf subjects ?

$$p = 14/101 = 13.86$$

$$np = 14; n(1-p) = 101 - 14 = 87$$

Expected values: 25.25 (25% of 101) and 75.75 (75% of 101)


25.25 and 75.75 are both larger than 5, Chi-square approximation is possible.

Observed values: 14 and 87

$$\chi^2 = (14 - 25.25)^2 / 25.25 + (87 - 75.75)^2 / 75.75 = 6.68$$

Table shows that test is S ($p < .01$). Prevalence in sample is significantly lesser than in population.

Chapter 6. Univariate significance tests for two or several samples

- 
- t-test and ANOVA for 2 means
 - ANOVA for several means
 - Contrasts for means
 - Chi-square test for 2 proportions
 - Chi-square test for several proportions
 - Contrasts for proportions
 - Non-Parametric tests

● test for a difference between 2 means

Purpose: is the difference between two sample means due to chance or do they come from different populations ?

In other words: is the relationship between a categorical and a quantitative variable due to chance or is there some relationship in the population ?

Null hypothesis (H_0) = the 2 population means are equal ($\mu_1 = \mu_2$)

Test Rationale: if population means are equal for the different levels (H_0) then the variance of level-means around the grand mean should only be due to random fluctuations (sampling variations). Variance between levels will then be of the same nature as variance of individual measurements within each level. Variance between levels will be smaller because it is the sampling variance of a mean (σ^2_x/n), but a simple relationship exists when this is taken into account.

If H_0 is true, then: $\sigma^2_m = \sigma^2_x/n$

$$n * \sigma^2_m = \sigma^2_x$$

$$n * \sigma^2_m / \sigma^2_x = 1$$

Estimations of between-category ($n * \sigma^2_m$) and within-category (σ^2_x) variances from the data.

Estimation of BETWEEN-category variance = $[n_1 * (m_1 - m)^2 + n_2 * (m_2 - m)^2]$

where n_1 n_2 are the sample sizes, m is the weighted grand mean. We take weighted mean because it is the best estimation of the population grand mean under H_0 . The null hypothesis says that data from the different categories are taken from populations with the same mean and each data should therefore equally contribute to the estimation of this common mean. This is obtained either by adding up the individual data or by weighting the category-means by the category-sizes.

Estimation of WITHIN-category variance = $s^2_x = [\sum(x_{i1} - m_1)^2 + \sum(x_{i2} - m_2)^2] / n - 2$

where x_{ij} are individual data, m_j are category means, n is the total sample size.

Sampling distribution of $\frac{\text{between category variance}}{\text{within category variance}}$

is a **Fisher F** distribution with **df = 1; n-2**

and

Sampling distribution of $\sqrt{\frac{\text{between category variance}}{\text{within category variance}}}$

is a **Student t** distribution with **df = n-2**

Application condition for F- test and t-test : within category variance should be the same for all categories.

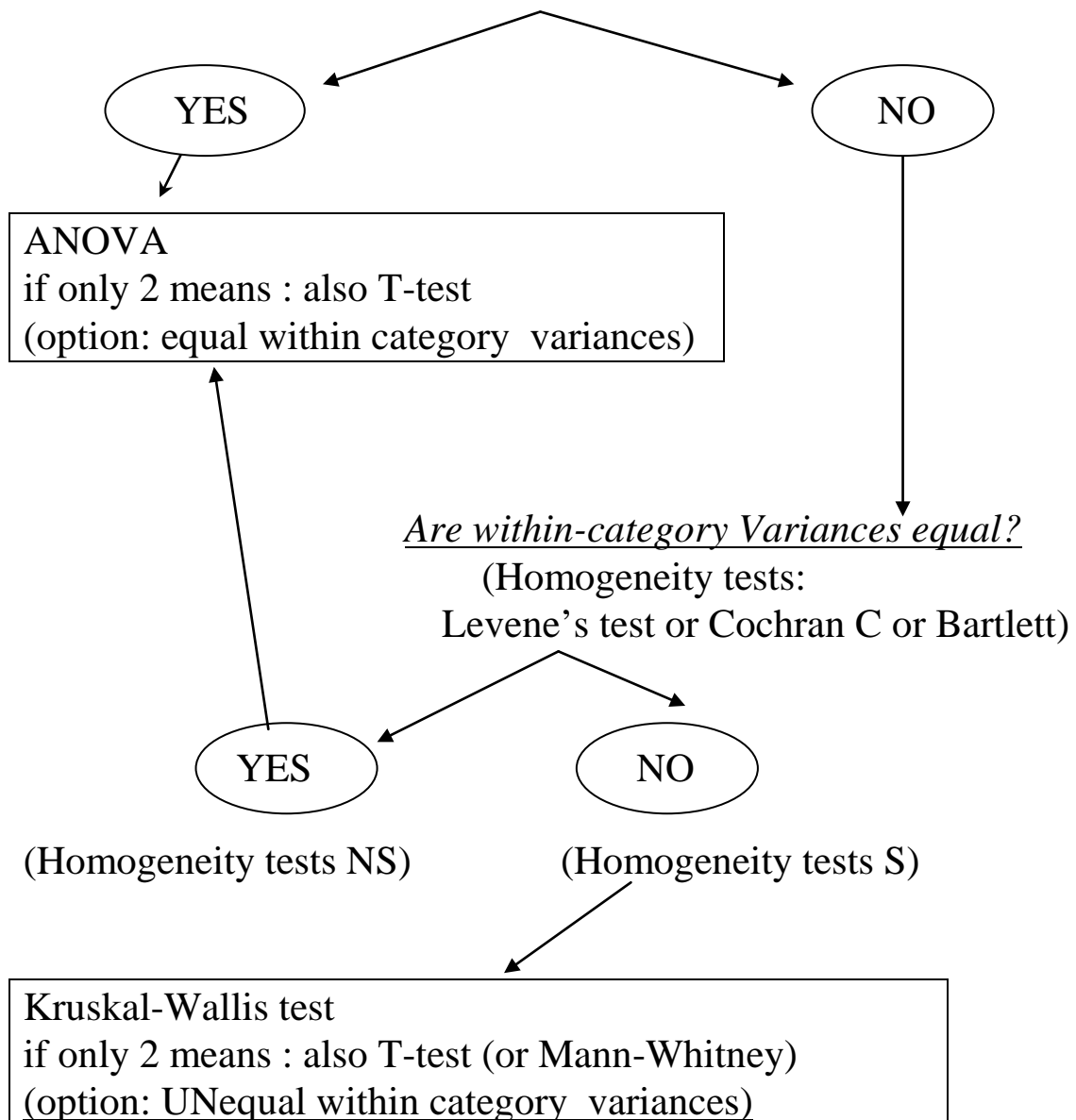
Before testing the significance of the between- versus within- variance difference we must verify that within-category variance is more or less the same for the different categories. If this is not true within-category variance estimation will depend on the number of observations per category. As the number of observations per category generally changes from study to study, this will make that estimation of within-category variance will also change. Therefore it is necessary to test the equality of within-category variances (of the category-components of the total within-group variance) with either a Cochran-C or a Bartlett-Box-F test, or a Levene's test (the latter is the most resistant to non-Normality). These "HOMOSCEDASTICITY" tests should be NON-significant for drawing exactly the right conclusions from the ANOVA. For small samples, ANOVA should only be applied if at least the B-Box is NON significant. For large samples, we can be more tolerant because even small differences between within-category variances are then significant. For equal-size categories, or small differences between size-categories (so long as the ratio of the largest to the smallest category-size is only about 1,5. Note¹), ANOVA can be applied whatever the issue of the homoscedasticity tests.

Alternative test: if homogeneity of variance tests are significant, use t-test for means with unequal variances.

¹ According to Hays (1988), p.373.

Test of a difference between means (one factor case)

*Are Sample Sizes fairly similar ?
(below limit ratio sample sizes 1.5)*



Example in which ANOVA test (or t-test) can be used for comparing two means:
effect of sex of newborn on weight at birth.

Data: $n = 749$

boys	$n_1 = 398$	$m_1 = 3317.61 \text{ g}$	$s_1 = 521.2753 \text{ g}$
girls	$n_2 = 351$	$m_2 = 3135.88 \text{ g}$	$s_2 = 516.9227 \text{ g}$

Calculations for F-test: see SPSS Output 6.1

Bartlett-Box test: NS ($p = .871$), we can use ANOVA.

$F(df=1,747) = 22.85$; S at $p < .0005$.

Difference in mean weight between boys and girls is indicated in SPSS Output “Estimates - Sex- parameter coeff”. The corresponding t-value (4.78) is the square root of the F value (22.85).

Conclusion: the 181.72 g weight difference between boys and girls is highly significant ($p < .0005$).

Calculations for t-test: see SPSS Output 6.2

Levene’s test: NS ($p = .803$), we can use “equal variance option”.

$t(df=747) = 4.78$; S at $p < .0005$.

Same conclusion.

Conclusion: S at $p < .0005$. There is less than .0005 chance that mean weight at birth difference between boys and girls is due to random variation (sampling variation).

Example in which ANOVA test (or t-test) can NOT be used for comparing two means: Application of t- test for unequal variances

Effect of smoking (no coded 0; yes coded 1) on pregnancy duration in a sample of 569 deliveries.

See SPSS Output 6.3.

Cochran's C and Bartlett-Box are both S ($p < .0005$) so we cannot use ANOVA, t-test for unequal variances should be used instead.

See SPSS Output 6.4.

Levene's test for equality of variances is S ($p = .042$), confirming the use of unequal variances option.

Conclusion: pregnancy duration is not significantly shorter for smokers ($p = .197$).

● ANOVA test for differences between several means

T-test can not be used for testing equality between several means. ANOVA test is then available and is calculated in much the same way as above. The only difference lies in the number of degrees of freedom.

Null hypothesis (H_0) = the k population means are equal ($\mu_1 = \mu_2 \dots = \mu_k$)

**Sampling distribution of between category variance
within category variance**

is a **Fisher F** distribution with **df = k-1; n-k**

Application condition for F- test: within category variance should be the same for all categories.

As above, use Cochran-C or a Bartlett-Box-F test for testing homogeneity of variance.

Alternative test: if homogeneity of variance tests are significant, use Non-parametric **Kruskal-Wallis test**.

Example: neurological data (Ref²) : relationship between Cerebral Blood Flow (CBF) and visuo-spatial neglect ?

NO NEGLECT		MODERATE NEGLECT		SEVERE NEGLECT	
SUBJECT	CBF	SUBJECT	CBF	SUBJECT	CBF
16,00	98,82	1,00	88,90	10,00	91,42
17,00	96,22	2,00	82,66	11,00	91,64
18,00	98,84	3,00	94,44	12,00	87,34
19,00	100,56	4,00	91,70	13,00	88,06
20,00	102,96	5,00	90,38	14,00	90,72
21,00	95,84	6,00	95,40	15,00	91,90
22,00	95,62	7,00	99,02		
23,00	92,96	8,00	90,86		
24,00	100,66	9,00	92,14		
25,00	100,24				
26,00	102,40				
27,00	94,92				
28,00	99,74				
m1	98,44	m2	91,72	m3	90,18
S1	3,07	S2	4,57	S3	1,97

² Demeurisse, G., Hublet, Cl., Paternot, J., Colson, C. and Serniclaes, W. (1997) "Pathogenesis of subcortical visuo-spatial neglect. A HMPAO SPECT study" *Neuropsychologia*. 35, 731-735.

Homogeneity of variance tests are NS (See SPSS Output 6.5). ANOVA can be used and shows that relationship between CBF and neglect is S ($F(df=2,25)=16.03$; $p<.0005$).

• **Example of Kruskal-Wallis test**

Relationship between pregnancy duration and environment (in 4 categories, from town center=1 to periphery =4). See SPSS Output 6.8.

As homogeneity tests are highly significant ($p<.0005$), we check the apparent significant relationship between pregnancy duration and environment ($p=.017$) with Kruskal-Wallis test.

Conclusion : relationship between pregnancy duration and environment is NS ($p=0.20$).

● **Contrasts between means**

If **ANOVA** is significant, not all the differences between means are necessarily significant. Tests of differences between individual means or between specific combinations of means called "**contrasts**" are then possible.

A "contrast" is any combination of means or proportions of the form:

$$\text{contrast} = \sum c_j m_j \text{ with } \sum c_j = 0$$

Examples:

$0.5 * m_1 - 0.5 * m_2$ is a contrast

$m_1 - m_2$ is a contrast

$2 * m_1 - m_2$ is NOT a contrast

Contrast coefficients: these are the c_j scaling values.

Contrast types:

"**DEVIATION-last**" contrast (default option)= each level of the factor except the last is compared to the grand mean

"**DEVIATION-first**" contrast = each level of the factor except the first is compared to the grand mean

"**SIMPLE-last**" contrast = each level of the factor except the last is compared to the last level

"**SIMPLE-first**" contrast = each level of the factor except the first is compared to the first level

"**DIFFERENCE**" contrast = each level of the factor except the first is compared to the mean of previous levels

"**HELMERT**" contrast = each level of the factor except the last is compared to the mean of subsequent levels

"**REPEATED**" contrast = comparisons between adjacent levels.

Example: with the CBF-neglect data taking simple-first contrasts allow to compare mean CBF for each of the two neglect categories to CBF of the no-neglect category (coded 1, hence first category). (see SPSS Output 6.5).

Contrast coefficients show that the -6.72 difference between mean CBF of neglect category 2 vs. 1 (moderate neglect vs. no neglect) is S ($p = .00014$).

The -8.26 difference between category 3 vs 1 (severe neglect vs. no neglect) is also S ($p = .00006$).

Comparison between the moderate and severe neglect are obtained by taking repeated contrasts (SPSS Output 6.6) which shows that the mean CBF difference between degrees of neglect is not significant ($p=.40573$).

● **options for confidence intervals when testing multiple contrasts**

There are 3 options for confidence intervals when testing multiple contrasts (all pairwise contrasts in SPSS-ANOVA, see “POST-HOC”):

“INDIVIDUAL” (default option): the p value is not corrected

“BONFERRONI”: the p value is corrected for the number of contrasts tested; this is achieved by taking a p-value corresponding to $.05 \times \text{number of comparisons}$ ($.05 \times 3 = .15$ in the CBF-neglect example)

“SCHEFFE”: the p value is corrected for testing all possible contrasts.

Example: changing the options for confidence intervals with the CBF-neglect data taking simple-first contrasts (see SPSS Output 6.7).

● **Chi-square and Fisher-exact tests for a difference between 2 proportions**

The relationship between two related categorical variables can be represented in frequency tables with one variable in line and the other in columns. Each variable can display two or several categories and the complexity of the table is measured by the number of possible sources of independent variation inside the table, or **degrees of freedom (DF)**. The 2 by 2 table with 2 categories for each variable is the simplest one. A 2 by 2 table has only 1 DF because for a given sample size and a given pattern of marginal frequencies, there is only one free frequency inside the table. Knowing the frequency of 1 out of the 3 inner cells allows to know the 3 other ones.

We saw that conformity of a proportion to a population value can be tested by a **Pearson Chi-square**. This test can also be used for testing differences between 2 proportions. Expected frequencies for this application of Chi-square are calculated from the data. They correspond to the frequencies that would be obtained in each sample if there were no differences between the two sample proportions.

Null hypothesis (H_0): $P_1 = P_2$

$$\chi^2_{(DF=1)} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i are observed frequencies and E_i are expected frequencies
 $E_i = \text{line total} \times \text{column total} / \text{grand total}$

Application condition: at least 80% of the expected frequencies (E_i) must be larger than or equal to 5

Alternative test: Fisher Exact Probability test.

Example: effect of sex of newborn on proportion of low-weight at birth (below 2.5kg).

See SPSS Output 6.9

Data: n= 1095

boys	n1 = 594	prop(lowweight) = 125 /594 = 21 %
girls	n2= 501	prop(lowweight) = 95 /501 = 20 %

No cells with expected frequencies lower than 5 (Min. E = 26.712). We can use the chi-square test. (otherwise we should have used the Fisher exact test, also given in SPSS output).

Chi-square (df=1) = .734 ; NS (p=.39)

Conclusion: There is more than 5 % chance that the higher proportion of low-weight at birth for girls vs. boys (10.0 vs 5.5 %) is due to random variation (sampling variation). Therefore we conclude that the proportion of low-weight births does not depend on sex.

● **Chi-square test for differences between several proportions**

Instead of a 2 by 2 table we now have a 2 by k table, where k is the number of samples. The complexity of the table is measured by the number of possible sources of independent variation inside the table, or **degrees of freedom** (DF). The 2 by 2 table with 2 categories for each variable is the simplest one. A 2 by 2 table has only 1 DF because for a given sample size and a given pattern of marginal frequencies, there is only one free frequency inside the table. Knowing the frequency of 1 out of the 3 inner cells allows to know the 3 other ones. A table with 2 lines and 3 columns (or just the same 3 lines and 2 columns) has 2 DF because it is possible to deduce all cell values from 2 out of them, not less. A general rule for calculating DFs is to take the following product:

$$DF = (L-1)*(C-1)$$

L= number of lines in table (say predictor's categories)

C= number of columns in table (say dependent variable categories).

For each DF there is a corresponding Chi-square, the formula remaining unchanged.. Only DF change which means that threshold value for significance gets larger as DF increase (check in Table).

Example of Chi-square between several proportions (from Colton p.179).

	A	B	AB	O	Total
THR+	32	8	6	9	55
THR-	51	19	5	70	145
Total	83	27	11	79	200

$$DF = (2-1)*(4-1) = 3$$

As number of cells with expected frequencies lesser than 5 is lesser than 20% (12.5%, see SPSS Output 6.10) Chi-square is applicable. Relationship between bloodgroup and throembolism rate is significant ($p < .001$).

• contrasts between proportions

Pairwise differences between proportions can also be tested by Chi-square. More generally, contrasts between proportions can be tested by Chi-square. Contrasts are defined exactly in the same way as for means:

$$\text{contrast} = \sum c_j p_j \text{ with } \sum c_j = 0$$

In SPSS, contrasts between proportions are not provided automatically in Crosstabs Command. We will see later (Chapt. 7) how to obtain automatic contrasts with the Logistic Regression command.

Chapter 7. Univariate Regression

- taxonomy of bivariate relationships
- Linear regression
- Non-parametric tests
- Logistic regression

• taxonomy of bivariate relationships

A common issue in scientific research is to see if there is a relationship between two or several variables. Consider the following examples with two variables.

Examples involving bivariate relationships.

- (I) Does weight at birth depend on sex ?
- (II) Does the rate of low-weight at birth depend on sex ?
- (III) Does systolic pressure depend on age ?
- (IV) Is low-weight at birth related to skull perimeter ?

Some of these variables are quantitative (weight at birth, systolic pressure, age). Others are categorical (sex, low-weight at birth). But in each example the question is to know whether one variable depends on the other. In statistical terms:

- (I) whether mean weight at birth is different for females vs. males;
- (II) whether proportion of low-weight at birth is different for females vs. males;
- (III) whether systolic pressure is correlated with age;
- (IV) whether low-weight at birth is correlated to skull perimeter.

Differences between means and proportions were treated in the previous chapter. In this chapter we will consider regression between 2 variables.

<i>variable types</i> (independent-dependent)	<i>description</i>	<i>statistical models</i>
categorical - quantitative	difference between means example (I)	• ANOVA (or t-test) for means
categorical - categorical	difference between proportions Odds Ratio (OR) example (II)	• Pearson Chi-square • -2LL Chi-square for logistic regression
quantitative - quantitative	correlation coefficient (R) example (III)	• ANOVA (or t-test) for linear regression
quantitative - categorical	Odds Ratio (OR) example (IV)	• -2LL Chi-square for logistic regression

• Linear Regression

Linear regression consists in predicting of the value of a quantitative variable with another quantitative variable with the help of a linear equation. In order to obtain a linear equation which provides the "best" description of the relationship between y and x (see Fig.7.1 **scatter** diagram), least square estimations of the **slope** of the line (symbol **b**, also called "**regression coefficient**") and of the **intercept** (symbol **a**) are taken. These LSE minimize the squared differences between the **observed** values (**y**) and the **predicted** values (**y'**).

The regression equation does not provide a measure of the strength of the linear relationship between the variables. The slope of the regression line cannot be used in this purpose because it depends on the variances. The slope (b gets smaller either when the variance of x gets larger or when the variance of y gets smaller. (In other words, the higher the s^2_x/s^2_y ratio the smaller the regression coefficient). The latter is thus not a good index of relationship between the variables because it depends on the units of measurement. The **correlation coefficient** (symbol R) provides a measure of strength of relationship which is independent of the variances. The R does not depend on measurement units and varies between -1 and +1.

+1 indicates a perfect linear relationship

0 indicates the absence of relationship

-1 indicates a perfect inverse linear relationship

The r does not give a proportionnal measure of relationship. This is given by taking **R²** (which is lower than r except for r=0 or r=1). R² is the proportion of variation of one variable which is explained by the other, or **proportion of explained variation**..

The R gives an index of LINEAR relationship. A strong curvilinear relationship is always possible even with r=0 .

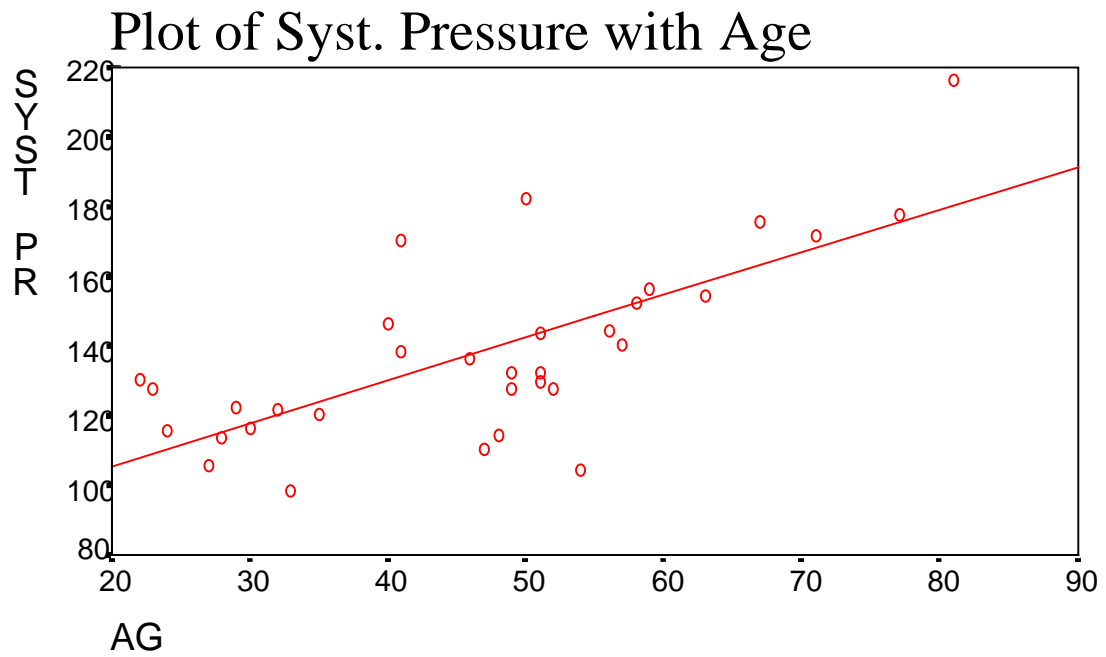


Fig.7.1 Relationship between age and blood pressure (from Colton, pp.191 & 192).

Linear Regression

Equation $y' = a + bx$

where y and x are two quantitative variables

y is the dependent variable

y' is the linear estimation of the “dependent variable”

x is the “independent” variable

Residual SS $\sum (y_i - y'_i)^2$

Explained SS $\sum (y'_i - m_y)^2$

Total SS $\sum (y_i - m_y)^2$

Least squares estimations

$$b = \frac{\sum (x_i - m_x)(y_i - m_y)/(n-1)}{\sum (x_i - m_x)^2/(n-1)} = \frac{\text{covariance (x,y)}}{\text{variance (x)}}$$

$$m_y = a + b m_x$$

$$a = m_y - b m_x$$

$$y' = m_y - b m_x + bx$$

$$y' = m_y + b (x - m_x)$$

Equation

Correlation Coefficient

$$R = \frac{\text{covariance (x,y)}}{\sqrt{\text{variance (x)} * \text{variance (y)}}}$$

Proportion of Explained variation

$$R^2 = \frac{\sum (y'_i - m_y)^2}{\sum (y_i - m_y)^2} = \frac{\text{explained SS}}{\text{total SS}}$$

Example: prediction of systolic pressure as a function of age (see Colton, p.189, for a sample of 33 women)

y = systolic pressure in mm Hg

x = age in years

y' = linear estimation of pressure from age.

b = 1.2 mm Hg per year of age

a = 81.5 mm Hg

predicted pressure = 138.6 + 1.2*(age - 46.7)

where 138.6 = mean pressure and 46.7 = mean age

r=.72 and r²=.52, which means that 52% of blood pressure variation is explained by age differences, and vice-versa.

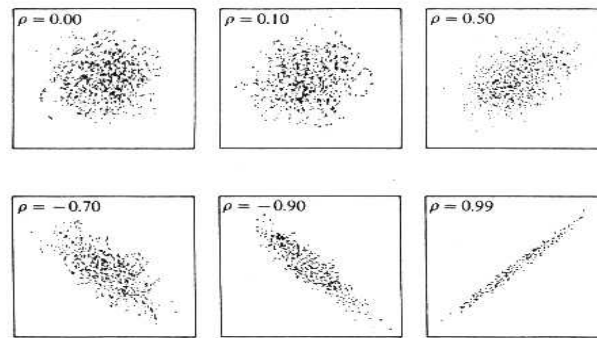


Fig. 2.4 Scatter plots showing examples of correlation. The scales and origin of the axes are irrelevant (see text) and are therefore not shown.

from R.J. Barlow "Statistics" J.Wiley, Chichester 1989,
p.16

Fig. 7.2: Correlation strength.

• test for Linear regression

Purpose: is the correlation due to chance only or is there some correlation in the population

t-test and ANOVA for correlation

Observed correlation = R

Null hypothesis (H_0) = $\rho = 0$

Sampling distribution of
$$\frac{R}{\sqrt{(1-R^2)/(n-2)}}$$

is a **Student's t** distribution with **df = n - 2**

Alternatively sampling distribution of
$$\frac{R^2}{(1-R^2)/(n-2)} = \frac{\text{explained SS}}{\text{residual SS}/(n-2)}$$

is a **Fisher F** distribution with **df = 1; n - 2**

Application condition for F test :

- Relationship should be fairly **linear**.
- Scatter distribution should be **free of deviant values** points located outside the bulk of the scatter diagram because they have larger effects than others on R value.
- Residual y variance should be fairly constant for the different x values: **no outliers** ("**homoscedasticity**" requirement).

Alternative nonparametric test

Kendall's Tau (τ) coefficient: is only based on ordinal information.

The H_0 is the absence of any relationship between rank orders when subjects are separately classified as a function of each variable. Then there is 50 %

chance for any two subjects to be classified in the same order on both dimensions.

The calculation of Kendall τ coefficient is based on the difference between the number of agreements and disagreements between classifications. For small samples (n smaller than 30), the type I error is obtained from specific sampling distributions (Siegel & Castellan, 1988, Tables R). For larger samples, the sampling distribution can be approximated by a Normal deviate (z value).

- Examples for Linear regression tests

Example 1: bloodpressure-age data

see SPSS output 7.1

Pearson $R = .72$ ($n=33$)

t value ($df=31$) = $(.72)/\sqrt{(1-.72^2)}/31 = 5.78$ (S at $p < .001$)

F value ($df=1; 31$) = $(.72^2)/(1-.72^2)/31 = 33.37$ (S at $p < .001$)

Application condition: scatter (see Chapter 4) shows relation is fairly linear, without deviant points. Residual analysis (SPSS output) does not reveal outliers more than 3 SD apart.

Example 2: data for which Pearson R is not relevant.

Correlation between cerebral bloodflow in two different cerebral regions ($n=28$).

See SPSS Output 7.2

Pearson $R = 0.44$ and is significant (S at $P = .018$) but this is due to the outlier with low CBF in both regions, outside the bulk of the data. Kendall $\tau = 0.19$ is NS ($p = .16$).

➤ Linear Regression: guidelines

The output of the correlation procedure can be distorted by: non-linearity, deviant cases, influential cases, colinearity (among others).

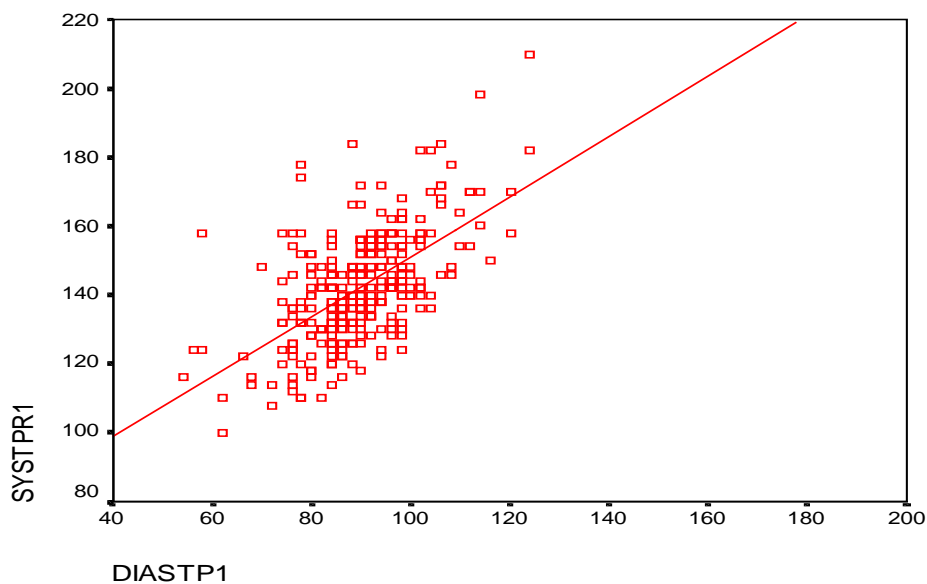
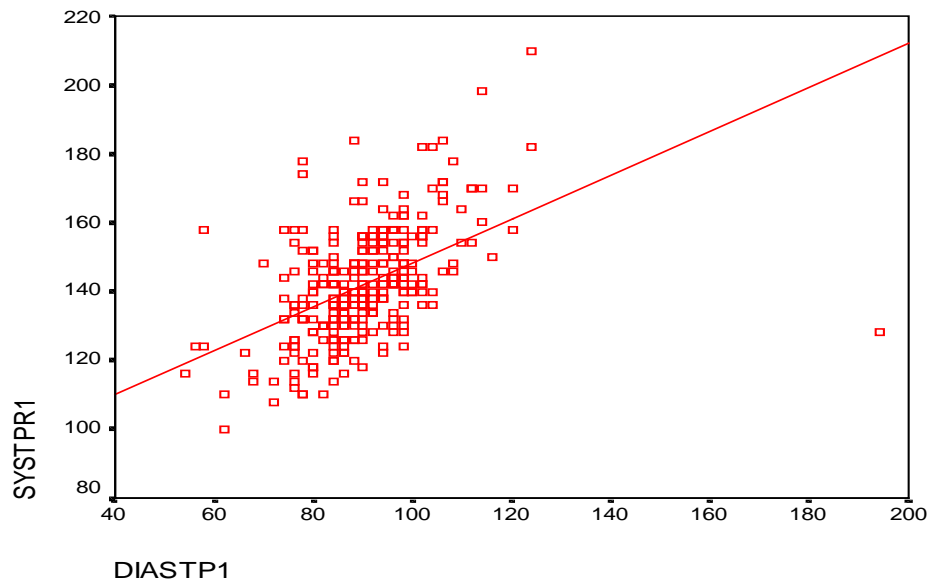
Non-linearity is detected by fitting the data with a quadratic function; a non-linear regression is suspected if the latter is visibly different from a straight line. Non-linearity can deflate the R. Solution: use a nonlinear transform of the X or Y variables (log, or exponential). Growth curves are typically exponential. Taking the logarithm of age allows then to linearize the regression.

Deviant cases are detected by examining the differences between observed and predicted values (or “residuals”) scaled in number of SD (Zresiduals). ZResiduals larger than 3 can either inflate or deflate the R, although their effect depends on the sample size (the larger the sample, the more extreme the residuals have to be for having substantial effects). An example is given below.

Influential cases can also inflate the R. They can be detected by comparing the residual with the “deleted residual”, which is the residual calculated for a case when it is not included. The case is influential if the difference is fairly large. Another way of detecting influential cases is to examine Cook’s distance, which considers the changes in all residuals when the case is omitted. Influential cases contribute to inflate the R. An example is given below.

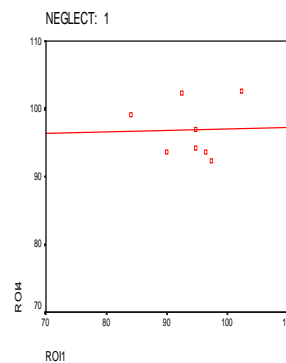
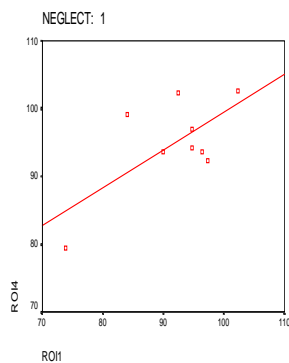
Colinearity makes the selection of two (or several) predictors hazardous. Colinearity is detected by a high R^2 between one of the predictors and the other ones, or just the same, by a low “Tolerance” ($1-R^2$). Colinearity means that different predictors explain much the same part of the variance of the independent variable. It can be avoided by taking only one of the correlated predictors (e.g.: only the number of living children, not gestity, parity ...). Another solution is to create a new variable which combines the different correlated predictors. The formula for the linear combination can be based on “Factorial Analysis”.

Case Number	Std. Residual	SYSTPR1	Predicted Value	Residual
96	-5.687	128	208.31	-80.31
250	3.079	178	134.52	43.48
262	3.273	210	163.78	46.22
272	3.053	184	140.88	43.12



Example of **deviant case**. Relationship between diastolic pressure and systolic pressure in a sample of 295 male adults. A deviant value appears in the original data (upper figure), seemingly due to an inversion between the two measurements. The deviant value is excluded in the lower figure. The R^2 increases from 0.24 with the deviant value included to 0.35 after excluding this value.

residual	deleted residual	Cook's distance
-0.29	-0.33	0.00
-5.50	-17.03	3.26
-2.38	-2.72	0.02
0.42	0.49	0.00
-3.93	-4.62	0.05
7.01	7.89	0.12
1.78	2.57	0.03
8.60	10.98	0.43
-5.73	-6.87	0.13



Example of **influential case**. Relationship between bloodflow in two brain areas in a sample of 9 patients suffering from cognitive neglect after stroke. An influential value appears in the original data (left figure). The value is quite visible on graph as well as in the table of residual values: the **deleted residual**¹ is much larger than the residual and **Cook's distance**² is large. This value is excluded in the lower figure. The R^2 decreases from 0.46 (S, $p=.046$) with the influential value included to 0.0007 (NS, $p=.95$) after excluding this value.

¹ The “**deleted residual**” is the residual calculated when the case is not included.

² The “**Cook's distance**” considers the changes in all residuals when the case is omitted.

• Logistic Regression

Logistic regression can be used for predicting of the value of a categorical variable with a quantitative variable. Regression is not linear in this situation. Let us take the case of single proportion corresponding to a binary variable. The relationship between the proportion and the quantitative variable is generally S-shaped.. The proportion changes slowly for extreme values close to either 0 or 100% and changes more and more rapidly as values get closer to 50% (Fig.7.3). S-shaped curves can be fitted either with **Logistic functions** or with **Cumulative Normal functions**. Any Logistic function can be transformed into a linear function by transforming the proportion into a **Logit**. Any Cumulative Normal function can be transformed into a linear function by transforming the proportion into a **Probit**. Logistic fitting is often preferred because Logistic equation is much more simple than the Cumulative Normal.

LOGISTIC function and LOGIT

$$P'(\text{disease} / x) = \frac{e^{y'}}{e^{y'} + 1}$$

where $y' = a + bx = \text{logit } P' = \ln [P'/(1-P')]$

Examples:

logit 0.5 = 0

logit 0.9 = 2.197

logit 0.1 = -2.197

logit 0.95 = 2.944

MAXIMUM LIKELIHOOD FITTING

The fit of the Logistic function to the observed proportions is based on “likelihood” calculations.

The likelihood of an observed proportion is the probability to find this proportion in a sample for a given theoretical value.

For example, with a theoretical value of .60 and a sample of size 10, the likelihood of any proportion p is the probability to obtain p in a sample of size 10 as given by the Binomial formula.

Thus:

$$\text{likelihood } (.7) = 10! / 7!3! (.6)^7 (.4)^3 = .2150$$

$$\text{likelihood } (.6) = 10! / 6!4! (.6)^6 (.4)^4 = .2508$$

The likelihood of an observed proportion gets larger when the theoretical proportion is closer.

If there is only one observed proportion, and if we are free to choose any theoretical value (suppose we just want to estimate the “best” population value for an observed proportion), then the best “likelihood” estimation of the theoretical value is simply the observed value (in the above example: .6). This value gives the highest possible likelihood.

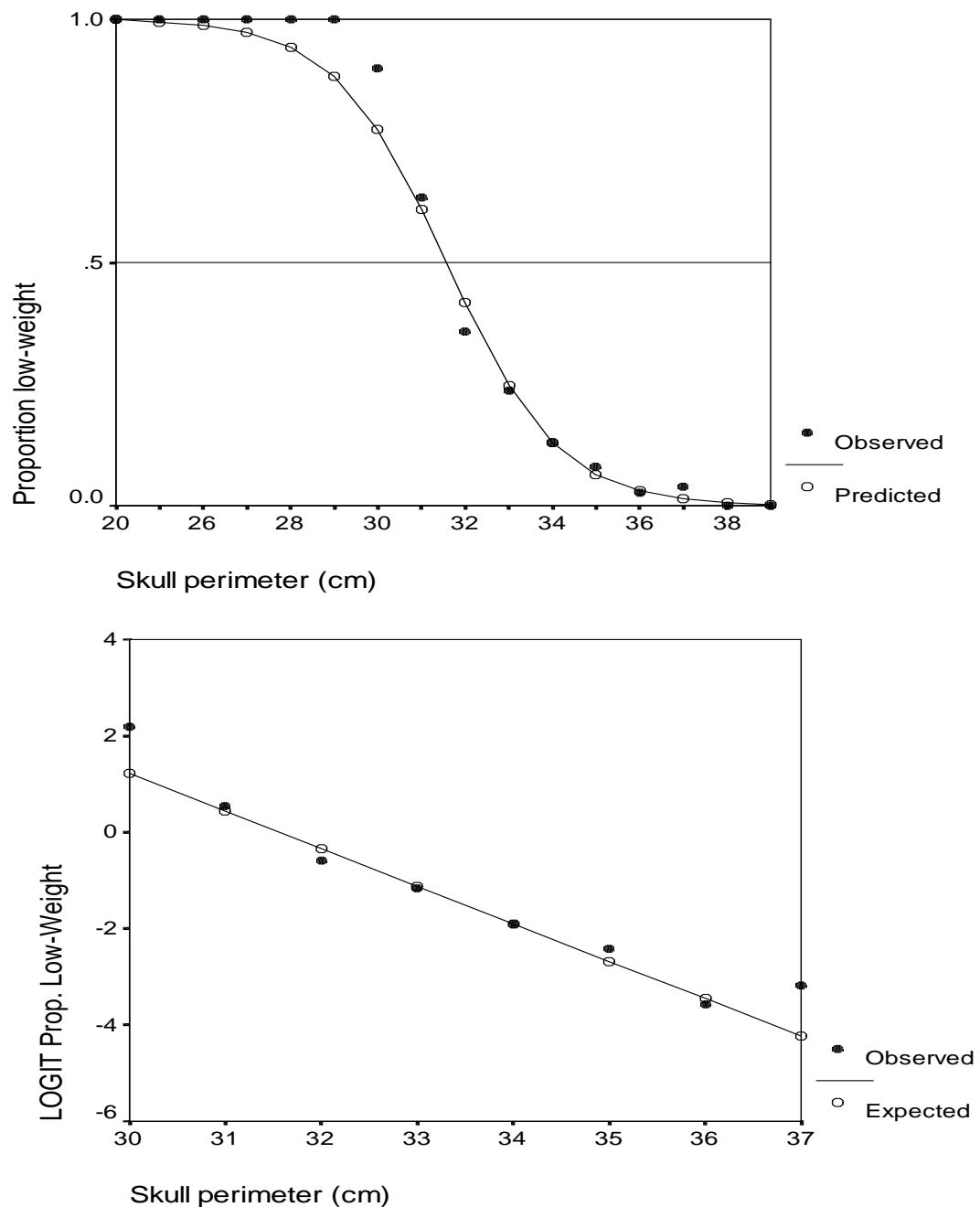
Now if there are several proportions and if we want estimations which are linked to some quantitative variable by a logistic curve, then the best possible “likelihood” estimations are those who give the highest joint likelihood.

Joint likelihood is simply the product on individual likelihoods:

$$\text{JOINT LIKELIHOOD: } L(P'_1 P'_2 \dots P'_n) = L(P'_1) * L(P'_2) \dots * L(P'_n)$$

Summary: fitting of a logistic curve to proportions as a function of some predictor (X) is achieved by calculating P' values which jointly maximize their likelihood.

Comparison with linear regression: fitting of a linear regression curve was obtained by maximizing explained variance (or minimizing residual variance). It can be shown that this procedure is a special case of maximum likelihood fitting, but which can only be used for quantitative dependent variables, not for proportions.



PROBABILITY of LOW WEIGHT at birth
as a LOGISTIC function of SKULL perimeter alone

$$P(\text{LOW WEIGHT}) = \frac{e^{Y'}}{e^{Y'} + 1}$$

where $Y' = 22.3 - 0.71 \cdot \text{skull} = 0.71 \cdot (31.4 - \text{skull})$
 Y' is the "LOGIT" of the Probability

Fig.7.3 LOGIT of P(LOW WEIGHT) at birth as a LINEAR function of SKULL perimeter alone

Logistic Regression tests with a quantitative predictor

Null Model (Model 0): Logit $p(y) = \alpha$
X₁ Model (Model 1): Logit $p(y) = \alpha + \beta_1 x_1$

Null hypothesis (H_0) = $\beta_1 = 0$

is tested by comparing likelihood for Model 1 with likelihood for Model 0. Likelihood is always better for Model 1 (because a predictor is included). But is the difference large enough for being significant ?

For assessing significance, we take the “improvement Chi-square”

$$= -2 \ln (L_0/L_1)$$

where

L₁ = Likelihood with Model 1

L₀ = Likelihood with Model 0

Sampling distribution of this expression is approx. **Chi-square** distribution with **df = 1** (as the model contains a single predictor).

Alternative test of improvement takes less computer processing time:

WALD approximate χ^2 test

Application condition :the logistic curve should fit the data. This is the case
“Hosmer-Lemeshow” Chi-square is not significant.

Example of logistic regression: example with a quantitative predictor: Effect of skull perimeter on weight at birth in two categories .

see Output 7.3

Hosmer-Lemeshow Chi-square is not significant ($p=.46$), so we can use the Logistic model.

Slope is negative ($b=-0.71$) which means that proportion low-weight at birth diminishes with increasing skull perimeter (presence of low-weight at birth is coded 1). OR = .492 ($=\exp(-.71)$) which means that odds of lowweight gets 49.2 % lower for each 1 cm increase of skull perimeter. As the constant is 22.3 the skull perimeter for which a 50% proportion low-weight is expected can calculated as follows:

50% is 0 in logits

$$0 = 22.3 - 0.71 * \text{skull}$$

$$\text{skull} = 22.3 / 0.71 \cong 31.4 \text{ cm (see Fig.7.3).}$$

Both Improvement Chi-square ($X^2(df=1)= 227$) and Wald test ($X^2(df=) =148$) are highly significant ($p <.001$).

Conclusion: relationship between skull perimeter and low-weight at birth is highly S ($p <.001$). As expected, proportion low-weight babies (below 2.5 kg) is inversely related to skull perimeter.

- **Logistic Regression tests with a categorical predictor**

Comparing proportions in 2 samples can also be handled by Logistic regression. Take the example of proportion decrease at birth as a function of sex. Sex, as any dummy variable, can be assigned two numerical values (0 and 1 for instance) and can then be treated as an elementary quantitative variable. We can therefore perfectly use Logistic regression for describing the relationship between a proportion (say rate of decrease) and a dummy variable (say sex). The model writes as follows:

$$\text{logit } p(\text{decrease}) = \alpha + \beta * (\text{Sex})$$

where sex takes values 0 or 1.

The interest of proceeding in this way is that it makes it possible to combine categorical and quantitative predictors within the same model (see Chapter 9). For the while, using the Logistic model for proportions has the practical advantage of giving automatic tests for contrasts between proportions in SPSS.

- **Coefficients**

Delta percentage, RR (constant = 2 here), OR the best one for research (see Fleiss, pp. 90...).

Up to some 10 %, OR (Logistic function) can be approx. by RR (logarithmic function).

Interest of RR: easier to understand, way open to other epidemiological coefficients (see RL).

P	Logit (P)		RR	loge OR	OR
50%	0	50% vs 25%	2	1,10	3,00
25%	-1,10				
20%	-1,39	20% vs 10%	2	0,81	2,25
10%	-2,20	10% vs 5%	2	0,75	2,11
5%	-2,94				
4%	-3,18	4% vs 2%	2	0,71	2,04
2%	-3,89	2% vs 1%	2	0,70	2,02
1%	-4,60				

- **Contrasts and Odds Ratios**

But there more about this. With a categorical predictor, the slope (β) is directly related to the Odds Ratio (O.R.).

Odds Ratio (OR)

$$OR = ad/bc = \frac{a/c}{b/d} = \frac{a/b}{c/d}$$

Relationship between ODDS RATIO and Logistic Regression coefficients

IF E is coded (0,1) difference between E+ and E- correspond to 1 unit

$$\begin{aligned} \ln(OR) &= \ln(a/b) - \ln(c/d) \\ &= \text{logit}(p(D+/E+)) - \text{logit}(p(D+/E-)) \\ &= \text{increase logit D for 1 unit increase of E} = \beta_E \end{aligned}$$

$$\text{Thus } \ln(OR) = \beta_E$$

$$OR = e^{\beta_E}$$

IF E is coded (-1,1) difference between E+ and E- correspond to 2 units

$$\begin{aligned} \ln(OR) &= \text{logit}(p(D+/E+)) - \text{logit}(p(D+/E-)) \\ &= \text{increase logit D for 2 units increase of E} \\ &= 2\beta_E \end{aligned}$$

$$\text{Thus } \ln(OR) = 2\beta_E$$

$$OR = e^{2\beta_E}$$

Example with a 2 by 2 table: association between perinatal health & mother's age

	diseased	healthy	
age ≤ 20 years	10	40	50
age > 20 years	15	135	150
	25	175	200

$$\text{O.R.} = 10 \cdot 135 / 40 \cdot 15 = 2.25$$

See SPSS Output 7.4 for comparing OR and slope of logistic curve with two SPSS commands (Crosstabs and Logistic Regression)

- **Contrasts for several proportions**

Just as for means, we can test different kinds of contrasts between proportions (see previous Chapter). Logistic regression allows to do this. Contrasts available in SPSS Logistic Regression command are the same as those in ANOVA General Factorial plus a further one: INDICATOR contrast. With indicator contrast any category can be taken as reference and pairwise differences between reference and all other categories are tested.

Contrast types: the same as for ANOVA-means plus “indicator” type which is specific to Logistic-proportions. Interest of indicator contrasts for categorical predictors with $DF > 1$ (more than 2 categories) because any category can be taken as reference (as “indicator”). This is not possible with simple contrasts, which can only take the last or the first category as reference.

“DEVIATION-last” contrast (default option)= each level of the factor except the last is compared to the grand mean

“DEVIATION-first” contrast = each level of the factor except the first is compared to the grand mean

“**INDICATOR**” contrast = each level of the factor except one taken as reference is compared to the reference

“SIMPLE-last” contrast = each level of the factor except the last is compared to the last level

“SIMPLE-first” contrast = each level of the factor except the first is compared to the first level

“DIFFERENCE” contrast = each level of the factor except the first is compared to the mean of previous levels

“HELMERT” contrast = each level of the factor except the last is compared to the mean of subsequent levels

“REPEATED” contrast = each level of the factor except the last is compared to the next level

Example of indicator contrast with several proportions: bloodgroup- throembolism (see SPSS Output 7.5). With bloodgroup AB which has the highest THR rate as indicator (group coded 3), contrasts are NS for bloodgroup A ($p=.3156$) and for B ($p=.1557$). Only contrast with bloodgroup O is S ($p=.0015$).

Chapter 8. Sampling Methods

8.1 Concepts and methods

There are two key concepts in estimation: **bias** and **precision**.

The primary requirement for the obtention of **unbiased** estimates is that the population sampled corresponds to the target population (see Introduction). Classical instances of bias are "Berkson's fallacy" for case-control studies carried exclusively in hospital and out-of-sight bias in follow-up studies (Kleinbaum et al., American J. of Epidemiology, 1981, 452-463). Berkson's fallacy is due to the fact that the risks of hospitalisation can combine within patients, which gives rise to a selection bias. Out-of-sight subjects in follow-ups do not occur independently of other characteristics, and can therefore also have filtering effects.

 Example of bias in the realization of a survey: Sample of households in Syracuse (USA) in 1930-1931: Distribution of households according to size, in the original sample and in the census tract. Households of one were not included in the survey (from Kiser (1934) in Yates, F.R.S. (1981) "Sampling Methods for Censuses and Surveys" 4th edition, High Wycombe, Bucks, England: Ch. Griffin; p.11)

Number in household	Original sample Number	%	Census tracts Number	%
2	254	19.4	1762	26.8
3	338	25.9	1745	26.5
4	307	23.5	1438	21.9
5	201	15.4	853	13.0
6	106	8.1	388	5.9
7	46	3.5	208	3.2
8	25	1.9	96	1.5
9 and more	29	2.2	86	1.3

As can be seen, the households with 2 members are underrepresented. This arises from the fact that women without

children are more often absent and enumerators did not return to these places to collect the information.

The **precision** of the estimates depends on their **variance** and on the **size of the sample**. Given the tradeoff between these 3 parameters, the size of the sample can be specified beforehand if the required precision is provided and variance is estimated on some other grounds.

Choice of the **method**:

Simple random sampling: each individual is extracted at random and independently of the others.

Systematic sampling: an individual is extracted at regular intervals.

Stratified sampling: the population is partitioned into groups, or "strata", and individuals are thereafter sampled within each stratum.

Cluster sampling: clusters of individuals, grouped as a function of spatial or temporal proximity, are sampled first and individuals are thereafter sampled within each cluster.

Although SRS provides a simple basis for theoretical developments, it also has several drawbacks. Some of these are practical:

- SRS requires the enumeration, and hence the identification, of all the units of the population. This is not possible if there is no available file containing all the units.

- the items of the sample can be largely dispersed, which is time consuming.

Other problems are of theoretical nature, in the sense that they can give rise to biases:

-some subset of the population, characterized by a specific feature which may affect the variable under study, may be underrepresented, or overrepresented, in the sample (because there is no direct control).

8.2 Simple random sampling

A sample of n elements extracted from a given population is a "SIMPLE RANDOM SAMPLE" if the extraction procedure is conceived in such a way that all the possible samples of n elements have the same probability to be extracted from the population.

Let us consider a finite population of N elements and a sample of n elements which are extracted without replacement from the population. The total number of possible samples of size n is:

$$T = C_N^n = N! / n! (N-n)!$$

and, for the sample to be simple and random, the probability of extracting a given sample of size n must be :

$$\frac{n! (N-n)!}{N!}$$

This condition will be fulfilled if elements are extracted at random and independently of each other from the population.

.....
Mathematical development

If each element of the sample is taken at random from the population, the probability of an element being selected is $1/N$ for the first, $1/(N-1)$ for the second, $1/(N-2)$ for the third..., $1/(N-n+1)$ for the last element. If the elements are extracted independently, the probability that the n elements are extracted in a specific order is, on the basis of the law for combination of independent events:

$$\frac{1}{N} * \frac{1}{N-1} * \frac{1}{N-2} * \dots * \frac{1}{N-n+1} = \frac{(N-n)!}{N!}$$

As there are $n!$ possible orders in which the same elements of the population can enter the sample, the probability of extracting a given sample of n elements is:

$$n! (N-n)! / N!$$

.....

Procedure for simple random sampling

A list containing the N elements of the population is constructed. This is the "SAMPLING BASIS". Each element is given a number from 1 to N . The n elements of the sample are extracted by using a table of random numbers or a computer with a random numbers generator.

Random number generation: see SPSS or EpiInfo.

8.3 Systematic Sampling

The idea is to subdivide the population into zones, a single item being extracted at random within each zone. The main advantage is that the sample is more uniformly spread over the population.

For a finite population, n zones each containing k items are created ($k = N/n = \text{"SAMPLING INTERVAL"}$). In each zone, the i th item is extracted, i being taken at random between 1 and k .

Example: Take a systematic sample of 9 students in a classroom of 27. The sampling interval is $27/9 = 3$. Take a number between 1 and 3 at random. Suppose 2 is taken. Extract students number 2, 5, 8, 11, 14, 17, 20, 23, 26.

Another advantage of systematic sampling is that the population need not to be known before the initiation of the sampling.

Examples:

In a study on intensive care, we can, for instance, decide to take one patient over 10 to enter an emergency room

Systematic sampling can be a source of bias in case of cyclic variations. This is especially the case if the zone size corresponds to the cycle size.

Example of bias with systematic sampling: in a comparative study on hospital work in different departments, a zone size of 7 days, from Sunday to Sunday, would certainly overemphasize the degree of activity of the emergency room.

8.4 Stratified sampling

8.4.1 A stratum is a subgroup of the population, in which the individuals share a common characteristic which is or could be related to the variable under study. After the population has been subdivided into strata, individuals are extracted at random within each stratum. This allows to control a possible biasing effect from the confounding characteristic. Indeed, random sampling alone does not guarantee that distribution of the confounding feature in the sample will be equivalent to its distribution in the population. On the contrary, sampling variation will also affect the confounder, which makes that the corresponding categories will be either underrepresented or overrepresented in the sample, to a degree which depends on chance. Subdividing the population into strata allows to neutralize the sampling variability of the confounding feature.

Example: In a study on the Reception of Patients in the hospital, individuals belonging to different communities are extracted separately from the population, in order to control the proportion of people from each community in the sample.

Different sampling methods can be used for extracting the items in each strata: simple random sampling, systematic sampling...

In the previous example: for each community, patients can be extracted at random and independently from the hospital file (SRS), or a specific proportion of those leaving the day after can be taken each day during one week (systematic sampling).

Two different approaches can be taken for specifying the number of items per stratum: equal or proportional allocation. Equal allocation means that the same number of items is taken for each stratum, which makes that the distribution of the confounding feature in the sample is generally not representative of the population. This method is preferable when the aim of the study is to make comparisons between strata, rather than obtaining an

estimation of the mean value across strata. Indeed, the statistical power of comparisons tests will be higher with equal sized samples.

On the contrary, proportionnal allocation should be used if the major objective of the survey is to provide a population estimation of a mean value, or of the proportion of some other attribute than the one used for stratification. Indeed, the practical task of collecting the sample and the formulas for estimating parameters will be simpler. Concerning the formulas, the samples will be "self-weighted", which means that weighting coefficients will not be required for the obtention of unbiased estimates.

In the previous example on the Reception of Patients; let us suppose that there are 4 communities with relative frequencies of 60 %, 20 %, 15 % and 5 %. In order to make comparisons between the degrees of satisfaction between communities, the 4 subsamples should have the same size (e.g. 500 for a total of 2000). On the contrary, if the primary aim of the study is to estimate the overall degree of satisfaction, the subsamples should be proportionnal to the population frequencies (e.g. 1200, 400, 300 and 100 for a total of 2000). This will also facilitate the practical realization of the survey.

8.4.2 Several simultaneous stratifications can be performed, each corresponding to a possible confounder. This is called "MULTIPLE STRATIFICATION".

 In the previous example on the Reception of Patients, the size of the hospital can also be taken into account. For practical reasons, hospitals can be grouped into 3 categories (according to the size) and the number of patients per category can be taken to be proportionnal to relative number of beds per category. If these are of 30% for small or mean sized hospitals and of 40% for large sized hospitals, then the sharing out of a total sample size of 2000, with proportionnal allocation between hospital type and equal allocation between communities, is as follows:

Community	A	B	C	D	Tot.
small hosp.	150	150	150	150	600
mean hosp.	150	150	150	150	600
large hosp.	200	200	200	200	800
total	500	500	500	500	2000

8.5 Cluster Sampling

The primary interest of cluster sampling is for investigating a population which is highly dispersed. Instead of directly taking individuals at random from the whole population, it is first divided into mutually exclusive and exhaustive clusters (for example: towns and villages). A simple random sample of clusters is then extracted and individuals are thereafter taken within the selected clusters.

Several methods are available for extracing the individuals in the selected clusters. In "ONE STAGE" cluster sampling, all the individuals of the cluster are taken into the sample. In "TWO STAGE" cluster sampling, individuals are sampled out with the help of one of the previous methods (SRS, systematic sampling, stratification). More than two stages can be involved. For instance, sampling units could be villages at the first stage, blocks of houses within the selected villages at the second stage, households within the blocks at the third stage. For two stage sampling, the number of individuals within each cluster can either be fixed, or taken with a "PROBABILITY PROPORTIONAL TO SAMPLE SIZE" (PPS). (This is the same as for stratified sampling: equal or proportionnal allocation).

 Example from Kaamugisha, J., and Feksi, A.T. (1988) "Determining the prevalence of epilepsy in the semi-urban population of Nakura, Kenya, comparing two independent methods not apparently used before in epilepsy studies" Neuroepidemiology 7, 115-121.

See Method I, pp. 116-117.

- creation of 30 clusters of about 100 households each
- starting list of all pupils in the first year of primary education. Total number of pupils is 3043. Each pupil is represented by a number, name and school name
- 30 pupils are selected by systematic sampling. Sampling interval is $3043/30 \cong 113$. First pupil is selected by SRS. Second by adding 113 to first pupil's number, etc...

- starting household for each cluster is adjacent to the one of selected children. A total of about 99 households are taken in each cluster.

8.6.Specification of sample size

The number of observations to be collected depends on the aim of the study: to provide an estimation of the population value or to test a specific hypothesis.

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple Random sampling

VARIABLE: Qualitative (proportion)

POPULATION: Infinite

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: 1

Starting from the formula of the confidence interval, and supposing that one has some preliminary idea about the value of the proportion in the population (Π). Otherwise, to put things at worse take $\Pi = .5$.

$$p \pm z_{\alpha} * \sqrt{\frac{\Pi(1-\Pi)}{n}}$$

The precision, d = half of the width of confidence interval =

$$z_{\alpha} * \sqrt{\frac{\Pi(1-\Pi)}{n}}$$

and for a given precision, the sample size is:

$$n = \frac{z_{\alpha}^2 * \Pi(1-\Pi)}{d^2}$$

Example: suppose you know the rate of arthritism among women aged between 50 and 60 years is around 20 % in a given region. How many subjects should you take in order to estimate the rate of arthritism with a precision of 2% ?

$$n = \frac{(1.96)^2 * (.2) * (1-.2)}{(.02)^2} = 1537$$

for a precision of 10%

$$n = \frac{(1.96)^2 * (.2) * (1-.2)}{(.10)^2} = 61$$

without any idea of the true proportion

$$n = \frac{(1.96)^2 * (.5)^2}{(.10)^2} = 96$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple Random Sampling

VARIABLE: Qualitative (proportion)

POPULATION: Finite (size N)

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: 1

Starting from the formula of the confidence interval,

$$p \pm z_{\alpha} \sqrt{\frac{\mathbf{N - n}}{\mathbf{N - 1}} * \frac{\pi(1-\pi)}{n}}$$

where the expression in bold characters is the "**finite population correction**".

If d is the required PRECISION then:

$$d = z_{\alpha} \sqrt{\frac{\mathbf{N - n}}{\mathbf{N - 1}} * \frac{\pi(1-\pi)}{n}}$$

and:

$$n = \frac{N z_{\alpha}^2 \pi(1-\pi)}{N d^2 + z_{\alpha}^2 \pi(1-\pi)}$$

Notice that (1)for a very large population, this formula becomes almost the same as the previous one, for infinite populations (proof: divide each term by N);

(2)sample size is lesser for a finite than for an infinite population.

Example: suppose you want to estimate the proportion of smokers for the 500 medical doctors working in a hospital, without any idea of the true proportion. How many people should you take for a precision of 10 % ?

Taking the largest possible variance, which corresponds to a proportion of one half:

$$n = \frac{500*(1.96)^2*(.5)^2}{500*(.1)^2 + (1.96)^2*(.5)^2} = \text{about } 81$$

With EPI program (Statcalc/ sample size/ Population Survey)/

Population Survey or Descriptive Study Using Random (Not Cluster) Sampling

Population Size : 500

Expected Frequency : 50.00 %

Worst Acceptable : 40.00 %

Confidence Level	Sample Size
-----	-----
80 %	38
90 %	60
95 %	81
99 %	125
99.9 %	176
99.99 %	215

Formula : Sample Size = $n / (1 - (n / \text{population}))$
 $n = Z * Z(\Pi(1 - \Pi)) / (D * D)$

Reference : Kish & Leslie, Survey Sampling, John Wiley & Sons, NY/ see Epiinfo manual p.258.

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple random sampling

VARIABLE: qualitative (proportion)

POPULATION: Infinite

AIM OF THE STUDY: Test of Hypothesis

NUMBER OF SAMPLES: 1

Sample size for hypothesis testing:

The investigator has to define the values of the 3 following parameters at the start:

type I error (α)

type II error (β) and hence the minimal deviation from the null hypothesis (Π_1 versus Π_0) to be detected by the test. Usually, type I error is considered to be 4 times as serious as type II error (Cohen, J.(1977) "Statistical Power Analyses for the Behavioral Sciences" New York: Academic press). Hence $\beta = 4\alpha$, and for $\alpha=.05$, $\beta=.20$ ($z_\beta=.84$, caution unilateral !)

the SD of the population = $\Pi_0 (1-\Pi_0)$.

With these ingredients in hand, the number of observations is :

$$n = \frac{[|z_\alpha| \sqrt{\Pi_0 (1-\Pi_0)} + |z_\beta| \sqrt{\Pi_1 (1-\Pi_1)}]^2}{[\Pi_0 - \Pi_1]^2}$$

where z_α corresponds to α in the bilateral Normal table (if two tailed test)

and z_β corresponds to β in the unilateral Normal table (even for two tailed test).

.....
Mathematical development: if we take the m value which corresponds to the threshold of rejection, we have:

$$z_\alpha = \frac{|p - \Pi_0|}{\sqrt{\Pi_0 (1-\Pi_0)/n}}$$

$$z_\beta = \frac{|p - \Pi_1|}{\sqrt{\Pi_1 (1-\Pi_1)/n}}$$

Eliminating m by combining these equations gives the above formula.

.....

 Example: the spontaneous rate of recovery is of .20 for a given disease (Π_0) and we want to detect a +.10 difference at least with type I error equal to .05 and type II error equal to .20. How many observations are required?

$$n = \frac{(1.96\sqrt{(.2)(.8)} + 0.84\sqrt{(.3)(.7)})^2}{.1} = \text{about } 137$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple random Sampling

VARIABLE: Qualitative (Proportion)

POPULATION: Infinite

AIM OF THE STUDY: Test of Hypothesis

NUMBER OF SAMPLES: 2

The size of one of the two samples, the smallest one if the 2 samples do not have the same size, is:

$$n_1 = \frac{[z_{\alpha/2} \sqrt{(r+1)\pi(1-\pi)} + z_{\beta/2} \sqrt{r\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}]^2}{r(\pi_2 - \pi_1)^2}$$

π stands for $(\pi_1 + r\pi_2)/(r + 1)$

This is an approximative formula (for the exact one see Fleiss p.44; also in Epiinfo, "cohort or cross-sectional" option and equivalently with "unmatched case-control"). r is for taking account for unequal sample sizes and is equal to n_2/n_1 . For equal sample sizes $r=1$ (see formula p.14- Table 7 in Lemeshow et al., 1990)

 Example: In a research on the Reception of Patients from two different ethnic communities in Hospitals which sample sizes should be taken for assessing the differences in proportions of "satisfied or not" with a precision of .1 if one community is 3 times larger than the other, and assuming that, if there is a difference, satisfaction will be 10% higher in the largest community ?

$$n_1 = \frac{[1.96 \sqrt{4(.575)(.425)} + .84 \sqrt{3(.5)(.5) + (.6)(.4)}]^2}{3 (.1)^2}$$

= about 256

This is for the smallest community. For the largest $n = 256*3 = 768$, and the total is about 1024. (With the exact formula - Epiinfo- 272 for the smallest sample and total=1088).

then either COHORT... or UNMATCHED-CASE CONTROL

Values to be entered with the above example: 95%, 80%, 3 /1, 50%, 0, 60%

Then for calculation and saving: F5/ F4/ F6/ F10 ...

Unmatched Cohort and Cross-Sectional Studies (Exposed and Nonexposed)

Sample Sizes for 50.00 % Disease in Unexposed Group

Conf.	Power	Unex:Exp	Disease in Exposed	Risk Ratio	Odds Ratio	Sample Size		Total
-----	-----	-----	-----	-----	-----	-----	-----	-----
95.00 %	80.00 %	3:1	60.00 %	1.20	1.50	816	272	1,088
90.00 %	"	"				651	217	868
95.00 %	"	"				816	272	1,088
99.00 %	"	"				1,197	399	1,596
99.90 %	"	"				1,734	578	2,312
95.00 %	80.00 %	"				816	272	1,088
"	90.00 %	"				1,077	359	1,436
"	95.00 %	"				1,320	440	1,760
"	99.00 %	"				1,845	615	2,460
"	80.00 %	1:1				407	407	814
"	"	2:1				612	306	918
"	"	3:1				816	272	1,088
"	"	4:1				1,020	255	1,275
"	"	5:1				1,225	245	1,470
"	"	6:1				1,428	238	1,666

Formula : $m' = \frac{\{c(a/2) \cdot \sqrt{(r+1) \cdot PQ} - c(1-b) \cdot \sqrt{r \cdot P_1 Q_1 + P_2 Q_2}\}}{(r \cdot \sqrt{P_2 - P_1})}$

$m = .25m' \cdot \sqrt{1 + 2 \cdot (r+1) / (m' \cdot r \cdot \text{Abs}[P_2 - P_1])}$

Reference : Fleiss, "Statistical Methods for Rates and Proportions",
2nd Ed., Wiley, 1981, pp. 38-45.

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple Random Sampling

VARIABLE: Quantitative (mean)

POPULATION: Infinite

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: 1

Starting from the formula of the confidence interval, with some idea about the population variance (σ^2):

$$m \pm z_{\alpha} * \frac{\sigma}{\sqrt{n}}$$

if d is the PRECISION required, then

$$d = z_{\alpha} * \sigma / \sqrt{n}$$

and:

$$n = \frac{z_{\alpha}^2 * \sigma^2}{d^2}$$

 Example: suppose you want to estimate the systolic pressure for males aged between 18 and 22, with a precision of 10 mm Hg, assuming a standard deviation of 15 mm Hg. How many subjects should you take ?

$$n = \frac{(1.96 * 15)^2}{(10)^2} = \text{about 9 subjects}$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple random Sampling

VARIABLE: Quantitative (mean)

POPULATION: Finite (size = N)

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: 1

Starting from the formula of the confidence interval,

$$m \pm z_{\alpha} * \sqrt{\frac{\mathbf{N - n}}{\mathbf{N - 1}}} \sigma^2/n$$

where the expression in bold characters is the "**finite population correction**".

if d is the PRECISION required, then:

$$d = z_{\alpha} * \sqrt{\frac{\mathbf{N - n}}{\mathbf{N - 1}}} \sigma^2/n$$

and:

$$n = \frac{N z_{\alpha}^2 \sigma^2}{N d^2 + z_{\alpha}^2 \sigma^2}$$

Notice that for a very large population, this formula becomes almost the same as the previous one, for infinite populations (Proof: divide each term by N).

 Example: suppose you want to estimate the B.M.I. with a precision of 2 for children below 10 years of age in a village where the corresponding population of children amounts 600, supposing a standard deviation of 3 ?

$$n = \frac{600 * (1.96 * 3)^2}{600 * (2)^2 + (1.96 * 3)^2} = \text{about 8 subjects}$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple Random Sampling

VARIABLE: Quantitative (mean)

POPULATION: Infinite

AIM OF THE STUDY: Test of Hypothesis

NUMBER OF SAMPLES: 1

The investigator has to define the values of the 3 following parameters at the start:

type I error (α)

type II error (β) and hence the minimal deviation from the null hypothesis ($\delta = \mu_1 - \mu_0$) to be detected by the test

the SD of the population (σ), a value which is generally unknown but can be estimated from previous investigations or with the help of a pilot study.

With these ingredients in hand, the number of observations is :

$$n = \frac{[\sigma(z_{\alpha} + z_{\beta})]^2}{d^2}$$

where z_{α} corresponds to α in the bilateral N table (if two tailed test)

and z_{β} corresponds to β in the unilateral N table (even for two tailed test).

and $d = \text{precision required} = \mu_1 - \mu_2$

.....
Mathematical development: if we take the m value which corresponds to the threshold of rejection, we have:

$$z_{\alpha} = \frac{|m - \mu_0|}{\sigma/\sqrt{n}}$$

$$z_{\beta} = \frac{|m - \mu_1|}{\sigma/\sqrt{n}}$$

Eliminating m by combining these equations gives the above formula.

.....

Example: weight of babies. Suppose that in a further study we want to detect a difference of +50 g from the present mean (3251 g) which we take as μ_0 . Let us take the SD of the previous sample (525 g) as estimate for σ , and choose a .05 value both for α and β .

$$n = \frac{(\sigma(1.96 + 1.65))^2}{50} = 1437 \text{ observations}$$

For a difference of +20 g, n= about 8980 observations.

For a difference of 300 g, n = about 40 observations.

For $\beta = 20\%$: n = 864

SPECIFICATION OF SAMPLE SIZE

METHOD: Simple Random Sampling

VARIABLE: Quantitative

POPULATION: Infinite

AIM OF THE STUDY: Test of Hypothesis

NUMBER OF SAMPLES: 2

The size of one of the two samples, the smallest one if the 2 samples do not have the same size, is:

$$n_1 = \frac{(r + 1) \sigma^2 (z_{\alpha/2} + z_{\beta/2})^2}{r d^2}$$

This is an approximative formula .r is for taking account for unequal sample sizes and is equal to n_2/n_1 , where n_2 is the size of the largest sample. For equal sample sizes $r=1$ (see formula p.39 in Lemeshow et al., 1990).

 Example: for $\sigma=15$, $d=5$ mm Hg, $\alpha=1\%$, $\beta= 4\%$, $r=1$

$$n_1 = \frac{2*(15)^2*(2.57+1.75)^2}{5^2} = \text{about } 337$$

SPECIFICATION OF SAMPLE SIZE - SUMMARY TABLE

SIMPLE RANDOM SAMPLING

PROPORTIONS

Finite Populations

Single sample

$$\text{Estimation} \quad n = \frac{N z_{\alpha}^2 \Pi(1-\Pi)}{N d^2 + z_{\alpha}^2 \Pi(1-\Pi)}$$

Infinite Populations

Single Sample

$$\text{Estimation} \quad n = \frac{z_{\alpha}^2 \Pi(1-\Pi)}{d^2}$$

Single sample

$$\text{Test} \quad n = \frac{[z_{\alpha} / \sqrt{\Pi_0(1-\Pi_0)} + z_{\beta} / \sqrt{\Pi_1(1-\Pi_1)}]^2}{[\Pi_0 - \Pi_1]^2}$$

Two Samples

$$\text{Test} \quad n_1 = \frac{[z_{\alpha} / \sqrt{(r+1)\Pi(1-\Pi)} + z_{\beta} / \sqrt{r\Pi_1(1-\Pi_1) + \Pi_2(1-\Pi_2)}]^2}{r(\Pi_2 - \Pi_1)^2}$$

n_1 = size of smallest sample

r = n_2/n_1

SPECIFICATION OF SAMPLE SIZE - SUMMARY TABLE

SIMPLE RANDOM SAMPLING

MEANS

Finite Populations

Single sample

$$\text{Estimation} \quad n = \frac{N z_{\alpha}^2 \sigma^2}{N d^2 + z_{\alpha}^2 \sigma^2}$$

Infinite Populations

Single Sample

$$\text{Estimation} \quad n = \frac{z_{\alpha}^2 \sigma^2}{d^2}$$

Single sample

$$\text{Test} \quad n = \frac{[\sigma(1/z_{\alpha} + 1/z_{\beta})]^2}{d^2}$$

Two Samples

$$\text{Test} \quad n_1 = \frac{(r + 1) \sigma^2 (1/z_{\alpha} + 1/z_{\beta})^2}{r d^2}$$

n_1 = size of smallest sample

$r = n_2/n_1$

SPECIFICATION OF SAMPLE SIZE

METHOD: Systematic sampling

VARIABLE: Qualitative or Quantitative

POPULATION: Finite or Infinite

AIM OF THE STUDY: Estimation or Test of Hypothesis

NUMBER OF SAMPLES: 1 or 2

Provided that the units of the population are ordered at random in the list from which systematic sampling was taken, then estimations for systematic sample is equivalent to those for SRS. Practically: do not assign a number to each item, otherwise there is no advantage versus SRS, but mix them first and then extract at random (Example: students mixed in the classroom).

SPECIFICATION OF SAMPLE SIZE

METHOD: Stratified sampling

VARIABLE: Qualitative (proportion)

POPULATION: Infinite

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: strata

Confidence interval:

$$p \pm z_{\alpha} \sqrt{\sum P_s * \frac{\sigma_s^2}{n}}$$

where P_s is the probability for an item in the population to belong to stratum s .

$$n = \frac{z_{\alpha}^2 \sum P_s * \sigma_s^2}{d^2}$$

 Example: suppose that a preliminary survey shows that the rate of arthritism among women aged between 50 and 60 depends on the region. The data are as follows:

Region	Population	rate of arthritism
A	65 %	around 20%
B	35 %	around 5%

How many subjects should be taken in order to estimate the rate of arthritism with a 2% precision?

$$n = \frac{(1.96)^2 [(.65)(.2)(.8) + (.35)(.05)(.95)]}{(.02)^2}$$

= about 1158

SPECIFICATION OF SAMPLE SIZE

METHOD: Stratified Sampling; proportional allocation

VARIABLE: Qualitative (proportion)

POPULATION: Finite (size N, strata of size N_S)

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: strata

Confidence interval:

$$p \pm z_{\alpha} \sqrt{\frac{(N-n)}{N} * \frac{\sum (n_S/n) \Pi_S(1-\Pi_S)}{n}}$$

where $\Pi_S(1-\Pi_S)$ is the variance per stratum.

Sample size:

$$n = \frac{N z_{\alpha}^2 \sum (N_S/N) \Pi_S(1-\Pi_S)}{z_{\alpha}^2 \sum (N_S/N) \Pi_S(1-\Pi_S) + Nd^2}$$

Example: suppose that a preliminary survey shows that the rate of arthritism among women aged between 50 and 60 depends on the region. The data are as follows:

Region	Population	rate of arthritism
A	55000	around 20%
B	30000	around 5%

How many subjects should be taken in order to estimate the rate of arthritism with a 2% precision?

$$n = \frac{85000(1.96)^2 [(55/85)(.2)(.8)+(30/85)(.05)(.95)]}{(1.96)^2 [(55/85)(.2)(.8)+(30/85)(.05)(.95)] + 85000(.02)^2}$$

= about 1140

SPECIFICATION OF SAMPLE SIZE

METHOD: Stratified Sampling; proportional allocation

VARIABLE: Quantitative (mean)

POPULATION: Finite (size N, strata of size N_S)

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: strata

Confidence interval:

$$m \pm z_{\alpha} \sqrt{\frac{N-n}{N} \frac{\sum (n_S/n) \sigma_S^2}{n}}$$

where σ_S^2 is the variance per stratum.

Sample size:

$$n = \frac{N z_{\alpha}^2 \sum (n_S/n) \sigma_S^2}{z_{\alpha}^2 \sum (n_S/n) \sigma_S^2 + Nd^2}$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Stratified Sampling; optimal allocation

VARIABLE: Quantitative (mean)

POPULATION: Finite (size N, strata of size N_S)AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: strata

"OPTIMAL ALLOCATION" is the sharing out of subjects in strata which will give the highest precision, for a given sample size, when the aim of the study is to make a global estimation for all the strata, and not to test differences between strata. If the variances of the strata are equal, the optimal allocation corresponds to PPS. Otherwise, the more general formula is as follows:

$$n_S = \frac{n N_S \sigma_S}{\sum N_S \hat{\sigma}_S}$$

Taking account of sampling cost per elementary unit differences between strata: if C_S is the cost per stratum, the global cost is:

$$C = \sum n_S C_S$$

and , for a given sample size:

$$n_S = \frac{n N_S \sigma_S / \sqrt{C_S}}{\sum N_S \sigma_S / \sqrt{C_S}}$$

For a fixed total cost C:

$$n_S = \frac{C N_S \sigma_S / \sqrt{C_S}}{\sum N_S \sigma_S / \sqrt{C_S}}$$

 Example: suppose measurements of BMI are taken in 2 strata, one for people living in town, the other for people in the countryside. 60 % of the population live in the countryside and

the cost per unit is 250 BF in town against 500 BF outside.
Which sample size should be taken for a global cost of 50000 BF?

$$n_{\text{countryside}} = \frac{50000(.6/\sqrt{500})}{(.6/\sqrt{500}) + (.4/\sqrt{250})} = \text{about } 68$$

$$n_{\text{stown}} = \frac{50000(.4/\sqrt{250})}{(.6/\sqrt{500}) + (.4/\sqrt{250})} = \text{about } 64$$

$$\text{Total cost} = 68*500 + 64*250 = 34000 + 16000 = 50000$$

SPECIFICATION OF SAMPLE SIZE

METHOD: Cluster Sampling; fixed or proportional allocation

VARIABLE: Quantitative (mean)

POPULATION: infinite

AIM OF THE STUDY: Estimation

NUMBER OF SAMPLES: clusters

sample size for cluster sampling = sample size for SRS * design effect

$$\text{design effect} = \frac{\text{variance with clustering}}{\text{variance without clustering}}$$

Epi-Info: how to obtain the design effect in a pilot study which can then after be used for calculating sample size in further studies.

Example: “Expanded Program on Immunization” (EPI : Lemshow & Robinson (1985) see Epi-Info manual, pp. 135...)

- data are in epil.rec file (vaccinal coverage, 30 clusters, 7 subjects per cluster, sample size =210)

- run CSAMPLE with epirec1.rec, main variable VAC ...- analysis TABLES

- see output $SE = 3.034 \% = \sqrt{p(1-p)/n} = \sqrt{0.7381*0.2619/210}$

this is the within-cluster SE calculated with the Binomial formula. Logic: if simple random sample (SRS) then Binomial distribution.

- run CSAMPLE with epirec1.rec, main variable VAC - **psu CLUSTER**- analysis TABLES
- see output $SE = 4.599 \%$

this is the observed between-cluster SE. Logic: if there was no cluster effect then between-cluster SE should be the same as within-cluster SE.

- design effect = $(4.599/3.034)^2 = 2.298$

- conclusion: sample size for cluster sampling of vaccinal coverage should be about 2.3 times larger than sample size for SRS, for obtaining estimates the same confidence and precision.

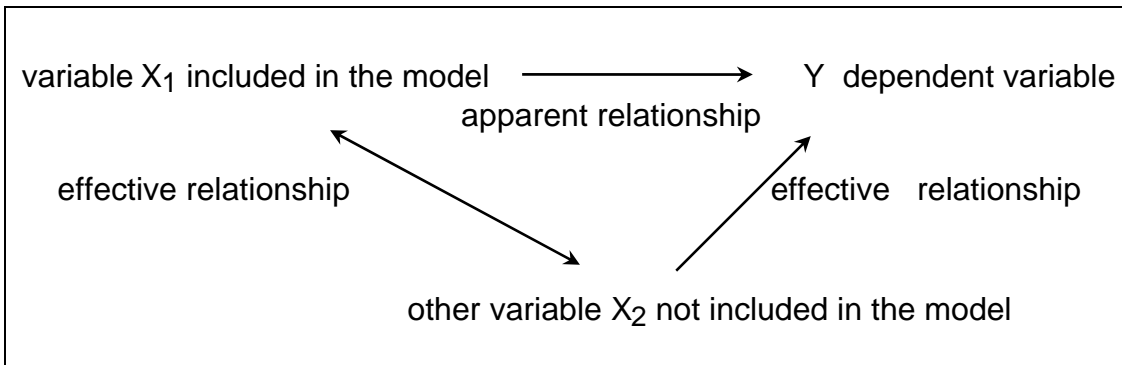
Chapter 9. Generalized Linear Model

- Generalized Linear model
- Risk Coefficients
- Contrasts
- Stratification
- Bias (confounder) & Interaction (effect modification)
- Multivariate model
- Guidelines for the choice of a test
- Strategies for model building
- TABLES

9.1 Confounding, Effect Modification and Stratification.

Taking several predictors allows to cope with BIAS (confounding) and INTERACTION (effect modification).

BIAS means that effect of a predictor is due to its relationship with another predictor. The effect will then disappear with stratification of the other predictor. INTERACTION means that the effect is not constant for the different strata of the other predictor.



The effect of a single predictor (X_1) on a dependent variable (Y) can be partially or totally due to the relationship between X_1 and one or several other possible predictors ($X_2, X_3 \dots X_k$). When this is the case, other possible predictors are then **confounding variables** or **confounders**. One of the purposes of multivariate statistics is to take account of possible confounders. Consider the following examples with three variables.

Examples of possible confounders.

(I) Does systolic pressure depend only on age or also on BMI which covaries with age?

Does the effect of age remain when BMI is taken into account? Taking account of BMI might reduce or even falsify the apparent effect of age. BMI is a possible confounder.

(II) Is weight at birth related to ethnic group?

We should take account of social class before drawing conclusions on weight differences between ethnic samples. If the proportions of babies from different social classes are not similar in the samples then we might attribute to ethnic differences what is in fact caused by social class differences. This latter factor would be a confounder.

(III) Height at birth might also be a confounder for the ethnic - weight relationship.

(IV) Is perinatal decease related to skull perimeter? Or does it in fact depend on a related variable, weight at birth.

(V) Is decease rate lower with drug vs. placebo? Or are subjects which were given the drug in better initial health than subjects in the “placebo group”.

(VI) Age might also be a confounder for the drug-decease rate relationship

In the simplest case (3 variables), stratification is describing the relationship between two variables for the different levels of a third variable. We are then “controlling” the possible confounding effect of the third variable.

Example (V) with proportions: effect of drug vs. placebo on recovery is described for each initial health condition. (SPSS command: crosstabs)

Example (II) with means: effect of ethnic group on weight at birth is described for each social class. (SPSS command: means)

If there is confounding (bias) then stratified description will reveal a weakening (or a strengthening) of the effect for each strata. For example (V) effect of drug will be weaker (or stronger) for each initial health condition taken separately. HINT: compare before / after stratifying.

If there is effect modification (interaction) then stratified description will reveal a difference in the strength of the effect between strata. For example (V) drug will have a larger effect for subjects in relatively good initial health condition than for those in bad initial condition. HINT: compare between strata.

9.2 Generalized Linear Model with a single DF

A linear equation can be used for representing any relationship between two variables, either categorical or quantitative. This is the Generalized Linear Model (GLM). GLM is very simple when the predictor (X) only has one degree of freedom (DF=1), that is for a single quantitative predictor or a binary categorical predictor. The GLM is then simply a way for representing differences between means or proportions with a linear equation, already used before for representing regression lines.

$$Y' = \alpha + \beta * X$$

The α coefficient is the “constant”.

For linear regression, α is the mean of the y values when x values are centered on the mean : $y' = \text{mean } y + \text{slope} * (x - \text{mean } x)$

For logistic regression, just the same as linear regression if x is quantitative:

α is the mean of the logit y values when x values are centered on the mean.

If x is categorical, α is the mean of the logit y values *if the categories are coded as deviation contrasts*.

For ANOVA, just the same as logistic regression with categorical x: α is the mean of y values *if the categories are coded as deviation contrasts*.

The β coefficient is a slope, or just the same, a contrast.

The X values are either genuinely quantitative (e.g. weight in kg) or dummy (binary categorical: e.g. sex coded as 0, 1).

Y' represents predicted values. It stands for individual values in a regression model (Linear or Logistic). Then there will usually be at least some difference between expected (Y') and observed (Y) values even if the contrast is highly significant. However when Y' stands for a mean there will be no difference with the observed value if the contrast is significant.

Consider the following examples for MEANS:

1) equation is trivial (predicted means are the same as observed means) if all the contrasts are significant. Consider the following examples with only one contrast (df=1).

Relationship between weight at birth and sex, with female coded 0 and male coded 1, deviation first contrasts (and female coded 0, male coded 1):

$Y' = \text{general mean} + (\text{mean category} - \text{general mean}) * X$

Predicted mean male = general mean + (mean male – general mean) * 1

Predicted mean female = general mean + (mean male – general mean) * (- 1)

An example with simple first contrast:

$$Y' = \text{general mean} + (\text{mean category} - \text{mean weight female}) * X$$

Predicted mean weight male =

$$\text{grand mean} + (\text{mean weight male} - \text{mean weight female}) * 0.5$$

Predicted mean weight female =

$$\text{grand mean} + (\text{mean weight male} - \text{mean weight female}) * (-0.5)$$

2) equation gives predicted values different from observed values if at least one contrast is NS. Consider the following case with only one contrast (df=1), supposing that weight difference between males and females is NS.

predicted mean weight female = grand mean

predicted mean weight male = grand mean

As we can see the effect of category difference on a mean or a proportion can be represented by a slope provide that a numerical value is assigned to each category. This is only possible for binary (dummy) categorical variables (see Chapter 2: equal interval requirement). Again we see the interest of dummy variables. We already saw that variance calculation makes sense with dummy variables. We now see that their effect on other variables can be quantified by slopes.

Interpretation of the slope (contrast) depends on the units used for describing the variables. This is evident not only for a quantitative predictors (slope is multiplied by 10 if skull perimeter is measured in cm rather than in mm) but also for categorical ones.

The magnitude of the slope will for instance be divided by 2 if male-female contrast is coded -1/+1 instead of 0/1. Notice that the coding will not depend on the original units but on the contrast type. Thus with deviation contrasts, the coding is -1/+1 and the slope is for half the increment between categories (for 1 unit increase, whereas the categories are 2 units apart). Then: $OR = e^{2\beta}$. For indicator or simple contrasts: $OR = e^{\beta}$.

The sign of the slope (contrast) will change from + to - if male-female contrast is coded 1/0 instead of 0/1. If two categorical variables are coded differently magnitude or direction of their effects might be judged according to different standards if we do not take care.

<i>variable types</i>	<i>Y</i>	<i>α</i>	<i>β</i>	<i>X</i>
quantitative - quantitative	values of quantitative dependent variable (in kg, years...)	intercept	slope of linear regression line	values of quantitative predictor (in kg, years...)
quantitative - categorical	logit of proportion in a category of the dependent variable	intercept	slope of logistic regression line	values of quantitative predictor (in kg, years...)
categorical - quantitative	values of quantitative dependent variable (in kg, years...)	global mean	contrast between means	numerical values assigned to predictor's categories
categorical - categorical	logit of proportion in a category of the dependent variable	global proportion	contrast between logits of proportions	numerical values assigned to predictor's categories

Risk Coefficients: which one and when?

	OR	RR
Basic reason for using	does not depend on prevalence ⇒ invariant	easier to understand & explain
Drawback	very abstract	depends on prevalence ⇒ if prevalence in the sample is arbitrary, sample RR does not represent population RR
Underlying mathematical function	Logistic function	Exponential function
Linear transform	Logit	Log
Examples of application	Case-control studies Cohort studies	Cohort studies
Approximation	$OR \cong RR$ if prevalence < 10%	

9.3 Generalized Linear Model with several DFs per predictor and several predictors.

$$Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 \dots$$

- when a predictor has several DFs there is one slope (contrast) per DF.
Examples: ANOVA for means in 3 samples (e.g. CBF- neglect), 2 slopes. Logistic Regression for proportions in 4 samples (e.g. bloodgoup- throembolism), 3 slopes.

- effects of different predictors can be represented into the same model. Each predictor is then represented by a number of slopes (contrasts) corresponding to its DF.

In the examples above:

- (I) effect of age and BMI on systolic pressure; one slope for each of the 2 quantitative predictors (Multiple Linear Regression);
- (II) effect of ethnic group (say 4 categories) and social class (say 5 categories) on weight at birth; 3 slopes for ethnic group, 4 slopes for social class (ANOVA for means);
- (III) effect ethnic group (say 4 categories) and height on weight at birth; 3 slopes for ethnic group, 1 slope for height (Analysis of Covariance);
- (IV) effect of skull perimeter and weight at birth on decease rate; 1 slope for skull per. 1 for weight (Multiple Logistic Regression);
- (V) effect of drug and initial health condition on decease rate; 1 slope for drug, 1 for health condition (Multiple Logistic Regression);
- (VI) effect of drug and age on decease rate; 1 slope for drug, 1 for age (Multiple Logistic Regression).

- an INTERACTION is introduced as the product of two (or more) predictors (e.g. $X_1 * X_2$) in the Multivariate Model and represented with a specific slope (e.g. β_3).

$$Y = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2$$

Just as for main effects, interactions are represented by as many slopes (contrasts) as there are Dfs.

Example (II): effect of ethnic group (say 4 categories) and social class (say 5 categories) on weight at birth; 3 slopes for ethnic group, 4 slopes for social class plus 12 slopes for in the 4*3 DF interaction (ANOVA for means);

- *NUMBER OF SLOPES* for a given predictor depends on the kind of variables

- for a *QUANTITATIVE PREDICTOR*: only one slope

$$Y' = m_Y + b(X - m_X) \quad \text{where } b \text{ is the slope}$$

example: b represents difference in mm Hg bloodpressure for 1 year of age increase

- for a *CATEGORICAL PREDICTOR AND A QUANTITATIVE DEPENDENT VARIABLE*: as many slopes as degrees of freedom (DF = number of categories-1)

$$Y' = m_Y + b_1 X_1 + b_2 X_2 + b_3 X_3 \dots \text{ where } b_1 \ b_2 \ b_3 \text{ are slopes}$$

example: $b_1 = m_1 - m_Y$; $b_2 = m_2 - m_Y$; $b_3 = m_3 - m_Y$
each category has a specific *COEFFICIENT* for each x variable

- relationship between *two CATEGORICAL VARIABLES*: as many slopes as degrees of freedom

$\text{logit}(Y') = m \text{logit}(y) + b_1 X_1 + b_2 X_2 + b_3 X_3 \dots$ where $b_1 \ b_2 \ b_3$ are slopes

and $e^{b_1} \dots$ are Odds Ratios

example: $b_1 = \text{logit}(p_1) - m \text{logit}(y)$; $b_2 = \text{logit}(p_2) - m \text{logit}(y)$

9.4 Generalized Linear Model with dependent variables with more than 1 DF

Several quantitative dependent variables. The relationship can be tested with a multivariate ANOVA model. (Multivariate ANOVA in SPSS).

Categorical dependent variable: more than 2 categories. The relationship can be tested with a logistic model. But this is not available in SPSS.

Remember we can also use a Pearson Chi-square with $DF = (L-1)*(C-1)$.

9.5 Multivariate statistical tests

The tests for GLM are subdivided into two broad categories.

When dependent variable is quantitative all methods are instances of ANOVA.

When dependent variable is categorical all methods are instances of LOGISTIC REGRESSION.

DEPENDENT VARIABLE INDEPENDENT VARIABLES	quantitative	categorical
all quantitative	ANOVA for linear regression example (I)	logistic regression example (IV)
all categorical	ANOVA with several factors example (II)	logistic regression example (V)
quantitative and categorical	Analysis of Covariance (ANCOVA) example (III)	logistic regression example (VI)

- **Probability distributions**

DEPENDENT VARIABLE	QUANTITATIVE	CATEGORICAL		COUNT	
DF = 1	Student's t (df) $t^2(df) = F(1, df)$ DF = 1 df = n-2	Normal z $z^2 = \text{Pearson } \chi^2(1)$ DF = 1	Binomial if all $E_i \geq 5$	Normal z $z^2 = \text{Pearson } \chi^2(1)$ DF = 1	Poisson if all $E_i \geq 5$
DF ≥ 1	Fisher F (DF, df) ($\neq t^2$) DF = k-1 df = n-k	Pearson χ^2 (DF) ($\neq z^2$) DF = (L-1)*(C-1)	if (80%) $E_i \geq 5$	Pearson χ^2 (DF) ($\neq z^2$) DF = (L-1)*(C-1)	if (80%) $E_i \geq 5$

INDEPENDENT- DEPENDENT VARIABLE TYPE	quantitative - quantitative	categorical - quantitative	quantitative - categorical	categorical - categorical
MODEL	linear regression	contrasts between means	logistic regression	contrasts between proportions
NULL HYPOTHESIS FOR DF=1	population R = 0	population contrast = 0	population OR = 1	population OR = 1
SAMPLING DISTRIBUTION FOR DF = 1	Student's t	Student's t	Log Likelihood χ^2	Pearson χ^2 or Log Likelihood χ^2
TESTS FOR DF = 1	$\frac{R}{\sqrt{(1-R^2)/(n-2)}}$	$\frac{m_1 - m_2}{SE}$	-2 ln (L0/L1)	$\frac{\sum (O_i - E_i)^2 / E_i}{\ln (L1/L0)}$
NULL HYPOTHESIS FOR DF ≥ 1	population multiple R = 0	all population contrasts = 0	all population OR = 1	all population OR = 1
SAMPLING DISTRIBUTION FOR DF ≥ 1	Fisher F	Fisher F	Log Likelihood χ^2	Pearson χ^2 or Log Likelihood χ^2
TESTS FOR DF ≥ 1	$\frac{R^2}{(1-R^2)/(n-k)}$	$\frac{s_m^2}{s^2}$	-2 ln (L0/L1)	$\frac{\sum (O_i - E_i)^2 / E_i}{\ln (L1/L0)}$
APPLICATION CONDITIONS	Linear relationship Unimodal distributions Equal dispersion	Unimodal distributions Equal dispersion	Logistic relationship	(80% of) $E_i \geq 5$ or Logistic relationship
NON- PARAMETRIC TESTS	Kendall rank coefficient (= τ coefficient)	Kruskal-Wallis test	Kendall rank coefficient (= τ coefficient)	Fisher exact probability test (only for DF=1)

9.6 Strategies for model building

Different strategies can be taken for building a model. The following strategy is fairly simple and should also be fairly reliable.

1) univariate selection: X variables which are significant not too far away from significance for predicting Y in univariate tests will be kept in the multivariate starting model

2) selection of non-redundant variables: only one among different redundant predictors will be kept in the multivariate starting model (e.g. if there are several variables related to mother's childbirth history, such as gestity, parity, nber. living children, only the most significant will be kept).

3) starting model without interaction terms: only main effects, without interactions, are entered in the starting model.

4) forward selection of variables in the starting model.

5) intermediate model with 2-way interactions: This model will contain all the predictors selected in the previous stage plus two-way interactions.

6) forward selection of variables and interactions in the intermediate model.

Procedure stops here if individual variables entering into a selected interaction are also selected (e.g. $X1 \cdot X3$ interaction selected and $X1$, $X3$ variables also selected). The model obtained is the final model.

7) otherwise if there are significant interactions without significant component variables the latter are also entered (e.g. as the $X1 \cdot X3$ interaction is significant the $X1$, $X3$ variables are entered). This is application of the "Hierarchy principle" which says that interaction should not be taken into account if main effect is not in the model.

Strategy for model building with logistic regression

STAGE 1

Test the effect of the variables without interaction term.

Method: stepwise ("forward conditional" in SPSS; option: "at last step").

Contrast: indicator (first).

Decision:

- put all the significant variables (that is those selected) into the model: still ADDITIVE model at this stage.
- If only one variable is significant, stop here: ADDITIVE model with a single variable.
- If more than one variable is significant go to stage 2.

STAGE 2

Test the effect of the significant variables and of their interaction(s).

Method: stepwise ("forward conditional" in SPSS; option: "at last step").

Contrast: indicator (first).

Decision:

- put all significant interactions (that is those selected) into the model.
- Also put all the variables entering into these interactions into the model (even the non significant variables provided that they have a significant interaction with another variable).
- If significant interaction(s): opt for the INTERACTIVE model.
 - Calculate the OR per stratum with the stratified chi-squares (Crosstabs in SPSS).
 - Obtain the significance levels for the variables with another contrast type, different from indicator (e.g. simple), because these significance levels are not calculated independently of the significance level of the interaction with indicator contrast. (Why not choosing another contrast at the very start then: because the relationship with individual strata OR is easier to see manually with indicator; calculations are more complex with other contrast types¹).
- If no significant interaction(s): opt for the ADDITIVE model.
 - Take a global OR for each variable (given by output of STAGE 1).

¹ Note that this is essentially for pedagogical purposes. In practice, a procedure based on simple contrasts would be quicker as it directly provides the significance levels for the variables with (or without) interaction.

-
- Significance levels are those obtained in stage 1 (they do not depend on contrast type when there are no interactions into the model).

Normal distribution. Probability area above the Z value.

Z first decimal on line header; second decimal on column header. (**NDIST formula in Excel**)

Z		0	0.01	0.02	0.03		0.04	0.05		0.06	0.07	0.08	0.09
0		0.500	0.496	0.492	0.488		0.484	0.480		0.476	0.472	0.468	0.464
0.1		0.460	0.456	0.452	0.448		0.444	0.440		0.436	0.433	0.429	0.425
0.2		0.421	0.417	0.413	0.409		0.405	0.401		0.397	0.394	0.390	0.386
0.3		0.382	0.378	0.374	0.371		0.367	0.363		0.359	0.356	0.352	0.348
0.4		0.345	0.341	0.337	0.334		0.330	0.326		0.323	0.319	0.316	0.312
0.5		0.309	0.305	0.302	0.298		0.295	0.291		0.288	0.284	0.281	0.278
0.6		0.274	0.271	0.268	0.264		0.261	0.258		0.255	0.251	0.248	0.245
0.7		0.242	0.239	0.236	0.233		0.230	0.227		0.224	0.221	0.218	0.215
0.8		0.212	0.209	0.206	0.203		0.200	0.198		0.195	0.192	0.189	0.187
0.9		0.184	0.181	0.179	0.176		0.174	0.171		0.169	0.166	0.164	0.161
1		0.159	0.156	0.154	0.152		0.149	0.147		0.145	0.142	0.140	0.138
1.1		0.136	0.133	0.131	0.129		0.127	0.125		0.123	0.121	0.119	0.117
1.2		0.115	0.113	0.111	0.109		0.107	0.106		0.104	0.102	0.100	0.099
1.3		0.097	0.095	0.093	0.092		0.090	0.089		0.087	0.085	0.084	0.082
1.4		0.081	0.079	0.078	0.076		0.075	0.074		0.072	0.071	0.069	0.068
1.5		0.067	0.066	0.064	0.063		0.062	0.061		0.059	0.058	0.057	0.056
1.6		0.055	0.054	0.053	0.052		0.051	0.049		0.048	0.047	0.046	0.046
1.7		0.045	0.044	0.043	0.042		0.041	0.040		0.039	0.038	0.038	0.037
1.8		0.036	0.035	0.034	0.034		0.033	0.032		0.031	0.031	0.030	0.029
1.9		0.029	0.028	0.027	0.027		0.026	0.026		0.025	0.024	0.024	0.023
2		0.023	0.022	0.022	0.021		0.021	0.020		0.020	0.019	0.019	0.018
2.1		0.018	0.017	0.017	0.017		0.016	0.016		0.015	0.015	0.015	0.014
2.2		0.014	0.014	0.013	0.013		0.013	0.012		0.012	0.012	0.011	0.011
2.3		0.011	0.010	0.010	0.010		0.010	0.009		0.009	0.009	0.009	0.008
2.4		0.008	0.008	0.008	0.008		0.007	0.007		0.007	0.007	0.007	0.006
2.5		0.006	0.006	0.006	0.006		0.006	0.005		0.005	0.005	0.005	0.005
2.6		0.005	0.005	0.004	0.004		0.004	0.004		0.004	0.004	0.004	0.004
2.7		0.003	0.003	0.003	0.003		0.003	0.003		0.003	0.003	0.003	0.003
2.8		0.003	0.002	0.002	0.002		0.002	0.002		0.002	0.002	0.002	0.002
2.9		0.002	0.002	0.002	0.002		0.002	0.002		0.002	0.001	0.001	0.001
3		0.001	0.001	0.001	0.001		0.001	0.001		0.001	0.001	0.001	0.001

Z		0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0		1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0.1		0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0.2		0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0.3		0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0.4		0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0.5		0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0.6		0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0.7		0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0.8		0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0.9		0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1		0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1.1		0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1.2		0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1.3		0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1.4		0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1.5		0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1.6		0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1.7		0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1.8		0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1.9		0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2		0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2.1		0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2.2		0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2.3		0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.017	0.017
2.4		0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2.5		0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2.6		0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2.7		0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2.8		0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2.9		0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3		0.003	0.003	0.003	0.002	0.002	0.002	0.002	0.002	0.002	0.002
3.1		0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.001
3.2		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.3		0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
3.4		0.001	0.001	0.001	0.001	0.					

Student's t distribution. Double-tailed probability area: to be above the t value or below the $-t$ value
 t first decimal on line header; DF on column header. (**TINV formula in Excel**)

t values	df	5	10	15	20	25	30	40	50	60	100	5000
0.2		1.48	1.37	1.34	1.33	1.32	1.31	1.30	1.30	1.30	1.29	1.28
0.1		2.02	1.81	1.75	1.72	1.71	1.70	1.68	1.68	1.67	1.66	1.65
0.05		2.57	2.23	2.13	2.09	2.06	2.04	2.02	2.01	2.00	1.98	1.96
0.01		4.03	3.17	2.95	2.85	2.79	2.75	2.70	2.68	2.66	2.63	2.58
0.001		6.87	4.59	4.07	3.85	3.73	3.65	3.55	3.50	3.46	3.39	3.29

Chi-square distribution. Probability area: to be above the X value.

X first decimal on line header; DF on column header. (**CHIINV formula in Excel**)

X² values								
p	df	1	10	20	40	60	80	120
0.2		1.64	13.44	25.04	47.27	68.97	90.41	132.81
0.1		2.71	15.99	28.41	51.81	74.40	96.58	140.23
0.05		3.84	18.31	31.41	55.76	79.08	101.88	146.57
0.01		6.63	23.21	37.57	63.69	88.38	112.33	158.95
0.001		10.83	29.59	45.31	73.40	99.61	124.84	173.62

F distribution. Probability to be above the F value.

P on line header; DF on column header. (**FINV formula in Excel**)

F values	df1	1	1	1	1	1	1	1	1
p	df2	1	10	20	40	60	80	120	5000
0.2		9.47	1.88	1.76	1.70	1.68	1.67	1.66	1.64
0.1		39.86	3.29	2.97	2.84	2.79	2.77	2.75	2.71
0.05		161.45	4.96	4.35	4.08	4.00	3.96	3.92	3.84
0.01		4052.18	10.04	8.10	7.31	7.08	6.96	6.85	6.64
0.001		405311.58	21.04	14.82	12.61	11.97	11.67	11.38	10.84

F values	df1	10	10	10	10	10	df1	20	20	20	20
p	df2	20	40	60	80	120	df2	40	60	80	120
0.2		1.53	1.44	1.41	1.39	1.37	0.2	1.36	1.32	1.31	1.29
0.1		1.94	1.76	1.71	1.68	1.65	0.1	1.61	1.54	1.51	1.48
0.05		2.35	2.08	1.99	1.95	1.91	0.05	1.84	1.75	1.70	1.66
0.01		3.37	2.80	2.63	2.55	2.47	0.01	2.37	2.20	2.12	2.03
0.001		5.08	3.87	3.54	3.39	3.24	0.001	3.15	2.83	2.68	2.53

F values	df1	40	40	40	df1	60	60	df1	80
p	df2	60	80	120	df2	80	120	df2	120
0.2		1.27	1.25	1.23	0.2	1.22	1.20	0.2	1.18
0.1		1.44	1.40	1.37	0.1	1.36	1.32	0.1	1.29
0.05		1.59	1.54	1.50	0.05	1.48	1.43	0.05	1.39
0.01		1.94	1.85	1.76	0.01	1.75	1.66	0.01	1.60
0.001		2.41	2.26	2.11	0.001	2.10	1.95	0.001	1.86

Chapter 10

Multifactorial ANOVA

The incorporation of several factors in the ANOVA allows to solve the following problems:

The *confounder problem* : as we have already seen the presence of a statistical relationship between any two variables must be interpreted with care. When the ANOVA only includes a single factor, the effect of the factor can due to the confounding effect of a “hidden” factor, which is not included in the statistical analysis. Confounding can only occur if the two factors are related, i.e. **if relative frequencies of factor A categories depend on factor B categories** (non-orthogonal design). Confounding is *not* possible when relative frequencies of factor A are constant across factor B categories (orthogonal design).

Orthogonal design: confounding is *not* possible

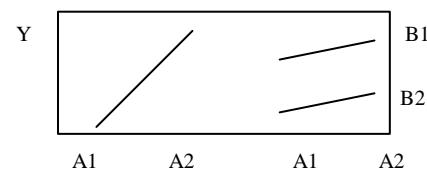
	A1	A2	total
B1	p_1	p_2	1
B2	p_1	p_2	1

Non-orthogonal design: confounding is possible

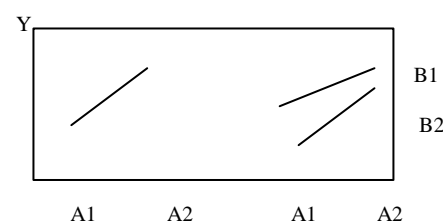
	A1	A2	total
B1	p_{11}	p_{12}	1
B2	p_{21}	p_{22}	1

$$p_{11} \neq p_{21} \text{ and } p_{12} \neq p_{22}$$

Example of confounder: effect of A in single factor ANOVA is larger than in two-factor ANOVA because larger proportion of B1 category in A2 vs A1.



Example of interaction: effect of B is smaller for A2 category.



The *interaction problem*: there is an interaction when **effect of a given factor on the dependent variable (not its frequency) is related to the other factor**. Interaction means that the effect of factor A depends on the value of factor B. Interaction is reflected in *non-parallelism*: contrast values (slopes) of factor A change as a function of factor B.

• tests for two-factor ANOVA

Effect of each factor is tested separately by F-test, with DF calculated in the same way as in one-factor ANOVA.

Effect of interaction is also tested separately with DF equal to the product of factor's DFs.

Tests for contrasts with Student's t-tests. For each factor and interaction there are as many contrasts as DF.

Denominator of F-tests is the within-cells MS

several H_0 :

for testing factor A: all contrasts

for factor A = 0

for testing factor B: all contrasts

for factor B = 0

for testing the interaction: all

contrasts for interaction AB = 0

Source of variation	df	SS	MS
F			
factor A	k_1-1
factor B	k_2-1
interaction AB	$(k_1-1)*(k_2-1)$
residue	$n - k_1 - k_2 - (k_1-1)*(k_2-1) + 1$
total	$n-1$
Test : $F_{\dots\dots\dots} = MS_{\dots}/MS_{\text{residue}}$			

• Anova Design

Number of factors: generalization of the preceding method with more than two factors. The only new element is the occurrence of the several levels of interaction. Besides the interactions between two factors (interactions of the first order), there are interactions of upper order (second, third...) between 3, 4 ... factors. A second order interaction that the amount of interaction between two factors depends on a third one.

Within vs residual error term: *Within* error term means that denominator of the F-ratio is the within cell MS. *Residual* error term means that only residual error is taken as denominator which is useful for repeated measurements (as explained after).

Within+residual error term means that non-significant factors are dropped out which makes that error term is increased by non-significant variance ("residual" variance). This is the usual option.

Unique Vs sequential testing design: *Unique* design means testing all factors and interactions simultaneously. Each factor (and interaction) is then corrected for confounding effects of all other factors. This is the usual option. *Sequential* design means that factors (and interactions) are included one after another and are corrected for those entered before into the model and confounded with those entered after into the model.

A NOTE ON TERMINOLOGY:

- Univariate: a single predictor ($X \rightarrow Y$)
- Multivariate: several predictors ($X_1, X_2, \dots \rightarrow Y$)
- Multivariate in SPSS slang: several dependent variables

($X \rightarrow Y_1, Y_2, \dots$)

($X_1, X_2, \dots \rightarrow Y_1, Y_2, \dots$)

- **● an example of two-factor ANOVA**

Example: relationship between district (4 categories) , salt consumption (in 2 categories) and systolic pressure (in mm Hg).

Descriptive and bivariate statistics (SPSS output 10.1)

There is a significant relationship between SP and district ($F(3,36) = 4.25$, $p = .01$) but also between SP and salt consumption ($F(1,38) = 204.19$, $p < .0001$), and seemingly between district and salt consumption ($\chi^2(3) = 12.14$, $p = .007$, but this test is only indicative as there are more than 50% cells with frequencies below 5).

ANOVA with two factors (SPSS output 10.2)

District effect is NS when salt is included into the model ($F(3,32) = 0.03$, $p = .99$). Interaction is also NS ($F(3,32) = 1.27$, $p = .3$) although contrast lines are not strictly parallel. Only salt effect remains significant ($F(1,32) = 130.46$, $p < .001$).

Conclusion

Salt is a confounder for SP-district relationship.

● Analysis of Covariance (ANCOVA)

ANCOVA allows us to test effects of both categorical and quantitative variables on a quantitative variable. ANCOVA designs are mixtures of ANOVA and regression designs.

Incorporation of quantitative predictors (“covariates”) allow to take account of their possible confounding effects.

● Application conditions:

Homogeneity of within-group variances (as in ANOVA) but also homogeneity of covariances. This means that both variances of the dependent variable and its covariance with the quantitative predictor should be constant over the different levels of the categorical predictor.

● tests for ANCOVA

Homogeneity of variance is tested by Bartlett-Box. Homogeneity of covariance is tested by introducing a factor-covariate interaction into the model.

Effect of each factor and each covariate is tested separately by F-test, with DF calculated in the same way as in ANOVA and in regression.

Effect of interaction is also tested separately with DF equal to the product of factor's DFs.

several H_0 :

for testing factor A: all contrasts

for factor A $= 0$

for testing covariate X: $\rho = 0$

for testing the interaction: ρ is constant over A levels.

Source of variation	df	SS	MS
F			
factor A	$k_1 - 1$
covariate X	1
interaction AX	$(k_1 - 1)$
residue	$n - 2 * (k_1 - 1) - 2$
total	$n - 1$
Test : $F_{\dots\dots\dots} = MS_{\dots} / MS_{\text{residue}}$			

● example of Analysis of Covariance

Effect of mother education on gestity, controlling for age (covariate). See SPSS output 10.3. First ANOVA is runned with the age*education interaction in the design. As interaction is NS ($p=.226$), analyze of covariance is applicable. A second ANOVA, without interaction with covariate, shows that effect of education is just NS ($p=.057$). Effect of age is S ($p<.0005$).

● Repeated measures ANOVA

To be used when the same quantitative variable is measured on several occasions on the same subjects.

Simplest situation: when the same variable is measured twice on the same subjects. A “**paired t-test**” can then be used. It is based on the mean and standard-deviation of the differences.

Alternative test gives exactly the same result: a repeated measures ANOVA ($F(DF=1, n-1) = t^2(DF=n-1)$).

With more than 2 measurements, two different ANOVA tests are available: univariate or multivariate.

Univariate ANOVA: two different factors are considered. Subjects are taken as a “random” factor which means that each subject is a different level of a factor with $n-1$ DF. This is a random factor because the same factor (“subjects”) will usually contain different levels (individuals) in two different studies. The second factor is a “fixed” factor, equivalent to all the factors we have considered up to now. This factor correspond to the different repetitions of the measure. each repetition is made in a different condition (day, bodily location...) and these conditions are the factor's levels. The denominator of the F ratio used for testing the effect of the fixed factor is the variance of the fixed factor effect across subjects (MS fixed-random interaction). This makes sense: the test will be less significant when fixed factor effect varies more across subjects.

Condition for using univariate ANOVA design: Mauchly Sphericity Test must be NS (H_0 : equality of

● example of repeated measures ANOVA

variances and covariances of individual pairwise differences between levels).

example of paired t-test: Is there a difference between skull perimeter of twins ? See SPSS output 10.4

Paired t-test

$$t \text{ (DF= } n_d - 1) = \frac{m_d}{s_d / \sqrt{n_d}}$$

where n is the number of pairs

m_d is the mean of the differences

s_d is the SD of the differences

Univariate ANOVA (“mixed design”)

$$F \text{ (DF=k-1, (k-1)*(n-1))} = \frac{\text{MS fixed}}{\text{MS interaction fixed*random}}$$

where k is the number of repetitions

n is the number of subjects

Mauchly Sphericity Test : has to be NS for using **Univariate ANOVA**. Sphericity means that variability of the fixed factor effect is constant for the different possible pairwise differences between factor's levels.

Example: variability of CBF differences between cerebral areas A & B should be the same as between A & C, B & C

See SPSS outputs 10.5 & 10.6

Cerebral blood flow (CBF) was measured in 5 different cervical regions on the same subjects (only subjects with cognitive neglect will be analysed here). Question: does CBF depend on the region ?

DATA¹

Cerebral Region	Superior Parietal	Inferior Parietal	Posterior Temporal	Frontal I	Frontal II	Mean 5 regions
Diagnostic	CBF 1	CBF 2	CBF 3	CBF 4	CBF 5	
neglect (n=15) code = 1	90.43	88.48	87.38	95.38	93.85	91.11
no neglect (n=13) code = 0	97.29	97.68	99.20	98.66	99.38	98.44
mean both groups	93.62	92.75	92.87	96.90	96.42	94.51

Table 10.1

Test of the effect of brain location on CBF for subjects with neglect (SPSS output 10.5). As Mauchly Sphericity test is NS ($p=.13$), we can use the univariate test for repeated measurements. Effect of brain location is highly significant with this procedure ($F(df=4,56) = 8.31$; $p<.0005$). Notice that the multivariate test although less powerful is also significant but with a higher type I error ($p=.025$). Simple contrasts are not available for repeated measurements. Difference contrasts were used instead. They show that differences between Parietal regions are not significant ($p=.16$), that difference between Temporal and the two Parietal regions is also NS ($p=.19$). But difference between non-Frontal regions and Frontal I region is highly significant ($p<.0005$). (Less interesting: Difference between Frontal II region and all 4 others is just significant ($p=.033$)).

We conclude from these tests and from the mean values presented in Table 10.1 that CBF of subjects with neglect is significantly lower for non-Frontal brain regions.

● **example of repeated measures ANOVA in which the within-subjects factor is (“random” factor) crossed with a between-subjects factor (“fixed” factor).**

Does CBF depend on brain location and on presence vs absence of neglect ? Does the effect of brain location depend on neglect ? To answer these questions, both for subjects with neglect and those without neglect are now included in the repeated measures ANOVA (SPSS output 10.6).

¹ Demeurisse, G., Hublet, Cl., Paternot, J., Colson, C. and Serniclaes, W. (1997) “Pathogenesis of subcortical visuo-spatial neglect. A HMPAO SPECT study” *Neuropsychologia*. 35, 731-735.

Effect of the between-subjects factor, neglect, is highly significant ($F(df=1,26)=31.68, p<.0005$). Effect of neglect is negative (-7.4) because CBF is lower for neglect (coded 1) than for no-neglect (coded 0). As Mauchly Sphericity test is NS ($p=.199$), we can use the univariate test for assessing the effects of the within-subject factor, brain location. Effect of brain location is significant ($F(df=4,104) = 4.26; p=.003$) and so does the location-neglect interaction ($F(df=4,104) = 3.23; p=.015$). Difference contrasts were used for location and simple (first) contrasts for neglect. For location, results are similar to those obtained in the previous analysis. Only two difference contrasts are significant: (1) between non-Frontal regions, on the one hand, and Frontal I region, on the other hand ($p=.011$); (2) between the Frontal II region and all 4 others ($p=.023$). Neglect-location interaction contrasts are also available in SPSS Output 10.5. A significant interaction contrast means that effect of neglect (simple contrast) is different in a given location versus the mean of the previous ones (difference contrast) and that the effect is larger in this location if the contrast coefficient is positive, lesser if the coefficient is negative. Only interaction contrast T4 is significant here ($p=.031$) and it has a positive value (5.224) which means that effect of neglect is lesser in Frontal I region vs. mean of the 3 non-frontal regions. This interpretation can be checked by looking at the data in Table 10.1. Mean of neglect effect in Frontal I is about -3 units CBF against -8 units in the three non-frontal regions, a difference of about +5.

We conclude from these tests and from the mean values presented in the Table above that CBF is significantly lower for subjects with neglect and for non-frontal brain regions. Further, CBF lowering for subjects with neglect is more important in non-frontal regions as revealed by the significant location-neglect interaction.

Chapter 11

Multiple linear regression

We use multiple linear regression when the value of a quantitative variable is predicted from the values of several other quantitative variables or “**predictors**”, using a linear equation.

A **partial regression coefficient** (b) is attached to each predictor.

Relative importance of the predictors is given by the **standardized** partial regression coefficients (beta), or by **partial correlation coefficients** ($r_{yx1 \cdot x2}$)

$r^2_{yx1 \cdot x2}$ is the proportion of variance explained by predictor x_1

If predictor and dependent variable are not correlated with the other predictor $r^2_{yx1 \cdot x2} = r^2_{yx1}$

Overall strength of the prediction is given by a **multiple correlation coefficient** (**R**).

R^2 is the proportion of variance explained by all predictors.

An unbiased estimation of population ρ^2 is given by “adjusted R^2 ”.

$$y' = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

$b_1 \ b_2 \ \dots \ b_k$ are **partial regression coefficients**

standardized (partial) regression coefficients:

$$\text{beta}_i = b_i \cdot s_{xi}/s_y$$

squared partial correlation coefficients (with two predictors) :

$$r^2_{yx1 \cdot x2} = \frac{(r_{yx1} - r_{yx2} \cdot r_{x1 \cdot x2})^2}{(1 - r^2_{yx2})(1 - r^2_{x1 \cdot x2})}$$

squared multiple correlation coefficient (with two predictors)

$$R^2 = \frac{r^2_{yx1} + r^2_{yx2} - 2r_{yx1} \cdot r_{yx2} \cdot r_{x1 \cdot x2}}{1 - r^2_{x1 \cdot x2}}$$

adjusted $R^2 = 1 - (\text{residual MS} / \text{total MS})$

• tests for multiple linear regression

Test of H_0 : multiple correlation coefficient is null, by F-test with number of degrees of freedom of the regression equal to the number of predictors.

Test of H_0 : a partial correlation coefficient is null, by Student's t-test or just the same with a F test..

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

or just the same : $\rho = 0$

Source of variation	df	SS	MS	F
regression	k
residue	n-k-1	
total	n-1	

Test : $F^{k,n-k-1} = MS_{\text{regression}}/MS_{\text{residue}}$

$$H_0 : \beta_1 = 0$$

or just the same : $\rho_1 = 0$

$$F^{1,n-k-1} = \frac{r^2_{yx1.x2\dots xk}}{(1 - r^2_{yx1.x2\dots xk})/(n-k-1)}$$

$$F^{1,n-k-1} = (t^{n-k-1})^2$$

• an example of multiple linear regression

Example: prediction of systolic pressure from weight and cholesterol level in a sample of 300 (or slightly less depending on the missing values) male subjects.

Univariate correlations (SPSS output 11.1):

$R(\text{SP}, \text{Weight}) = 0.150$ (S, $p = .01$)

$R(\text{SP}, \text{chol}) = 0.147$ (S, $p = .01$)

$R(\text{chol}, \text{Weight}) = 0.174$ (S, $p = .003$)

Descriptive statistics (SPSS output 11.2):

$SD(\text{SP}) = 16.20$

$SD(\text{WEIGHT}) = 10.39$

$SD(\text{CHOL}) = 39.77$

Multivariate equation: (SPSS output 11.3):

predicted pressure = $116 + 0.18 * \text{weight} + 0.05 * \text{chol}$

Standardized regression coefficients :

$\text{beta}_{\text{pressure weight}} = 0.18 * 10.39 / 16.20 \cong 0.11$

$\text{beta}_{\text{pressure chol}} = 0.05 * 39.77 / 16.20 \cong 0.12$

Both predictors have about the same strength.

Multiple correlation coefficient:

$$R^2 = \frac{(.150)^2 + (.147)^2 - 2(.150)(.147)(.174)}{1 - (.174)^2} \cong .035$$

Percentage of explained variance is about 3 %.

Tests:

Prediction of systolic pressure from weight and number of cigs. is significant ($F(2,292) = 3.98$; S at $p = .02$).

Effect of weight alone is significant (S at $p=.02$).

Effect of number of cigs. is non significant (NS, $p= 0.31$).

Conclusions: systolic pressure is related to weight; effect of number of cigarettes a day is NS when weight is included in the regression analysis which shows that this is not a confounder.

As weight alone is significant ρ estimation should be given by separate correlation between systolic pressure and weight :

$$\rho_{\text{est}} = 0.15 \pm 1.96 * \sqrt{(1-.15)^2 / (297-2)} = (0.04; 0.26)$$

• **predictor selection for multiple linear regression**

Forward inclusion in the regression equation: predictors are included by order of decreasing partial correlation with the dependent variable. The predictor is included if the partial correlation is significant and if the multiple correlation with the other predictors already in the equation is not too large (if the R^2 between candidate predictor and other predictors is not too large or just the same if $1-R^2$, which is called "**tolerance**", is not too small).

Backward inclusion: all the predictors are included at first and are thereafter selectively excluded. A predictor is excluded if the partial correlation is not significant

Stepwise inclusion: starts as in forward but each time a new variable is entered into the equation the variables already in the equation are checked as in backward.

See SPSS output 11.4 as example.

Chapter 12

Multilogistic regression

Multilogistic regression allows to test the effects of several predictors, categorical or quantitative, on a categorical dependent variable.

Purpose: to cope with possible confounding effects and interactions in OR estimation.

• Application condition

Relationship between predictors and dependent variable should fit a logistic regression curve reasonably well, i.e. the differences between observed and predicted values should be non significant.

Fit is always perfect when all predictors are categorical and with all terms (factors and interactions) into the model ("saturated" model).

• tests of factors and interactions

Effect of each factor is tested separately by -2LL Improvement Chi-square (or with Wald test), with DF calculated in the same way as in one-factor logistic regression.

Effect of interaction is also tested separately with DF equal to the product of factor's DFs.

• predictor selection for multilogistic regression

Forwards stepwise and backward stepwise selection of both factors and interactions are possible. Different criteria can be used for each option. LR option is the most rigorous criteria but takes the longest calculation time.

A model with 3 terms: 2 factors (1 DF each) and their interaction.

$$\text{Logit}(p) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2$$

With 2 DF for factor 1:

$$\begin{aligned} \text{Logit}(p) = & \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_2 \\ & + \beta_4 x_{11} * x_2 + \beta_5 x_{12} * x_2 \end{aligned}$$

Test that the data fit the logistic model:

Hosmer-Lemeshow χ^2

Practically: exclude large outliers.

Wald test is always available.

-2LL χ^2 improvement test is only available with stepwise selection.

Improvement $\chi^2 =$

$$\{-2LL \chi^2 (\text{model})\} - \{-2LL \chi^2 (\text{model at previous step})\}$$

$$\text{Model } \chi^2 = \{-2LL \chi^2 (\text{model})\} - \{-2LL \chi^2 (\text{model with constant only})\}$$

- **Mantel-Haenszel procedure**

“Mantel-Haenszel” procedure
gives results similar to those of the
logistic model without interactions (see
EPI-INFO Output 12.1).
Using the “M-H” method requires that
interaction is NS which can be tested
by the “Wulf” test.

• **Example of multilogistic regression with 2 categorical predictors**
(SPSS output 12.1)

Effect of smoking and ethnic group on lowweight at birth in a sample of 189 births. Smoking is in two categories (non-smoker, smoker) and ethnic group in 3 categories (white, black, other).

Data description: OR-smoker is largest for whites (5.76), lower for blacks (3.30) and lowest for others (1.25).

Global OR-smoker without stratification is 2.02 (SPSS output 12.2) . But this value does not take account of size differences between ethnic group samples (96 whites, 26 blacks, 67 others) which affect the global OR calculation because lowweight prevalence is smaller for whites although they display the highest smoking rate. We therefore expect a larger OR-smoker estimation when controlling for ethnic group.

When ethnic group is incorporated into the model (SPSS output 12.3), OR -smoker is estimated as 3.05. As expected this is above the 2.02 value obtained without controlling for ethnic group. Notice that ethnic group effect is S ($p=.01$) and should therefore be taken into account for OR-smoker estimation. Also note that OR-ethnic group (other vs. black) is 1.025 and OR (white vs. black) is 0.338. Risk of lowweight is thus almost the same for black and other groups and smaller for the white group with this model where effect of ethnicity is assessed without taking account of the interaction between smoking and ethnicity. These OR cannot exactly conform to the data because there is always some interaction between variables, whatever they are. In the present data, prevalence of lowweight at birth amounts to 42%, 37%, and 24%, respectively for blacks, others and whites. The OR (white vs. black) of 0.338 overestimates the 18% decrease for white vs. black (the empirical OR is 0.43), whereas the OR-ethnic group (other vs. black) of 1.025 slightly distorts the 5% decrease for others vs. blacks (underestimates the empirical OR is 0.81).

As shown in Figure 12.1, effect of smoking is not constant across ethnic categories. Should the smoker-ethnic group interaction also be taken into the model ? No because

it is non significant ($P=.22$; see SPSS output 12.4). Therefore, confidence limits are based on the model without interaction.

95% CI for $\ln(OR)$ smoke are: $1.1159 \pm 1.96 \cdot 0.3692 = (0.39 ; 1.84)$

95% CI for OR smoke are: (1.48 ; 6.29)

Had the interaction been significant, we should have a different OR estimation for each ethnic category (see above OR-smoker = 5.76 for whites, 3.3 for blacks and 1.25 for others). These OR can also be retrieved from the Logistic regression output. Start from the fact that “non-smoke” and “black” are taken as reference categories here (because we used indicator first contrasts) and that other ethnic groups are in the following order: other (2nd) and white (third). Therefore the OR-smoke (3.30) corresponds to the risk for the blacks. OR-smoke for others is obtained by $\text{EXP}(1.194 - 0.971) = 1.25$, where 1.194 is the slope the smoke and -0.971 is the slope for the first contrast for interaction . Similarly, OR-smoke for whites is obtained by $\text{Exp}(1.194 + .557) = 5.76$.

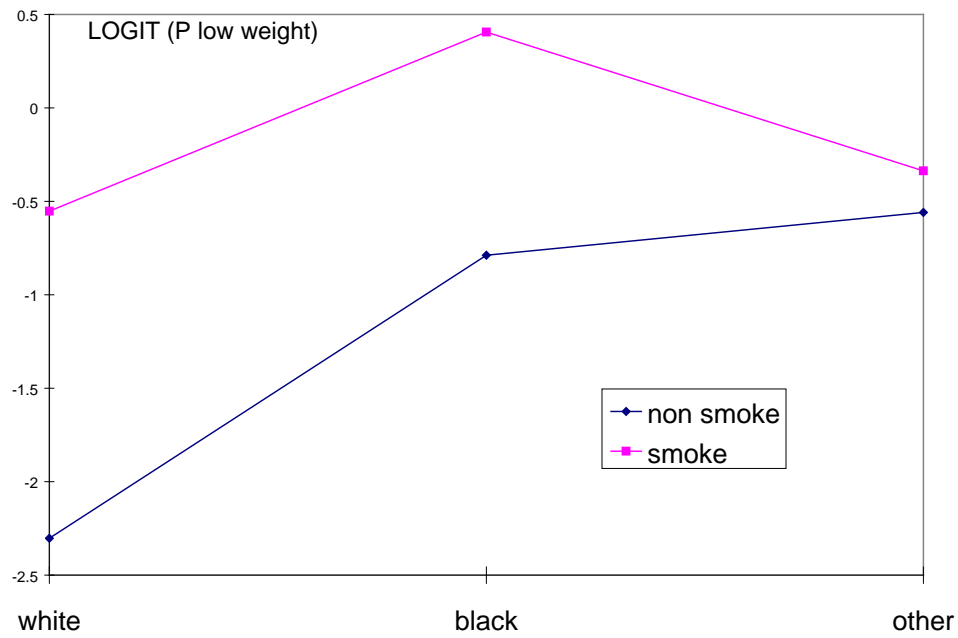
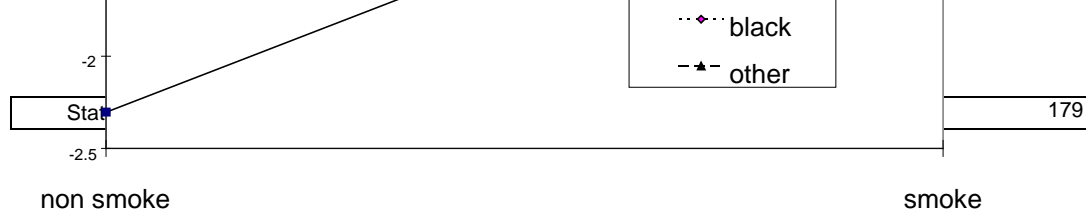


Figure 12.1 - Non-parallelism of logit lines reveals the presence of ethnic group-tabagism interaction. The interaction is however non-significant in this example ($P = .22$; see text).

• **Repeated measures for a categorical dependent variable**

• **McNemar Chi-square**

When a categorical variable is measured twice on the same subjects ("two related samples") data can be represented as follows. We construct a contingency table with 2 times 2 entries, in which the rows correspond to the number of events (1) or non-events (0) in sample 1, and the columns to the number of events and non-events in sample 2. Examples of events are: diseased, vaccinated, ... Examples of factors which make the distinction between samples are: drug vs. placebo, district management policy, ... In what follows we will take the diseased and drug example. The contents of the 4 cells within the table correspond to the number of individuals who are:

- diseased with placebo and not with drug (s frequency)
- diseased with drug and not with placebo (r frequency)
- diseased with both treatments (not relevant)
- non-diseased with both treatments (not relevant).

Only part of the data in the table are useful for the test. Frequencies of diseased or non-diseased with both treatments are *not* used because they do not provide any information on the difference between treatment effects. Relevant information is only provided by frequencies of subjects which exhibit a change in one or another direction, i.e. by **r** and **s** frequencies. Difference between directions of change is tested by "**MCNEMAR** test".

		drug	
		1	0
placebo	1	not relevant	S
	0	r	not relevant

MCNEMAR χ^2 (DF=1)

test of H0:

population r = population s.

Provided that (r+s)/2 is larger than or equal to 5, the following ratio:

$$\frac{(r-s)^2}{r+s}$$

follows an approximate χ^2 (DF=1) distribution.

• **Non-parametric alternative to**

McNemar: Binomial test

Used when $(r+s)/2$ is lesser than 5.

• **Cochran's Q test**

This test is a generalized McNemar test. Cochran's Q test is used when the same categorical variable is measured twice or more on the same subjects ("two or several related samples").

• **Logistic models with within-subjects factors**

A logistic model with treatment (drug vs. placebo) as within-subject factor would give similar results to those obtained with McNemar test.

Logistic models can also be applied to several repetitions (e.g. drug1, drug2, placebo), and is then similar to Cochran's Q test.

Logistic models can also be used for designs which cannot be treated with Cochran's Q test:

for repeated measures with quantitative predictors

example of within-subject factor with quantitative predictors:
effect of varying the amount of quinine absorption for each subject in a sample of malaria patients (quantitative within subject) on malaria symptoms (present vs. absent).

for a mixture of within-subjects and between subjects predictors

example of mixture of within-subjects and between subjects predictors:
prevalence of melanoma before / after treatment (treatment is within subjects) in different countries (between subjects).

Logistic models with within-subject factors are not available in SPSS.

• **Example of Mc Nemar test (SPSS Output 12.5).**

example: from "Basic medical statistics" A.Bahn, Grune & Stratton eds., N.Y. & London, 1972, p.240/

Results from skin testing 282 patients with 2 types of penicillin placed on the 2 arms of each patient by random allocation.

Critical information is provided by the cases for which one of the 2 factors is active and the other not.

Difference between frequencies of toxic reactions for the two penicillin fails to reach significance (McNemar $\chi^2 = 0.12$; $p > .50$). (SPSS: use the Cochran's Q command)

Strictly speaking χ^2 (z^2) applies to continuous variables. When we use χ^2 for testing differences between counts then we do as if the count was obtained by rounding a continuous value. The r-s difference might in fact be lesser than it really is. If r comes from rounding r-0.5 and if s comes from s + 0.5 then r-s is in fact 1 unit smaller or larger.

Continuity correction: we subtract 1 to $|r-s|$. This is a "conservative" procedure because $|r-s|$ might in fact be 1 unit smaller than it really is. (With SPSS you get continuity correction by using the McNemar command)

		penicillin G	
		react	no
peni- cillin BT	react	10	16
	no react	18	238

Compare 18 and 16.

As $(18+16)/2 = 17 > 5$, we can use the McNemar test

McNemar =
 $(18-16)^2 / (18+16) = 0.1176$
 χ^2 NS ($p = .73$)

Continuity correction

McNemar- corrected =
 $(|18-16|-1)^2 / (18+16) =$
 0.0294
 χ^2 NS ($p = .86$)

• **example of Binomial test (see SPSS output 12.6)**

31 medical students are given caffeine (one day) and a placebo (another day). 22 sleep well both with caffeine and placebo. 8 sleep well with placebo but not with caffeine. 1 sleeps well with caffeine but not with placebo. Does caffeine have an effect on sleep quality ?

		Caffeine	
		sleep well	sleep bad
Placebo	sleep well	22	8
	sleep bad	1	0

Conclusion : caffeine affects sleep quality (Binomial test; S at $p=.0391$).

As $(8+1)/2 = 4.5 < 5$, we *cannot* use the McNemar test

Binomial test: $H_0: n*\pi = 4.5$
 $P((8 \text{ over } 9 / n*\pi = 4.5) \text{ or } (1 \text{ over } 9 / n*\pi = 4.5)) = .0391$

EPI-INFO output 12.1

Command: statcalc , 2*2*n tables, F6, give file name (mh.txt),
 type data 1st stratum, enter, F5, F2,
 type data 2nd stratum, enter, F5, F2,
 type data 3rd stratum, enter, F5
 enter, F5, F6
 F10 ...

+ Disease - +-----+-----+ + 6 4 10 +-----+-----+ - 5 11 16 +-----+-----+ E 11 15 26 x p o s u r e	Analysis of Single Table Odds ratio = 3.30 (0.49 <OR< 24.54*) Cornfield 95% confidence limits for OR *Cornfield not accurate. Exact limits preferred. Relative risk = 1.92 (0.79 <RR< 4.66) Taylor Series 95% confidence limits for RR Ignore relative risk if case control study. <table border="0"> <tr> <td style="text-align: center;">Chi-Squares</td> <td style="text-align: center;">P-values</td> </tr> <tr> <td style="text-align: center;">-----</td> <td style="text-align: center;">-----</td> </tr> <tr> <td>Uncorrected :</td> <td>2.08 0.1488556</td> </tr> <tr> <td>Mantel-Haenszel:</td> <td>2.00 0.1569067</td> </tr> <tr> <td>Yates corrected:</td> <td>1.07 0.3003814</td> </tr> <tr> <td>Fisher exact: 1-tailed P-value:</td> <td>0.1504106</td> </tr> <tr> <td></td> <td>2-tailed P-value: 0.2279715</td> </tr> </table>	Chi-Squares	P-values	-----	-----	Uncorrected :	2.08 0.1488556	Mantel-Haenszel:	2.00 0.1569067	Yates corrected:	1.07 0.3003814	Fisher exact: 1-tailed P-value:	0.1504106		2-tailed P-value: 0.2279715
Chi-Squares	P-values														
-----	-----														
Uncorrected :	2.08 0.1488556														
Mantel-Haenszel:	2.00 0.1569067														
Yates corrected:	1.07 0.3003814														
Fisher exact: 1-tailed P-value:	0.1504106														
	2-tailed P-value: 0.2279715														

An expected cell value is less than 5.
 Fisher exact results recommended.

F2 More Strata; <Enter> No More Strata; F10 Quit

+ Disease - +-----+-----+ + 5 7 12 +-----+-----+ - 20 35 55 +-----+-----+ E 25 42 67 x p o s u r e	Odds ratio = 1.25 (0.29 <OR< 5.22*) Cornfield 95% confidence limits for OR *Cornfield not accurate. Exact limits preferred. Relative risk = 1.15 (0.54 <RR< 2.44) Taylor Series 95% confidence limits for RR Ignore relative risk if case control study. <table border="0"> <tr> <td style="text-align: center;">Chi-Squares</td> <td style="text-align: center;">P-values</td> </tr> <tr> <td style="text-align: center;">-----</td> <td style="text-align: center;">-----</td> </tr> <tr> <td>Uncorrected :</td> <td>0.12 0.7307388</td> </tr> <tr> <td>Mantel-Haenszel:</td> <td>0.12 0.7326783</td> </tr> <tr> <td>Yates corrected:</td> <td>0.00 0.9882324</td> </tr> <tr> <td>Fisher exact: 1-tailed P-value:</td> <td>0.4867171</td> </tr> <tr> <td></td> <td>2-tailed P-value: 0.7510270</td> </tr> </table>	Chi-Squares	P-values	-----	-----	Uncorrected :	0.12 0.7307388	Mantel-Haenszel:	0.12 0.7326783	Yates corrected:	0.00 0.9882324	Fisher exact: 1-tailed P-value:	0.4867171		2-tailed P-value: 0.7510270
Chi-Squares	P-values														
-----	-----														
Uncorrected :	0.12 0.7307388														
Mantel-Haenszel:	0.12 0.7326783														
Yates corrected:	0.00 0.9882324														
Fisher exact: 1-tailed P-value:	0.4867171														
	2-tailed P-value: 0.7510270														

An expected cell value is less than 5.

Fisher exact results recommended.

F2 More Strata; <Enter> No More Strata; F10 Quit

<p>+ Disease -</p> <p>+-----+-----+</p> <p>+ 19 33 52</p> <p>+-----+-----+</p> <p>- 4 40 44</p> <p>+-----+-----+</p> <p>E 23 73 96</p>	<p>Odds ratio = 5.76 (1.62 <OR< 22.36*)</p> <p>Cornfield 95% confidence limits for OR</p> <p>*Cornfield not accurate. Exact limits preferred.</p> <p>Relative risk = 4.02 (1.48 <RR< 10.93)</p> <p>Taylor Series 95% confidence limits for RR</p> <p>Ignore relative risk if case control study.</p>
--	--

<p>x</p> <p>p</p> <p>o</p> <p>s</p> <p>u</p> <p>r</p> <p>e</p>	<p>Chi-Squares P-values</p> <p>-----</p> <p>Uncorrected : 9.86 0.0016931 □---</p> <p>Mantel-Haenszel: 9.75 0.0017903 □---</p> <p>Yates corrected: 8.41 0.0037386 □---</p> <p>F2 More Strata; <Enter> No More Strata; F10 Quit</p>
--	---

<p>+ Disease -</p> <p>+-----+-----+</p> <p>+ 19 33 52</p> <p>+-----+-----+</p> <p>- 4 40 44</p> <p>+-----+-----+</p> <p>E 23 73 96</p> <p>x</p> <p>p</p> <p>o</p> <p>s</p> <p>u</p> <p>r</p> <p>e</p>	<p>***** Stratified Analysis *****</p> <p>Summary of 3 Tables</p> <p>Crude odds ratio for all strata = 2.02</p> <p>Mantel-Haenszel Weighted Odds Ratio = 3.09</p> <p>Cornfield 95% Confidence Limits</p> <p>1.40 < 3.09 < 6.73</p> <p>Mantel-Haenszel Summary Chi Square = 8.38</p> <p>P value = 0.00379798 □---</p> <p>Crude RR for all strata = 1.61</p> <p>Mantel-Haenszel Weighted Relative Risk</p> <p>of Disease, given Exposure = 2.15</p> <p>Greenland/Robins Confidence Limits =</p> <p>1.29 < MHRR < 3.58</p> <p><Enter> for more; F10 to quit.</p>
---	---

Chapter 13. Classification Methods

Statistical Classification assigns subjects to categories in the absence of deterministic information. It can be used to answer questions such as:

- is the subject affected by some disease in the absence of totally reliable criteria (absence of “golden standard” which is too expensive for systematic measurements)? ;
- will the baby survive at birth (we do not know the issue with certitude before delivery)?

These are situations in which a set of predictors either quantitative or categorical can provide probabilistic answers.

Only an outline of classification methods is presented here: Elements of Logistic classification, a bit more on Discriminant analysis.

13.1 Logistic classification

Logistic regression can be used for assigning subjects to categories with a simple rule (with 2 categories):

if $P(\text{cat. } 0) > 50\%$, then assign subject to cat. 0

if $P(\text{cat. } 1) > 50\%$, then assign subject to cat. 1.

Example of logistic classification with data from Armitage (1971) "Statistical Methods in Medical Research"(p.340): prediction of presence versus absence of hemolytic disease from measurements of hemoglobin and bilirubin in a sample of 79 babies, among which 63 survived and 16 deceased. (See SPSS output 13.1).

Hemoglobin is significant as predictor ($p=.0006$) whereas bilirubin is not ($p=.1675$). However both were used for classification for the sake of comparison with Discriminant analysis (see below).

Outcome is coded 1 for survival and 0 for decease. Overall PCC (Percent Correct Classification) is 92.41%. Sensitivity is 75.00%, specificity is larger and amounts to 96.93%. As a rule, **specificity** is larger than **sensitivity** with automatic classification when prevalence is below 50% in the sample.

The reason therefore is that the method gives more weight to the largest subsmample (here survivors) because it has a larger influence on goodness-of-fit. (Trivial situation: if there were only 1 deceased then predicting survival for all 79 subjects would give 0% sensitivity but 100% specificity and 98.7 (78/79)PCC).

Sensitivity / specificity balance cannot be modified in SPSS Logistic regression procedure. This is however possible in SPSS Discriminant Analysis, by giving prevalence a value different from the sample value. Then the larger the prevalence, the larger the sensitivity vs. specificity. Technical aspects of prevalence manipulation in SPSS Discriminant procedure are given below (see “Priors” modification).

SPSS Output 13.1

LOGISTIC REGRESSION outcome

-> /METHOD=ENTER bili haemo

-> /CRITERIA PIN(.05) POUT(.10) ITERATE(20) .

Total number of cases: 79 (Unweighted)
 Number of selected cases: 79
 Number of unselected cases: 0

Number of selected cases: 79
 Number rejected because of missing data: 0
 Number of cases included in the analysis: 79

Dependent Variable Encoding:

Original Value	Internal Value
,00	0
1,00	1

Dependent Variable.. OUTCOME

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 79,614946

* Constant is included in the model.

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number

1.. BILI
 HAEMO

Estimation terminated at iteration number 5 because
 Log Likelihood decreased by less than ,01 percent.

-2 Log Likelihood 39,989
 Goodness of Fit 206,075

	Chi-Square	df	Significance
Model Chi-Square	39,626	2	,0000
Improvement	39,626	2	,0000

Classification Table for OUTCOME

		Predicted		Percent Correct
		,00 0	1,00 1	
Observed				
,00	0	12	4	75,00%
1,00	1	2	61	96,83%
Overall				92,41%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
BILI	-,4917	,3562	1,9056	1	,1675	,0000	,6116
HAEMO	,5343	,1561	11,7173	1	,0006	,3494	1,7063
Constant	-2,3587	2,4790	,9053	1	,3414		

13.2 Discriminant analysis

The purpose of discriminant analysis is to subdivide subjects into 2 or several groups as a function of 2 or several measurements (variables). The purpose is thus not to predict a continuous value (regression), but well to predict the subject's category.

(Note: The problem makes sense only if a classification independent of the predictors under study is available. Examples: authoritative but more arduous predictors; after a laps of time, nature itself makes the difference between e.g. damage and recovery).

There are two steps towards the solution: First, to find the "best" linear combination of variables for doing the job. Second, to find the best criteria for separating the subjects on the basis of the combined values.

First step: The appropriate linear combination of variables is called the "LINEAR DISCRIMINANT FUNCTION":

$$y = a + b_1x_1 + b_2x_2 + \dots\dots\dots b_nx_n$$

(y = discriminant score = D in SPSS)

Procedure for optimizing the b coefficients: the squared difference between the y scores of the 2 categories (let us suppose for the while that the subjects must be classified in only 2 categories) is maximized by comparison with the intra-categorical variance of the scores. In other words: maximization of the corresponding t^2 or F ratio:

$$\text{EIGENVALUE} = V^2 = \frac{(\overline{m}_{y\text{cat}1} - \overline{m}_{y\text{cat}2})^2}{\text{intragroup } s_y^2} = \frac{\text{between group var.}}{\text{within group var.}}$$

Related discrimination coefficients:

$$\text{ETA}^2 = \frac{\text{between group SS}}{\text{total SS}} = \text{\% explained var.}$$

$$\text{WILK'S LAMBDA} = \frac{\text{within group SS}}{\text{total SS}} = \text{\% of residual var.}$$

Wilk's lambda decreases for better discrimination and we have:

$$\text{Eta}^2 + \text{Wilk's lambda} = 1$$

When group means are equal (no discrimination at all), $\text{Eta}^2 = 0$ and Wilk's lambda = 1.

Application condition: hypothesis of equal variances and covariances of predictors (x variables) within groups.

(Mathematical procedure: solution of a set of n equations with n unknowns including the intra group variances and covariances).

Second step: we are now back to a classification problem with a single variable. The y value to be used as boundary between categories will determine the sensitivity and specificity of the classification. If we want to equalize these 2 coefficients, the criteria should be placed halfway between the means of the 2 categories ($y_0 = (\text{my}_{\text{cat1}} + \text{my}_{\text{cat2}})/2$). (Note: simple rule here because equal variances). For other values of sensitivity and specificity, the Normal table can be used under the assumption that y is Normally distributed within each category, a condition which is fulfilled if the 2 intra-group (multivariate) distributions of y are Normal.

In the SPSS program, the discriminant boundary (y_0) corresponds to a y value such that the posterior probability of belonging to one group (e.g. diseased or D+) is equal to the posterior probability of belonging to the other group (e.g. not diseased or D-):

$$p(D+/y_0) = p(D-/y_0).$$

With this boundary value, an item is classified in the group for which the posterior probability is the largest. Where is this boundary located? This depends on the prior probabilities of the groups. If priors are equal (PRIORS EQUAL in SPSS), then the boundary is located halfway between the group means. If the sizes of the groups in the sample are taken as estimations for the priors (PRIOR SIZE in SPSS), then the boundary is closer to the mean of the smallest group. If $P(D-) > P(D+)$, this implies that sensitivity will be lower than specificity.

Mathematical development:

See p.30: Prior probability is for the unconditional realization of an event (e.g. for a disease: prevalence) whereas posterior probability is for the conditional realization (e.g. diseased if test is positive).

If priors are equal ($P(d+) = P(d-)$), then the application of Bayes' Theorem allows to state that:

$$p(D+/y_0) = p(D-/y_0) \text{ implies } p(y_0/D+) = p(y_0/D-)$$

This means that the boundary is located at the intersection between the probability distributions of $D+$ and $D-$ over y .

If priors are not equal (e.g. $P(d+) < P(d-)$), then:

$$p(D+/y_0) = p(D-/y_0) \text{ implies } p(y_0/D+) * p(D+) = p(y_0/D-) * p(D-)$$

$$\text{and: } p(y_0/D+) = p(y_0/D-) * p(D-) / p(D+)$$

$$\text{as: } p(D-) / p(D+) > 1$$

$$\text{we have: } p(y_0/D+) > p(y_0/D-)$$

This means that the boundary is located closer to the mean of the probability distribution of $D+$ than to the mean of the $D-$ distribution.

Generalization to several categories

Several discrimination functions may be required because a single function is usually not optimal for separating all the categories.

Suppose a first discriminant function has been obtained. The second function will be the one which is not correlated with the first and which, together with the first, provides the best separation between groups. And so on for the third, fourth ... function. The maximal number of discriminant functions is equal to the number of categories minus 1 (that is: $k-1$) or to n if $k-1 > n$.

Example of discriminant analysis from Armitage (1971) "Statistical Methods in Medical Research"(p.340): prediction of presence versus absence of hemolytic disease from measurements of hemoglobin and bilirubin. (see SPSS OUTPUT 13.2)

 Note: 0 = outcome decease, 1 = outcome survival.

EQUATION

discriminant function: $y = -3.18 + .30 \cdot \text{hemoglobin} - .18 \cdot \text{bilirubin}$

boundary line:

$$-1.28 = -3.18 + .30 \cdot \text{haemoglobin} - .18 \cdot \text{bilirubin}$$

where -1.28 is a value of y between mean y values for decease and survival groups and such that:

$$p(\text{decease}/-1.28) = p(\text{survival}/-1.28).$$

The boundary value (-1.28) is obtained by resolving for y in:

$$p(\text{decease}/y) = p(\text{survival}/y).$$

From Bayes' Theorem, this is the same as solving for y in:

$$p(y/\text{decease}) \cdot p(\text{decease}) = p(y/\text{survival}) \cdot p(\text{survival}).$$

Where $p(\text{decease})$ is taken as 16/79 and $p(\text{survival})$ as 63/79 (cf. SPSS PRIORS SIZE).

The suggested rule is:

if $y > -1.28$, diagnosis = survival
 if $y < -1.28$, diagnosis = decease

How to calculate the boundary value?

Let $p(D+) = 16/79$

$$p(y/\text{decease}) \cdot p(\text{decease}) = p(y/\text{survival}) \cdot p(\text{survival})$$

$$p(y/\text{survival})/p(y/\text{decease}) = 63/16$$

from Normal probability density formula (Chapt.3):

$$p(y/\text{decease})/p(y/\text{survival}) = \frac{e^{-Z_d^+/2}}{e^{-Z_d^-/2}}$$

$$(-Z_{D+}^2/2) + (Z_{D-}^2/2) = \ln(63/16)$$

$Z_{D+} = y - (-1.73)$ (where -1.73 is the mean y for D+; see SPSS output 13.2)

$Z_{D-} = y - 0.44$ (where 0.44 is the mean y for D-; see SPSS output 13.2)

$$-(y + 1.73)^2 + (y - 0.44)^2 = 2\ln(63/16)$$

$$-(y^2 + 2(1.73)y + 1.73^2) + (y^2 - 2(0.44)y + 0.44^2) = 2\ln(63/16)$$

$$y = (2\ln(63/16) + 1.73^2 - 0.44^2) / -(2(1.73 + 0.44)) \cong 1.28$$

Formula: boundary = $((2\ln((p(D-)/p(D+)) + my_{D+}^2 - my_{D-}^2) / -(2(-my_{D+} + my_{D-})))$

IF $P(D+) = P(D-)$

THEN

$$\text{boundary} = (my_{D+} + my_{D-})/2$$

That is, the boundary is then located halfway between the means of the two distributions.

SPSS Output 13.2

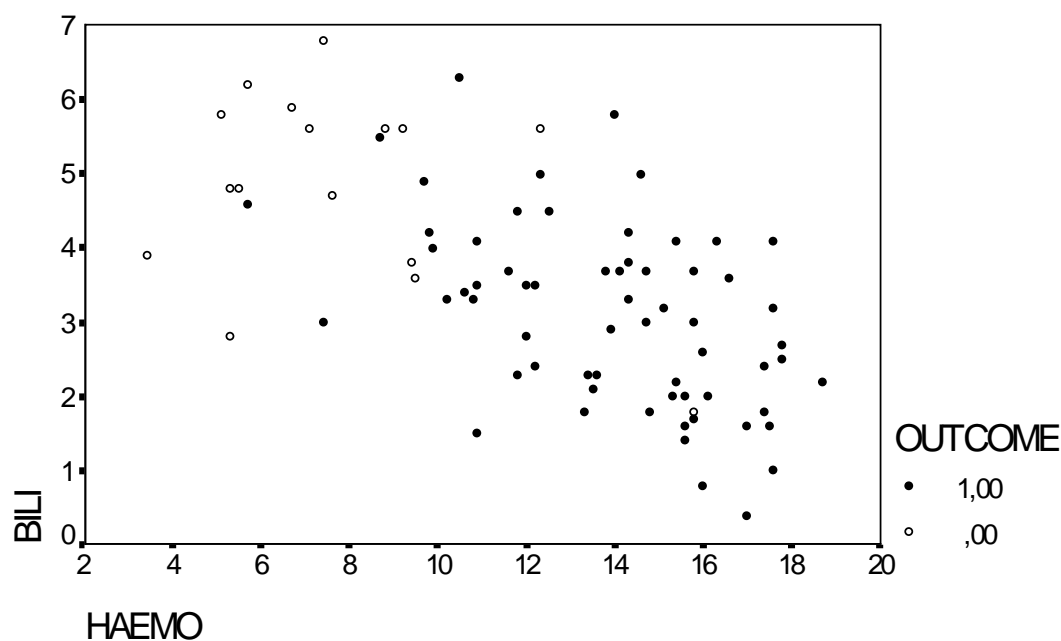


Fig.13.1. 79 cases plotted. As can be seen, hemoglobin level is higher for survivors whereas their bilirubin level is lower.

Example of discriminant analysis on SPSS: same data as above.

```
DISCRIMINANT GROUPS outcome (0,1) / VARIABLES hemo bili
/METHOD wilks /PRIORS size/ STATISTICS 1 2 7 10 11 13 14 15.
```

Since ANALYSIS= was omitted for the first analysis all variables on the VARIABLES= list will be entered at level 1.

```
- - - - DISCRIMINANT ANALYSIS - -
```

79 (unweighted) cases will be used in the analysis.

Number of Cases by Group

OUTCOME	Number of Cases		Label
	Unweighted	Weighted	
0	16	16.0	
1	63	63.0	
Total	79	79.0	

Group Means

OUTCOME	HEMO	BILI
0	7.75625	4.83125
1	13.89683	3.09048
Total	12.65316	3.44304

Group Standard Deviations

OUTCOME	HEMO	BILI
0	3.09170	1.34100
1	2.84168	1.24883
Total	3.79804	1.44264

-----On groups defined by OUTCOME

Analysis number 1

Stepwise variable selection

Selection rule: Minimize Wilks' Lambda
 Maximum number of steps..... 4
 Minimum Tolerance Level..... .00100
 Minimum F to enter..... 1.0000
 Maximum F to remove..... 1.0000

Canonical Discriminant Functions

Maximum number of functions..... 1
 Minimum cumulative percent of variance... 100.00
 Maximum significance of Wilks' Lambda.... 1.0000

-----Prior Probabilities

Group	Prior	Label
0	.20253	
1	.79747	
Total	1.00000	

---Variables not in the analysis after step 0 -----

Minimum

Variable	Tolerance	Tolerance	F to enter	Wilks' Lambda
HEMO	1.0000000	1.0000000	57.522	.57240
BILI	1.0000000	1.0000000	24.074	.76182

-----At step 1, HEMO was included in the analysis.

		Degrees of Freedom	Signif.	Between Groups
Wilks' Lambda	.57240	1 1	77.0	
Equivalent F	57.5215	1	77.0	.0000

----- Variables in the analysis after step 1 Variable Tolerance F to remove Wilks' Lambda

HEMO	1.0000000	57.522	
------	-----------	--------	--

---- Variables not in the analysis after step 1 -----

Minimum

Variable	Tolerance	Tolerance	F to enter	Wilks' Lambda
BILI	.7886300	.7886300	1.4438	.56173

Note for $p=.05$, $F(1,77) = 3.95$; thus 1.44 NS. However, BILI is included in the model **because $F > 1$ (default option in SPSS, previous VERSIONS)**.

At step 2, BILI was included in the analysis.

		Degrees of Freedom	Signif.	Between Groups
Wilks' Lambda	.56173	2 1	77.0	
Equivalent F	29.6484	2	76.0	.0000

note $\eta^2 = 1 - .56 = .44 = \% \text{ explained variance by groups} = \text{between group SS} / \text{total SS}$.

----- Variables in the analysis after step 2 -----Variable Tolerance F to remove Wilks' Lambda

HEMO	.7886300	27.072	.76182
BILI	.7886300	1.4438	.57240

F level or tolerance or VIN insufficient for further computation.

Summary Table

Action	Vars	Wilks'
Step Entered Removed	In	Lambda Sig. Label
1 HEMO	1	.57240 .0000
2 BILI	2	.56173 .0000

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chisquare	DF	Sig
1*	.7802	100	100	.6620	:				

note .66 = eta ($.66^2 = .44$)

* marks the 1 canonical discriminant functions remaining in the analysis.

Standardized Canonical Discriminant Function Coefficients

	FUNC 1
HEMO	.87172
BILI	-.23225

Structure Matrix:

Pooled-within-groups correlations between discriminating variables
and canonical discriminant functions
(Variables ordered by size of correlation within function)

	FUNC 1
HEMO	.97850
BILI	-.63302

Unstandardized Canonical Discriminant Function Coefficients

	FUNC 1
HEMO	.3014171
BILI	-.1832610
(constant)	-3.182906

Note: $.3014171 * \text{sd}(\text{hemo}) = .87172$; $\text{sd}(\text{hemo}) = 2.59 = \sqrt{\text{weighted mean of intragroup variances}}$
 $= \sqrt{((16 * (3.09)^2 + 63 * (2.84)^2) / 79)}$.

The unstandardised coefficient is multiplied by the SD of the predictor (Just as for regression)

Canonical Discriminant Functions evaluated at Group Means (Group Centroids)

Group	FUNC 1
0	-1.73042
1	.43947

Test of equality of group covariance matrices using Box's M

The ranks and natural logarithms of determinants printed are those
of the group covariance matrices.

Group Label	Rank	Log Determinant
0	2	2.719153
1	2	2.255052
Pooled Within-Groups Covariance Matrix	2	2.360296
Box's M	Approximate F	Degrees of freedom
1.1422	.36144	3,
		10372.6
		.7809

note: discriminant scores = $D = y$

if D1 is the highest predicted group, then the highest $P(D/G) = P(y/D1)$, $P(G/D) = P(D1/y)$, and the
2nd highest $P(G/D) = P(D0/y)$.

Case Number	Mis Val	Actual Sel Group	Highest Probability Group	P(D/G)	P(G/D)	2nd Highest Group	P(G/D)	Discrim Scores
1		1	1	.1072	.9993	0	.0007	2.0504

2	1	1	.2120	.9984	0	.0016	1.6875
3	1	1	.1989	.9985	0	.0015	1.7242
4	1	1	.3518	.9968	0	.0032	1.3707
5	1	1	.2730	.9978	0	.0022	1.5356
6	1	1	.1338	.9991	0	.0009	1.9388
7	1	1	.1741	.9987	0	.0013	1.7987
8	1	1	.1962	.9985	0	.0015	1.7319
9	1	1	.2370	.9981	0	.0019	1.6219
10	1	1	.1532	.9989	0	.0011	1.8679
11	1	1	.2269	.9983	0	.0017	1.6480
12	1	1	.4707	.9950	0	.0050	1.1609
13	1	1	.5896	.9926	0	.0074	.9788
14	1	1	.3876	.9963	0	.0037	1.3034
15	1	1	.4692	.9950	0	.0050	1.1633
16	1	1	.2920	.9976	0	.0024	1.4932
17	1	1	.6441	.9912	0	.0088	.9014
18	1	1	.5550	.9933	0	.0067	1.0297
19	1	1	.4074	.9960	0	.0040	1.2679
20	1	1	.4104	.9960	0	.0040	1.2626
21	1	1	.4757	.9949	0	.0051	1.1527
22	1	1	.4316	.9956	0	.0044	1.2260
23	1	1	.7886	.9867	0	.0133	.7075
24	1	1	.5377	.9937	0	.0063	1.0557
25	1	1	.5334	.9938	0	.0062	1.0623
26	1	1	.7319	.9887	0	.0113	.7821
27	1	1	.6109	.9921	0	.0079	.9482
28	1	1	.8963	.9821	0	.0179	.5699
29	1	1	.7959	.9864	0	.0136	.6981
30	1	1	.8902	.9685	0	.0315	.3015
31	1	1	.9932	.9760	0	.0240	.4310
32	1	1	.9348	.9720	0	.0280	.3577
33	1	1	.9338	.9803	0	.0197	.5226
34	1	1	.9598	.9738	0	.0262	.3890
35	1	1	.6416	.9379	0	.0621	-.0260
36	1	1	.9714	.9782	0	.0218	.4753
37	1	1	.8880	.9683	0	.0317	.2986
38	1	1	.9558	.9791	0	.0209	.4949
39	1	1	.9506	.9793	0	.0207	.5014
40	1	1	.9961	.9762	0	.0209	.4961
42	1	1	.4969	.9047	0	.0953	-.2399
43	1	1	.4058	.8723	0	.1277	-.3918
44	1	1	.5575	.9207	0	.0793	-.1470
45	1	1	.7003	.9473	0	.0527	.0546
46	1	1	.6041	.9308	0	.0692	-.0790
47	1	1	.5178	.9106	0	.0894	-.2073
48	1	1	.6261	.9351	0	.0649	-.0477
49	1	1	.3733	.8573	0	.1427	-.4509
50	1	1	.4214	.8787	0	.1213	-.3645
51	1	1	.3279	.8323	0	.1677	-.5389
52	1	1	.2765	.7963	0	.2037	-.6488
53	1	1	.5407	.9166	0	.0834	-.1724
54	1	1	.3311	.8342	0	.1658	-.5324
55	1	1	.2935	.8093	0	.1907	-.6110
56	1	1	.1070	.5565	0	.4435	-1.1726
57	1	1	.2490	.7727	0	.2273	-.7132
58	1	1	.1703	.6790	0	.3210	-.9319
59	1	1	.1504	.6466	0	.3534	-.9987
60	1	1	.1104	.5647	0	.4353	-1.1571
61	1	**	.8714	.6530	1	.3470	-1.5685
62	1	**	.8195	.6197	1	.3803	-1.5022
63	1	**	.5637	.9035	1	.0965	-2.3078
64	0	**	.4179	.9959	0	.0041	1.2496

65	0 **	1	.3466	.8432	0	.1568	-.5017
66	0 **	1	.1560	.6562	0	.3438	-.9792
67	0 **	1	.1374	.6228	0	.3772	-1.0460
68	0	0	.7685	.5854	1	.4146	-1.4361
69	0	0	.8621	.6472	1	.3528	-1.5567
70	0	0	.9816	.7376	1	.2624	-1.7535
71	0	0	.6397	.8808	1	.1192	-2.1986
72	0	0	.7348	.8479	1	.1521	-2.0691
73	0	0	.6071	.8909	1	.1091	-2.2447
74	0	0	.3840	.9465	1	.0535	-2.6010
75	0	0	.5001	.9203	1	.0797	-2.4048
76	0	0	.4626	.9294	1	.0706	-2.4650
77	0	0	.7128	.8560	1	.1440	-2.0985
78	0	0	.3280	.9571	1	.0429	-2.7086
79	0	0	.2533	.9696	1	.0304	-2.8728

Classification Results

Actual Group	No. of Cases	Predicted Group Membership	
		0	1
Group 0	16	12 75.0%	4 25.0%
Group 1	63	3 4.8%	60 95.2%

Percent of "grouped" cases correctly classified: 91.14%

MEANS TABLES = discore1 by outcome / STATISTICS 1.

Summaries of DISCORE1 FUNCTION 1 FOR ANALYSIS 1
By levels of OUTCOME

Variable	Value	Label	Mean	Std Dev	Cases
For Entire Population			6.7457E-16	1.3256692	79
OUTCOME	0		-1.7304188	1.0420656	16
OUTCOME	1		.4394715	.9895542	63

Total Cases = 79

Summaries of DISCORE1 FUNCTION 1 FOR ANALYSIS 1
By levels of OUTCOME

Value	Label	Mean	Std Dev	Sum of Sq	Cases
0		-1.7304188	1.0420656	16.2885095	16
1		.4394715	.9895542	60.7114905	63

Within Groups Total		6.7457E-16	1.0000000	77.0000000	79
NOTE		weighted mean about 0	weighted variance about 1		

$$(16*(1.04)^2 + 63*(.98)^2) / 79 = \text{about } 1$$