UPPSALA UNIVERSITY
Fredrik Wahlberg

5LN445 Mathematics for Language Technologists
23 September 2020

# Dugga 1: Sets, functions, and probability

Pass: at least 60% correct (i.e. 24p). Pass with distinction: at least 80% correct on the two tests added. When in doubt about the interpretation of a task, make reasonable assumptions and motivate those. If you get stuck on a task, try to solve other tasks first, then go back. Please read the whole exam before beginning.

**General rules:** Mobile phones must be switched off.

**Tools:** Pen, handwritten notes and a calculator. You will be given a probit table as a part of the exam.

**Formulas:**

Bayes' theorem: $P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1)+P(A|B_2)P(B_2)+...+P(A|B_n)P(B_n)}$

Theorem of total probability: $P(A|B) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_n)P(B_n)$

1. In a board game (of the adventure and power fantasy variety) you play the Hero and will have to fight orcs to get through a dungeon (probably below the pyramid of some zombie spider demigod). The encounters are fought by you throwing a 10 sided die, and another 10 sided die is thrown for the antagonist. The results are compared and the antagonist needs to score three points higher than the Hero in order to win the encounter (a-b ≥ -2). Additionally, a die score of 10 is a "botched" throw and counts as -10. (10p)
   *(Examples: The Hero wins on (3, 3), (1, 10), (10, 10). The antagonist wins on (10, 9), (1, 4), (4, 8) )*
   *Give probabilities as percentages and answer with relevant calculations/reasoning. If your calculator can't handle powers properly, write the simplest possible expression instead of real numbers.*
   a. Define a suitable sample space, Ω, for one fight in this game.
   b. Define one (or more) suitable events for the problem.
   c. Show (using a table or event tree) the number of outcomes in Ω that favour the Hero.
   d. Calculate the probability of the Hero or the antagonist winning an encounter, respectively.

   Our hero claims a warrior's heart and decides to fight more orcs.

   e. What is the probability of the Hero losing at least one of three encounters?
   f. Winning exactly 5 encounters out of 10.

   Let's assume the Hero is attacked by a horde of 100 orcs armed with wooden butter knives (probably due to cutbacks). This reduces an orc's chance of winning to 1%. *(Note that the Hero always miraculously survives, never gets tired, and fights orcs one-by-one.)*

   g. What is the probability of the Hero winning against each of the 100 orcs in the onslaught?

a. (1p): $D = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $\Omega = DxD = \{(1, 1), (1, 2), (1,3), \ldots, (10,9), (10,10)\}$
b. (1p): $H = \{(a, b) \in \Omega \mid a\text{-}b \geq -2\}$ (the hero wins)
c. (3p): Event H marked in bold, 'a' is the Heros die in rows and 'b' is the antagonists die in columns.

| **a - b** | **b=1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| **a=1** | **0** | **-1** | **-2** | -3 | -4 | -5 | -6 | -7 | -8 | **11** |
| **2** | **1** | **0** | **-1** | **-2** | -3 | -4 | -5 | -6 | -7 | **12** |
| **3** | **2** | **1** | **0** | **-1** | **-2** | -3 | -4 | -5 | -6 | **13** |
| **4** | **3** | **2** | **1** | **0** | **-1** | **-2** | -3 | -4 | -5 | **14** |
| **5** | **4** | **3** | **2** | **1** | **0** | **-1** | **-2** | -3 | -4 | **15** |
| **6** | **5** | **4** | **3** | **2** | **1** | **0** | **-1** | **-2** | -3 | **16** |
| **7** | **6** | **5** | **4** | **3** | **2** | **1** | **0** | **-1** | **-2** | **17** |
| **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** | **0** | **-1** | **18** |
| **9** | **8** | **7** | **6** | **5** | **4** | **3** | **2** | **1** | **0** | **19** |
| **10** | -11 | -12 | -13 | -14 | -15 | -16 | -17 | -18 | -19 | **0** |

d. (1p): $P(H) = |H|/|\Omega| = 70/100 = 70\%$
$P(H)+P(H^C)=1$ (from axioms) $\Rightarrow P(H^C) = 1\text{-}P(H) = 30\%$
e. (1p): Complement to winning all: $1\text{-}0.7^3 = 65.7\%$
f. (2p): Use binomial PMF: $\binom{n}{k}p^k(1-p)^{n-k} \Rightarrow \binom{10}{5} \cdot 0.7^5 \cdot 0.3^5 \approx 10.3\%$
g. (1p): This can be thought of as a tree diagram with two outcomes per level. Since each turn is independent the probability of 100 wins is simply a multiplication with the same probability at every level as: $P(H)^{100} = 0.99^{100} \approx 36.6\%$

2. For each of the following statements, *argue* that it is true for all A and B, false for all A and B, or true for some A and B (exemplify which) for *any finite sets* A and B. (8p)

a. $A \subseteq (A \cup B)$
b. $(A - B) \subseteq A$
c. $(B - A) = (A - B)$
d. $(A \cap B) \subseteq A$
e. $A - B = \varnothing$
f. $|A \cup B| + |A \cap B| = |A| + |B|$
g. $|A - B| = |B - A|$
h. $|A \cap B| \leq |A|$

a. (1p): Necessarily true for all A and B. The union operation never removes anything. Hence, each set going into the union is a subset of the union.
b. (1p): Necessarily true for all A and B. (A-B) must include A or less by definition.
c. (1p): Necessarily false if $A \neq B$
d. (1p): Necessarily true, since the intersection of A and B is the content of A or something less.
e. (1p): True if (and only if) $A \subseteq B$
f. (1p): Necessarily true for all A and B. This is the union cardinality counting rule or the general addition rule. Can be shown with a Venn diagram or derived.
g. (1p): Can be true for disjunct A and B where $|A|=|B|$, or empty sets.
h. (1p): Necessarily true, since the intersection of A and B is the content of A or something less.

3. Early on in the Covid-19 pandemic, the testing capacity was an important limitation in contact tracing for reducing the spread of the new virus. Early testing depended on polymerase chain reaction (PCR) where the virus' RNA could be detected. However, widely varying false negative rates were reported. In a meta-study[1] of different PCR testing approaches, sensitivity (ratio of positives correctly identified) for a saliva test were 91% and its specificity (ratio of negatives correctly identified) approximately 98%. Based on fatality data, infection fatality ratios (IFR), and SIR modelling, the number of people infected in Sweden in the beginning of april can be estimated to be between 130k (IFR 0.34%, Gangelt) and 65k (IFR 0.7%, Diamond Princess). Sweden's total population was approximately 10.2 million. While testing a random person in a screening test, investigate the probability of that person having the disease when given a positive test. (12p) *(Give probabilities as percentages and answer with relevant calculations/reasoning.)*
   a. Define relevant variables, a sample space and events.
   b. Find the probabilities given by the text in terms of your chosen probability space from (a).
   c. Draw an event tree
   d. What is the probability of having the infection, given a positive PCR result (i.e. P(being sick | positive test) for both estimates of the number of infected in the population?

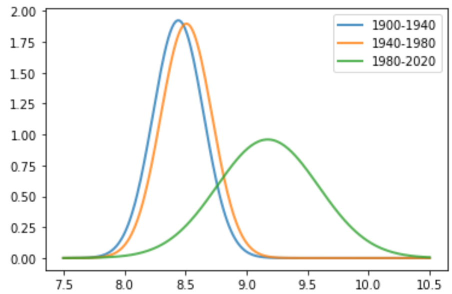
   The referenced meta-study also states that nasal swabs (as opposed to saliva tests) have a sensitivity of 98%:

   e. Redo the calculations from (d) using this new sensitivity?

a. (1p): A: Having covid, B: Positive test
b. (1p): $P(A) = 130000/10200000 \approx 1.27\%$ or $P(A) = 65000/10200000 \approx 0.63\%$
   $P(B|A) = 91\%$
   $P(B^C|A^C) = 98\%$
c. (2p): Draw a tree with 2x2 splits
   Since $P(A)+P(A^C)=1$: $P(A^C) = 98.73\%$ or $P(A^C) \approx 99.37\%$
   Since $P(B|A)+P(B^C|A)=1$: $P(B^C|A) = 9\%$
   Since $P(B|A^C)+P(B^C|A^C)=1$: $P(B|A^C) = 2\%$
d. (4p): By Bayes' theorem (or from the tree):
   $$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} = \frac{0.91 \cdot 0.0127}{0.91 \cdot 0.0127+0.02 \cdot 0.9873} \approx 0.3692$$
   $$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} = \frac{0.91 \cdot 0.0063}{0.91 \cdot 0.0063+0.02 \cdot 0.9937} \approx 0.2239$$
   $P_1(A|B) \approx 37\%$ and $P_2(A|B) \approx 22\%$
e. (4p): $P(B|A) = 98\%$
   $$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} = \frac{0.98 \cdot 0.0127}{0.98 \cdot 0.0127+0.02 \cdot 0.9873} \approx 0.3866$$
   $$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^C)P(A^C)} = \frac{0.98 \cdot 0.0063}{0.98 \cdot 0.0063+0.02 \cdot 0.9937} \approx 0.2370$$
   $P_1(A|B) \approx 39\%$ and $P_2(A|B) \approx 23\%$

[1] Laszlo Mark Czumbel, Szabolcs Kiss, Nelli Farkas, Ivan Mandel, Anita Emoke Hegyi, Akos Karoly Nagy, Zsolt Lohinai, Zsolt Szakacs, Peter Hegyi, Martin C. Steward, Gabor Varga, *"Saliva as a Candidate for COVID-19 Diagnostic Testing: A Meta-Analysis",* medRxiv 2020.05.26.20112565; doi: https://doi.org/10.1101/2020.05.26.20112565

4. The increasing temperature of the earth is clearly measurable, though anthropogenic climate change is a debate in many places. In the figure, temperature data[2] is shown for land mass average temperature over the summer months (of the northern hemisphere) at different time spans. The distribution of recorded temperatures form normal distributions. It should be noted that these distributions describe the observations and can not by themselves settle if climate change is anthropogenic. (10p)




*(Give probabilities as percentages and answer with relevant calculations/reasoning.)*

Period 1 temperatures (1900-1940): $X_1 \sim \mathcal{N}(\mu=8.44, \sigma=0.21)$

Period 2 temperatures (1940-1980): $X_2 \sim \mathcal{N}(\mu=8.51, \sigma=0.21)$

Period 3 temperatures (1980-2020): $X_3 \sim \mathcal{N}(\mu=9.17, \sigma=0.42)$

a. What are the mean summer temperatures for the respective time spans?

For a random day in the summer months and for each period, find:

b. The probability of observing a temperature of at least 9 degrees for each period.
c. Any temperature outside of the interval $\mu\pm3\sigma$ can be considered extreme. What span of temperature is then non-extreme for each period?
d. What is the probability of observing extreme temperatures for each period?
e. What is the probability of in period 3 observing temperatures that would have been considered extreme in period 2?

a. (1p): The means of the distributions are 8.44, 8.51. and 9.17, by definition.
b. (3p): $P(X_1>9) = 1-\phi((9-8.44)/0.21) \approx 1-\phi(2.67) \approx 0.38\%$
   $P(X_2>9) = 1-\phi((9-8.51)/0.21) \approx 1-\phi(2.33) \approx 0.98\%$
   $P(X_3>9) = 1-\phi((9-9.17)/0.42) \approx 1-\phi(-0.40) \approx 65.7\%$
c. (2p): $8.44\pm3\cdot0.21 \Rightarrow [7.81, 9.07]$
   $8.51\pm3\cdot0.21 \Rightarrow [7.88, 9.14]$
   $9.17\pm3\cdot0.42 \Rightarrow [7.91, 10.43]$
d. (2p): $1-P(\mu-3\sigma < X < \mu+3\sigma)$ is the same for all normal distributions.
   $1-P(-3 < Z < 3) \approx 0.27\%$
e. (2p): $P(X_3 > 8.51+3\cdot0.21) = 1-\phi((9.14-9.17)/0.42) \approx 1-\phi(-0.071) \approx 52.9\%$
   The probability of $P(X_3 < 8.51-3\cdot0.21)$ is negligible.

[2] Berkeley Earth project, http://berkeleyearth.org/data-new/ , raw_TAVG