INF5830, Exercises, 3 Sept.

Seminar room Java

Exercise 1

Say we have a population with 4 numeric observations: {3, 5, 5, 11}

1. What is the median, mode and mean?
2. What is the variance and standard deviation?

1. Mode:5, Median: 5, Mean: (3+5+5+11)/4=6
2. Var: ((3-6)**2 + (5-6)**2 + (5-6)**2 + (11-6)**2)/4=(3**2+1+1+5**2)/4=(9+1+1+25)/4=36/4=9
   s.d: sqrt(9) = 3

Exercise 2 – Outcomes and sample spaces. ("utfall og utfallsrom")

We will conduct some simple experiments. You should try to define suitable sample spaces for the following:

1. Form a sequence of the words *Kari*, O*la*, *liker*, where each word occurs exactly once.
2. Form a sequence of the same three words where we allow repetitions.
3. Read a sentence and determine whether it contains a conjunction.
4. Read a sentence less than 100 words long and count the number of verbs.
5. Listen to a person speak and count how many words she utters before the first occurrence of the personal pronoun "I".

1. Space
   a. Kari liker Ola
   b. Kari Ola liker
   c. Ola liker Kari
   d. Ola Kari liker
   e. Liker Kari Ola
   f. Liker Ola Kari
2. (Ola|Kari|liker)* (i.e. the infinite set of all finite strings over the alphabet of the three names)
3. {Yes, No}
4. {0, 2, …, 100}
5. {0, 1, …, n, …} Infinite

Exercise 3

Suppose that we know that a sentence chosen at random has a 0.3 probability of containing a conjunction and a 0.4 probability of containing a pronoun.

1. If we assume independence between containing a pronoun and a conjunction, what is the probability that a sentence contains both a pronoun and a conjunction?
2. And what is the probability it contains a pronoun but no conjunction?

3. It turns out that the probability for containing both a pronoun and a conjunction is 0.2. Are the two events independent.
4. What is the probability that a sentence which contains a pronoun also contains a conjunction?

1. $P(A)=0.3$, $P(B)=0.4$, $P(A \cap B)=P(A)*P(B)=0.12$
2. $P(-A \cap B)=P(-A)*P(B) = (1-P(A))*P(B) = 0.28$
3. No
4. $P(A|B)= P(A \cap B)/P(B)=0.2/0.4=0.5$

Exercise 4

Rare events may be surprisingly hard to observe correctly. Take for example the Norwegian word "med". It is a frequently occurring preposition. What many Norwegians do not know, is that it is also a rare noun. For this example, let us assume that 1 of 1000 occurrences is a noun.

Kari has made a tagger for Norwegian. Because "med" is rare as a noun, she has decided to tag all occurrences as preposition, and disregard the word as a noun.

Ola, however, does think this is not good enough. He thinks it is cheating. He has therefore made a classifier which for each occurrence of "med" tells whether it is a preposition or a noun. This classifier identifies at least 99% of the preposition occurrences as prepositions and at least 99% of the noun occurrences are identified as nouns. Ola thinks it will be better to augment the tagger with his classifier.

Which strategy do you support? Suppose we evaluate the tagger by counting its correct decisions. Which tagger will score best, Kari's or Ola's?

A – "med" is a preposition
-A – "med" is a noun
B – "med" is classified as a preposition
-B – "med" is classified as a noun

How often is Ola's classifier correct?
$P(A \cap B)+ P(-A \cap -B)=P(B|A)P(A)+P(-B|-A)P(-A) \leq P(B|A)+P(-A)= 0.99+0.001=0.991$
if $P(B|A)$ and $P(-B|-A)$ are 0.99 (We don't know that they are any better.)

How often is Kari's classifier correct?
$P(A \cap B)+ P(-A \cap -B)=P(B|A)P(A)+P(-B|-A)P(-A) \geq P(B|A)P(A) \geq 1* 0.999=0.999$

Exercise 5

Consider the sample space of all English wordforms. We may define several stochastic variables from this sample space. One categoric stochastic variable is the part-of-speech or word class of the wordform with

value space: {Noun, Verb, …}. One numeric stochastic variable is the number of characters in the wordform.

1. Define two other categoric stochastic variables and specify the value space for each of them
2. Define two other numeric random variables and specify their value space.

1. For example
    a. {Belongs to closed word class, Belongs to open word class}
    b. Etymologic descendant from {Latin, French, Old Norse, ..}
    c. First letter, {a,b,c,…,z}
2. For example
    a. Zipf rank in a given corpus {1,2, …,n}
    b. Frequency in a given corpus {0, 1, …, n}
    c. Relative frequency in a given corpus [0,1] (all reals in the interval)

Exercise 6

Consider the space of all sequences of English words.

1. Define three different categoric stochastic variables on this space and specify their value space.
2. Define three different numeric random variables on this space.

1. For example
    a. {Grammatical, ungrammatical}
    b. {More than 14 words, 14 words or less}
    c. {Contains a parasitic gap, Does not contain a parasitic gap}
2. For example
    a. Sentence length. {1, 2, …, n,…}
    b. Verb ratio: number of verbs/number of words [0,1]
    c. Average word length in the sentence $[1,\infty)$

Exercise 7

We are throwing 5 fair dices

1. What is the chance of getting 5 6s?
2. What is the chance of getting yatzy (five equal values, any value)?
3. What is the chance of getting at least 4 equals?
4. What is the chance of getting 4 – but not 5 – equals?
5. What is the chance of getting a house: 3 equals + a pair (with a different value)?

1. $\left(\frac{1}{6}\right)^5$

2. $6 \times \left(\frac{1}{6}\right)^5 = \left(\frac{1}{6}\right)^4$

Easier to do (4) before (3).

The chance of getting 66665 is $\left(\frac{1}{6}\right)^5$

The chance of getting 6666x, where x=/=6 is then $\left(\frac{1}{6}\right)^5 \times 5 = \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right)$

Since the dice that is x =/=6 can take any of five different positions (6666x, 666x6, 66x66, 6x666, x6666),

the chance of getting exactly four 6s is $5 \times \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right)$

Observe that this is the value of the binomial distribution.

To get not only 4 equal 6s, but any value, we multiply by 6 and get $6 \times 5 \times \left(\frac{1}{6}\right)^4 \times \left(\frac{5}{6}\right) \;=\; 5 \times 5 \times \left(\frac{1}{6}\right)^4$

For point (3) we add the results from (2) and (4).

Point 5:

The chance of getting 66655 is $\left(\frac{1}{6}\right)^5$

There are 10 different ways to order these numbers (try, or use the binomial coefficient $\binom{5}{2}$)

The three equals can take any value as long as they take the same, 6 possibilities.

The pair has to be different from the three equals, i.e. 5 possibilities.

Putting this together yields $6 \times 5 \times \binom{5}{2}\left(\frac{1}{6}\right)^5 = 50 \times \left(\frac{1}{6}\right)^4$

Exercise 8

We are considering the sum of the five fair dices.

1. What is the expectation of the sum?
2. What are the variance and the standard deviation of the sum?

Expectation of one dice is 3.5

Expectation of 5 dices is 5*3.5=17.5

Variance of one dice is 35/12

Variance of 5 dices are 5*35/12 = 14.583 because of independence

Standard deviation is 3.82