Practice test for Språkstatistik (HT:95,951019)

You are allowed to use all available literature and course notes while making this 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.

1. A normal deck of cards contains 52 cards of which 13 are hearts. We shuffle the deck and take out five cards.
   (a) What is the probability that the first card is an ace of hearts?
   (b) What is the probability that the fifth card is an ace of hearts?
   (c) What is the probability that the fifth card is a heart?
   (d) What is the probability that the fifth card is a heart given that the first four are a heart?
2. We have a box with one blue ball and five yellow balls. We make four draws at random *with* replacement. Give the probability for the event that
   (a) the blue ball is never drawn.
   (b) the blue ball is drawn twice.
   (c) the blue ball is drawn all times.
   (d) the blue ball is drawn at least two times.
3. At a roulette table you can bet SEK 1.00 on one out of 38 numbers. If the number appears you will receive SEK 35.00 plus your own bet otherwise you will lose your bet.
   (a) Compute the average and the standard deviation for your profit if you make only one bet.
   (b) Compute the average and the standard deviation for your profit if you make 100 bets.
   (c) What do you prefer to do here: making one bet or making 100 bets?
4. A box contains 6 cards with numbers: 2, 2, 2, 3, 3 and 6. We make 1000 draws with replacement from the box and we are interested in the sum of the numbers of the cards we have drawn.
   (a) What distribution is best for describing this sum?
   (b) Give the formula describing this sum. Assign number values to as many variables as possible.
5. An election will be held between two candidates A and B. We hold a pool among 100 people and we discover that 60 of them favor candidate A and 40 favor candidate B. Give the 95% certainty interval for the result of the election.
6. In a noisy channel model we can choose for using different context models, for example the unigram model (no context), the bigram model and the trigram model.
   (a) Why will the bigram model in general perform better than the unigram model?
   (b) Why will the bigram model in general perform better than a 25-gram model?
7. When we are clustering data, we make numeric representations of tokens. In the class and in the practical exercise we have used a representation of data that was based on bigrams. Make a trigram representation for the words that occur in the sentence: *to be or not to be.*

8. A stochastic grammar contains the rules:
0.50 S→ASB
0.25 S→BSA
0.15 S→AB
0.10 S→BA
1.00 A→a
1.00 B→b

   (a) Compute the probability that this grammar will generate *ab*
   (b) Compute the probability that this grammar will generate *abab*

Answers practice test for Språkstatistik (HT:95,951019)

1. (a) There is only one ace of hearts in 52 cards: $\frac{1}{52}$.
   (b) We have no information about the first four cards. The ace of hearts may be among them so since there is only one ace of hearts in 52 cards: $\frac{1}{52}$.
   (c) We have no information about the first four cards. The may be hearts so since there are 13 hearts in 52 cards: $\frac{13}{52} = \frac{1}{4}$.
   (d) The first four cards are hearts. When we draw the fifth card there are only nine hearts in a deck of 48 cards: $\frac{9}{48} = \frac{3}{16}$.

2. With replacement the probability of drawing the blue ball is $\frac{1}{6}$ and the probability of not drawing it is $1 - \frac{1}{6} = \frac{5}{6}$.

   (a) The probability of never drawing the blue ball is $\binom{4}{0} (\frac{1}{6})^0 (\frac{5}{6})^4 = (\frac{5}{6})^4$. Note that the $\binom{4}{0}$-term is necessary in case more than one combination of balls is possible.
   (b) The probability of drawing the blue ball twice is $\binom{4}{2} (\frac{1}{6})^2 (\frac{5}{6})^2$.
   (c) The probability of drawing the blue ball all times is $\binom{4}{4} (\frac{1}{6})^4 (\frac{5}{6})^0 = (\frac{1}{6})^4$.
   (d) The probability of drawing the blue ball at least two times is: $\binom{4}{2} (\frac{1}{6})^2 (\frac{5}{6})^2 + \binom{4}{1} (\frac{1}{6})^3 (\frac{5}{6})^1 + (\frac{1}{6})^4$.

3. The outcome of one bet can be described by a stochastic variable with 38 outcomes: 37 times -1 and 1 time +35. All outcomes have even probability. An easier stochastic variable is one that with two outcomes: -1 with probability $\frac{37}{38}$ and +35 with probability $\frac{1}{38}$

   (a) One bet:
   $\mu_1 = \sum x_i * P(x_i) = \frac{-2}{38} = -\frac{1}{19} = -0.053$
   $\sigma_1^2 = \sum (x_i - \mu)^2 * P(x_i) = (-1 + \frac{1}{19})^2 * \frac{37}{38} + (35 + \frac{1}{19})^2 * \frac{1}{38} = 33.2$
   (b) One hundred bets: With this amount of bets we may assume that the profit is distributed according to a normal distribution. In a normal distribution the average and the variance over *100*x* experiments are 100 times as large the average and the variance over x experiments. We may therefore assume that: $\mu_{100} = 100 * \mu_1$ and $\sigma_{100}^2 = 100 * \sigma_1^2$.
   (c) On average we lose 100 times as much money with hundred bets compared with one bet. Therefore we should prefer one bet.

4. (a) The number of draws is large so we'll choose the normal distribution.
   (b) The sum is a stochastic variable ξ_1 with a normal distribution. The frequency function for this variable is:

   $$f_{\xi_1}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

   We are looking for the sum and the variance of a random sample ξ_0 which describes the outcome of *one* draw from the box. In this case ξ_1=1000*ξ_0, $\mu_1 = 1000 * \mu_0$ and $\sigma_1^2 = 1000 * \sigma_0^2$. We will estimate μ_0 and the σ_0 values:

   $$\mu_0 = \bar{\xi_0} = \sum_i \xi_{0,i} * P(\xi_{0,i}) = (2 * \frac{3}{6} + 3 * \frac{2}{6} + 6 * \frac{1}{6}) = 3$$
   $$\sigma_0^2 \simeq s_0^2 = \frac{1}{n-1} \sum (\xi_{0,i} - \bar{\xi_0})^2 = \frac{1}{5}(3 * (2-3)^2 + 2 * (3-3)^2 + (6-3)^2) = 2.4$$

Therefore μ_1=3000 and $\sigma_1 = 1549$ (=1000*$\sqrt{2.4}$)

5. The formula for computing the interval for the election result is:

$$\bar{x} - u_{0.5*\alpha,n}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + u_{0.5*\alpha,n}\frac{\sigma}{\sqrt{n}}$$

$\bar{x}$ is the estimated average of the result and $u_{0.5\alpha}\frac{\sigma}{\sqrt{n}}$ describes the error. We will assume that the choice of a person is a stochastic variable that takes value 1 if the person supports candidate A and 0 if the person supports candidate B. In that case:

$\bar{x} = \frac{60}{100} = 0.6$
$0.5 * \alpha = 0.5 * (1 - 0.95) = 0.025$
$\sigma^2 \simeq s^2 = \frac{1}{n-1}\sum(\xi_i - \bar{\xi_1})^2 = \frac{1}{99} * (60 * 0.4^2 + 40 * 0.6^2) = 0.24$ (σ=0.49)
$u_{0.5*\alpha,n} = u_{0.025,100} = 1.98$ (closest value in table)
$u_{0.5*\alpha,n}\frac{\sigma}{\sqrt{n}} = 1.98 * \frac{0.49}{10} = 0.097$

So: 0.503< μ <0.697. Note that forgot to put part of the table in the question. You do not need to bring the table to the test.

6. (a) In general the bigram model will perform better than the unigram model because examines a context of the token that has to be corrected. In general the context of a token contains information that can be used to guess the token.
   (b) In general the bigram model will perform better than a 25-gram model because the context information that can be obtained from a 25-gram model is too specific. Even in a large body of data most 25-grams will be included only once. This means that the information that can be learned from a 25-gram can only be used for one specific part of the data and not for any new unseen data.

7. A trigram contains two tokens as context. We will choose the two preceding words as context and insert two dummy words in the beginning of the sentence to give the first two words a context.

	dummy_1-dummy_2	dummy_2-to	to-be	be-or	or-not	not-to	to-be
to	1	0	0	0	1	0	0
be	0	1	0	0	0	1	0
or	0	0	1	0	0	0	0
not	0	0	0	1	0	0	0

The representation of each words is equal to the row on the right of the word.

8. The probability of the sentences is equal to the probability of the rules used in their derivation(s):
   (a) P(ab)=0.15*1.00*1.00=0.15, since S→AB and two 1.00 rules have been used.
   (b) P(abab)=0.50*0.10*1.00*1.00*1.00*1.00=0.05, since S→ASB, S→BA and four 1.0 have been used here.

0.50 S→ASB
0.25 S→BSA
0.15 S→AB
0.10 S→BA
1.00 A→a
1.00 B→b