Tentamen Toegepaste Statistiek voor studenten Informatiekunde
vrije Universiteit, Faculteit Exacte Wetenschappen
29 maart 2005, 9.30-12.30u.

1. a) In a swedish study concerning hundreds of twins, it was determined that the level of mental skills was more similar in identical twins (twins coming from single egg) than in fraternal twins (twins coming from two separate eggs). What type of study is this? Motivate your answer.

   b) Use an example to explain what stratified sampling is.

   c) One hundred people are selected for a study of the effect of regular walking on heart rate. Each person is allowed to choose whether to be in the walker or non-walker group. The resting heart rate of each individual is recorded. After six weeks, the resting heart rate of each individual is compared with his / her initial heart rate. What is wrong with this set-up? Describe an alternative set-up of this study that is more adequate.

   d) Use an example to explain the concept *confounding.*

   e) Using numerous studies on the effect of early birth on later development of children, the conclusion is drawn that early birth has a negative effect on this development. What type of study is conducted here? Motivate your answer.

2. a) At some place the CO_2 level in the air is measured daily. What type of data comes from this process and on which measurement level?

   b) At a call center, telephone calls are taken care of for a number of different companies. Over a period of one year, the telephone calls are classified by the company they are addressed to. Which type of data comes from this process and what is the level of measurement?

3. Below are the lengths of sixteen year old people, given in centimeters.

163	180	170	164	170	164	185	162	167
167	159	171	161	168	178	164	168	173
175	195	158	162	171	159	164	180	182
160	158	158	181	160	165	174	169	158

   a) Give a table with four columns. In the first column, indicate the cells, starting with 155-160 and ending with 190-195; in the second column give the corresponding frequency; in the third column the relative frequency and en in the fourth column the cumulative relative frequency.

   b) Give, based on this table, a sketch of the histogram of the data and describe the shape of this histogram qualitatively.

   c) Give the modal cell and the median of the data.

4. Indicate which visualization technique is most appropriate in the situations described below. Motivate your answer.

   a) A producer of ice creams has produced four types of cream for already ten years. He is particularly interested in the market share of his four products with respect to each other over the past ten years.

   b) There are about 40 brands of 1.5V batteries in the market. These batteries are sold in various countries. Someone is interested in the differences in sales of the various brands over the countries The Netherlands and Indonesia.

   c) At a company selling natural stones, a new batch of stones arrives on a certain day. All stones are classified in classes numbered 1 to 5 (class 1-stones are the best, then 2 etc.). One wants to have an overval view of the quality of the arrived stones.

5. In the USA, the so-called SAT score is used to assess the ability of school children. Below SAT scores of 25 school children are given.

510	686	601	388	504
366	473	483	510	424
553	503	455	533	314
528	677	673	568	251
411	451	611	585	527

   a) Based on relative frequencies, give an estimate of the probability that a randomly selected school child has a SAT score below 450.

   The SAT score is a standardized score, that has a normal distribution with population mean $\mu = 500$ and standard deviation 100.

   b) Based on this fact, give the probability that the score of a randomly selected student is lower than 450. Table 1 at the end of this exam can be of use.

6. An honest coin is thrown twice. The individual outcomes (scores, points) are coded by -1 en +1.

   a) Construct a probability space for this experiment.

   b) Compute the probability that the score in the second throw is at least as favorable as in the first throw. Do this by formulating the event properly in the probability space constructed under a).

   c) Consider the random variable X that represents the total gain in this experiment. Construct the probability mass function for X based on its formal definition.

   d) Compute the expectation EX of X.

e) Formally describe the events 'first throw is +1' and 'second throw is -1' in the probability space and show (by the formal definition of independence) that these events are independent.

7. A company is interested in the percentage of people that usually watches a certain TV program. To get information on this, $n = 500$ people were randomly selected and 55 of these admitted to watch this program regularly.

   a) Give the usual point estimate $\hat{p}$ for the unknown percentage p that usually watches the program.

   The central limit theorem is a theorem that in this current situation states that for large sample size n the quantity

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}$$

   is approximately normally distributed.

   b) Derive, based on this result, the *general expression* for a 95% confidence set for p.

   c) Give the 95% confidence set for p in the current situation and describe the interpretation of this interval.

8. Under 5), the SAT scores of 25 school children are given. Based on these scores, we want to investigate the claim that the SAT score is standardized, with population mean 500. It is given that the mean of the sample is 504 and the standard deviation is 100.

   a) Formulate the null hypothesis H_0 and the alternative hypothesis H_a with respect to the mean SAT score of all school children in the USA.

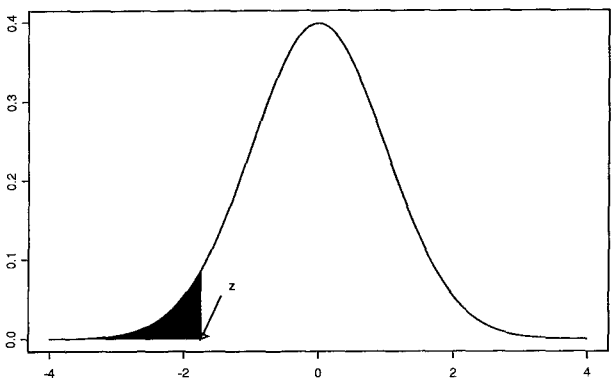
   b) Compute the standard score z (test statistic) based on the data.

   c) Compute the P-value corresponding to this z-score (table 1 can again be of use).

   d) Describe what happens if an *error of type 2* is made in this testing problem.

z	Area under the curve *to the left of* z
-2.932	0.002
-1.960	0.025
-0.500	0.309
0.040	0.516
0.250	0.599
1.188	0.883
1.645	0.950
1.960	0.975

Tabel 1: *areas under the standard normal curve of figure 1*



Figuur 1: *standard normal curve corresponding to table 1*

1a: 3	1d: 3	2b: 3	3c: 3	4c: 3	6a: 3	6d: 3	7b: 3	8b: 3
1b: 3	1e: 3	3a: 6	4a: 3	5a: 3	6b: 3	6e: 3	7c: 3	8c: 6
1c: 3	2a: 3	3b: 3	4b: 3	5b: 3	6c: 6	7a: 3	8a: 3	8d: 3

Tabel 2: *points to be earned for each part*

Grade=(number of points + 10)/10

Succes!