Math 10 – Homework 3 **ANSWER KEY**

Note to grader: earlier versions of the homework had a different point assignment.

Please follow the point assignment in this answer key.

Instructions

- Type your answers and paste images directly into this document.
- Answers are usually short, around 1-3 sentences.
- You will need to use a calculator for this homework.
- Print out and hand in homework in class on Tuesday.
- You may collaborate on the homework but you must write it up yourselves.

The Mathematics of DNA Matching – Part 1

- Quoted from http://dna-view.com/profile.htm:

"A typical DNA case involves the comparison of two samples – an unknown or *evidence* sample, such as semen from a rape, and a known or *reference* sample, such as a blood sample from a suspect.

If the DNA profile obtained from the two samples are indistinguishable (they 'match'), that of course is evidence for the court that the samples have a common source – in this case, that the suspect contributed the semen."

- Quoted from Wikipedia:

"Although 99.9% of human DNA sequences are the same in every person, enough of the DNA is different that it is possible to distinguish one individual from another, unless they are monozygotic ("identical") twins. DNA profiling uses repetitive ("repeat") sequences that are highly variable, called variable number tandem repeats (VNTRs), in particular, short tandem repeats (STRs). VNTR loci are very similar between closely related humans, but are so variable that unrelated individuals are extremely unlikely to have the same VNTRs."

When Wikipedia says the sequences are highly variable, it means that for well mixed heterogeneous populations, the occurrence of these genes in a person's DNA are *independent* events.

Please show your work for the following questions. Points will be taken away otherwise. Grade: please give points for the final answers, as well as the steps.

Question 1 - How would you try and check that these events are indeed independent in

such a population? *(3 pts)*

Hint: we did not learn a formal way to test for independence. For this question, start with the definition of independence in probability, then think about how you could at least try to see if this definition approximately holds.

Independence: P(A and B) = P(A)P(B). We can check by tallying the proportions P(A) and P(B) for two genes A and B. Then tally the proportion P(A and B) that a person has both A and B. Then see if P(A and B) = P(A)P(B). Repeat for all pairs.

Alternatively: P(A and B and C and D) = P(A)P(B)P(C)P(D). Do the same for all 4 genes at the same time.

An example of four genes from http://dna-view.com/profile.htm:

DNA profile locus	**Genotype frequency**
CSF1PO	.16
TPOX	.28
THO1	.07
vWA	.05

"DNA profile locus" are the events mentioned previously, and "Genotype frequency" is the probability that the event occurs for a person. E.g. if you pick a person at random from the population, there is a 0.28 probability that she has the gene TPOX.

Question 2 – Under the assumption of independence, what is the probability of a randomly selected person having all four of these genes, to 4 significant figures? *(3 pts)*

P(A and B and C and D) = P(A)P(B)P(C)P(D) = 0.16*0.28*0.07*0.05 = 0.0001568 by independence.

Question 3 – Under the assumption of independence, what is the probability of a randomly selected person NOT having this "gene profile"? *(2 pts)*

Note: having this "gene profile" means to have all four of these genes.

P(not having this profile) = 1 – P(have this profile) = 1 – 0.0001568 = 0.9998432.

Question 4 – Suppose that the DNA profile of the accused person is compared with the inmates of the 65,000 felons in the Arizona prison database. What is the probability that NONE of them match/has this profile, to 4 significant figures? *(3 pts)*

P(none match) = (0.9998432)^65000 = 0.00003743.

Question 5 – What is the probability that at least one person in the database matches the profile? *(2 pts)*

P(someone matches) = 1 – P(none match) = 1 - 0.00003743894 = 0.99996256106.

The Mathematics of DNA Matching – Part 2

The first criminal caught using DNA fingerprinting (England), using the DNA profiling method published in 1985 by Sir Alec Jerreys.

Two teenage girls were murdered in the small town of Narborough, Leicestershire, in 1983 and 1986. These events sparked a murder hunt that was only to be resolved by an innovative DNA mass intelligence screen. This eventually led to the conviction of a local man, but not before the prime suspect was proved to be innocent.

In 1983, 15-year-old schoolgirl Lynda Mann was found raped and murdered in the Narborough area. Forensic scientists visited the scene and a semen sample taken from her body was found to belong to a person with type A blood and an enzyme profile that matched only 10 per cent of the adult male population. Unfortunately, with no other leads or forensic evidence, the murder hunt was eventually wound down.

Three years later, Dawn Ashworth, also 15, was found strangled and sexually assaulted in the same area. Police were convinced that the same assailant had committed both murders, and the FSS recovered semen samples from Dawn's body that revealed her attacker had the same blood type as Lynda's murderer.

The prime suspect was a local boy, Richard Buckland, who revealed after questioning, previously unreleased details about Dawn Ashworth's body. Further questioning led to his confession, but he denied any involvement in the first murder of Lynda Mann. Convinced that he had committed both crimes, Leicestershire police contacted Dr Alec Jeffreys at Leicester University, who had developed a technique for creating DNA profiles.

Using this technique, Dr Jeffreys compared semen samples from both murders, against a blood sample from Richard Buckland, which conclusively proved that both girls were killed by the same man, but not by him.

The murderer was believed to be from one of three villages, with male population totaling 5000. Let's assume Jeffreys had a different gene profile, rather than the four described in part 1. We will use $0.000003 = 3 \times 10^{-6}$, or 3 over 1 million, as a proxy for the probability of having the profile found by Jeffreys (i.e. a randomly selected person having all the genes in the profile). For questions 1-3, do the calculations assuming that the murderer wasn't among the 5000.

Question 1 - What is the probability that no male in these villages would have this profile, to 4 significance figures? *(3 pts)*

P(no male matches) = (1-3/1000000)^5000 = 0.9851.

Question 2 - What is the probability that at least one randomly selected male in these villages has this profile, to 3 significance figures? *(2 pts)*

P(someone matches) = 1 – P(no matches) = 1 - 0.9851= 0.0149.

Question 3 - What is the probability that 2 randomly selected males have this profile? *(3 pts)*

P(male 1 matches AND male 2 matches) = P(male 1 match) P(male 2 match) by independence = (3/1000000)^2 = 9 x 10^(-12). It is fine if they say it is very close to or essentially 0.

Question 4 – Do you agree with the two assertions in Jeffrey's conclusion:

A) That because the semen samples from both murders had the same gene profile, both girls must be killed by the same man? *(3 pts)*

B) That because Richard Buckland's gene profile did not match those from the semen samples, he must not be the murderer? *(1 pt)*

Explain your answers. We are assuming that Jeffrey's tests/comparisons are perfectly accurate.

A) Yes, both girls were probably killed by the same man since two randomly selected males having the same profile has basically 0 probability from question 3.

***Note: students might (correctly) deduce that in the small village, people are more genetically related, and hence the occurrence of genes might not be independent. Using this point, they could cast doubts on assertion A), which should be given full credit.

B) Yes, if Richard Buckland was the killer, his DNA would have matched.

*** Note: there was a previous, more open-ended version of this question 4. Please be very generous with the grading if the student's argument makes statistical sense.