Practice test for Språkstatistik (HT97)

The material covered in this test are the chapters 3-6, 8-11, 13-18 and 26 from the book by Freedman et.al. plus the lecture notes for the second half of the course. You need to be able to compute averages, standard deviations, correlation coefficients and the rms error on your own. Formulas that are necessary for other computations will be supplied at the test.

No question involves massive computation so if you discover that you seem to need to perform one try something else or answer the question with an argumented guess.

This test contains 10 questions which contain two parts. Each part is worth 5 points. The maximal score is 100 points.

1. A box contains a set of cards each of which contains a number. The numbers are distributed according to the normal curve with an average of 100 and a standard deviation of 10.
   a. What SD-value corresponds with number 130? And what number corresponds with SD-value -2.5?
   b. We draw one card from the box. What is the probability that it contains a number between 90 and 130?
2. We are interested in the relation between the marks of students for test A and test B. We collect data for these two tests and find out that average(A)=8.0, SD(A)=1.0, average(B)=6.0, SD(B)=2.0 and r=0.6.
   a. What average B-mark do you expect for a student that got a 7 for test A?
   b. What average A-mark do you expect for a student that got a 7 for test B?
3. A box contains four red cards and one blue card.
   a. We make three draws with replacement from the box. What is the probability that the third card is blue?
   b. We make three draws without replacement from the box. The first card is red but the color of the second card is hidden. What is the probability that the third card is blue?
4. A box contains one green marble, two yellow marbles and three orange marbles. We make six draws from the box with replacement.
   a. What is the probability of obtaining exactly two yellow marbles in these six draws?
   b. Can you use the same formula as used in a. for computing the probability of obtaining exactly two yellow marbles in six draws **without** replacement?
5. In a simple gambling game you have 20% chance of winning 9 dollars and 80% of losing one dollar. You play the game 100 times.
   a. Compute the expected value and the standard error for your gain.
   b. What is the probability that you win more than 40 dollar?

6. You play a die game with a friend. The goal is to throw many sixes. Each of you throw the die 180 times.
   a. How many sixes do you expect for you and your friend? What is the standard error of this expectation?
   b. Your friend throws 50 sixes. Can the difference between this number and the result of the previous question be explained by a chance error?
7. a. We write a program that tries predict the next character in a text in some language. The program can use either unigram statistics, bigram statistics or ten-gram statistics. Which of these statistics will make the program perform better? Why?
   b. The mutual information of the the word pair *strong currents* is 9.04 while the mutual information of the the word pair *strong thunderstorm* is 8.84. What do these numbers mean in relation to each other?
8. a. Explain how the noisy channel model can be used in a part of speech tagger.
   b. Draw a Hidden Markov Model and explain the parts of the model.
9. a. Give both an advantage and a disadvantage of stochastic grammars.
   b. Give a reasonable application for which clustering could be used successfully.
10. a. The sentence alignment method of Gale and Church restricts the number of possible sentence alignment structures. Why? And does it have a negative influence on the performance of the method?
    b. Similarity methods and frequency methods are two groups of statistical techniques which can be used for extraction of translation equivalents from a sentence aligned corpus. Describe the difference between the two technique groups.

Normal table

z	Height	Area
0.00	39.89	0.00
0.15	39.35	11.92
0.25	38.67	19.74
0.50	35.21	38.29
1.00	24.20	68.27
1.50	12.95	86.64
2.00	5.40	95.45
2.50	1.75	98.76
3.00	0.44	99.73
3.50	0.09	99.95
4.00	0.01	99.99

Binomial formula

$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$

Answers practice test for Språkstatistik (HT:97)

1. a. For numbers with an average of 100 and an SD of 10 an SD value of 3.0 corresponds with number 130 and number 75 corresponds with SD value -2.5.

   b. 90 corresponds with -1.0 SD and 130 corresponds with +3.0 SD. So the probability is equal to half of the area under -1.0 SD and +1.0 SD plus half of the area under -3.0 SD and +3.0 SD. According to the table the first area is 68.27 and the second area is 99.73. The answer is thus (68.27+99.73)/2 = 84.00%.

2. a. A 7 for test A corresponds with -1.0 SD. So we expect a score for test B of 6.0 + -1.0*2.0*0.6 = 4.8.

   b. A 7 for test B corresponds with +0.5 SD. SO we expect a score for test A of 8.0 + 0.5*1.0*0.6 = 8.3.

3. a. The draws are independent. There is a chance of 1 out of 5 of drawing the blue card so the probability is 0.20.

   b. Since the first card is red there are only three red cards left. The blue card might be the second draw or the third one or one of the cards that remained in the box. The chance that it is the third draw is 1 out of 4 so the probability is 0.25.

4. a. We use the binomial formula with the values n=6 (total number of draws), p=2/6 (probability of drawing a yellow marble) and k=2 (number of yellow marbles we want to obtain). The result is $(6!/2!4!)*(2/6)^2*(4/6)^4 = (5*3)*(1/3)^2*(2/3)^4 = 15*1*16/(3^6) = 80/243$

   b. No, the binomial formula cannot be used for computing probabilities in draws without replacement because in these draws p is not constant.

5. We can represent this game as making draws with replacement from a box with one card with number 9 and four cards with number -1. The gain corresponds with the sum of the numbers in the draws.

   a. Expected value = 100*(average box) = 100*1.0 = 100.
   Standard error = sqrt(100) * SD(box) = 10 * 4 = 40. I have used the short cut formula for computing the SD of the box: SD(box) = (9-(-1))*sqrt((1/5)*(4/5)) = 10*sqrt(4/25) = 4.

   b. We may assume that the gain is distributed according to a normal curve with average 100 and SE 40. Winning 40 dollar or more corresponds with the area to the right of 40 dollar = -1.5 SE. This is equal to the area in the SD table for 1.5 SD plus half of the remaining value (for the right tail). The answer is thus: 86.64+(100-86.64)/2 = 93.32%.

6. We can represent this game as making draws with replacement from a box with one card with number 1 and five cards with number 0. The number of sixes corresponds with the sum of the numbers in the draws.

   a. Expected value = 180*(average box) = 180*(1/6) = 30.
   Standard error = sqrt(180) * SD(box) = sqrt(180) * sqrt(5/36) = sqrt(180*5/36) = 5. I have used the short cut formula for computing the SD of the box: SD(box) = (1-0)*sqrt((1/6)*(5/6)) = sqrt(5/36).

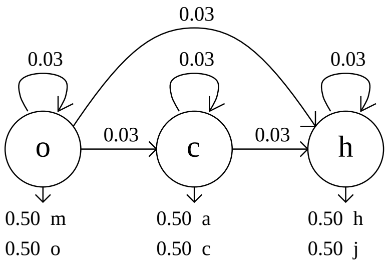
b. When we expect 30 sixes with a standard error of 5 then 50 sixes corresponds with +4.0 SE. According to the normal curve table the probability that this value is caused by a chance error is 0.01%. This is a highly significant result (<1.0%). The 50 sixes cannot be explained by a chance error.

7. a. Bigram statistics will give make the program perform better than unigram statistics because the bigram statistics pay attention to the context of the character and unigram statistics don't. Ten-gram statistics take more context in consideration but it will be almost impossible to collect meaningful statistics for ten-grams because there are so many of them. Thus bigram statistics will work out best.

   b. The two numbers mean that when we take in consideration the frequencies of the unigrams *strong*, *currents* and *thunderstorm* the bigram *strong currents* is relatively more frequent than the bigram *strong thunderstorm*.

8. a. We imagine that the words in a text are the observation of a message that contained the part of speech tags of the words and a noisy channel which converted the part of speech tags to words. If we apply the noisy channel model to reconstruct the original message from the observed words then it will function as a part of speech tagger.

   b. A Hidden Markov Model consists of cells which are connected with each other by links that contain weights. The weights give the probabilities for going from one cell to another (or the same) cell. Each cell can produce strings. The strings are produced according to probabilistic rules.



9. a. An advantage of stochastic grammars is that they can deal with long distance dependencies.
   A disadvantage of stochastic grammars is that they cannot be derived from an untagged corpus.

   b. Clustering can be used for deriving the class of unambigious words in a partially tagged corpus.

10. a. The alignment method restricts the number of alignment structures to keep down the number of computations that have to be performed. Certain alignment structures cannot be recognized by the algorithm. Its error percentage will go up but only with a very small amount since the neglected alignment structures are infrequent.

    b. Similarity techniques use the resemblances between words for grouping them together (for example the Swedish *adress* and the English *address*). Frequency methods use the distribution of the words in the corpus for grouping them together.