MATH 4330 — Assignment 1 Solutions

Instructor: Kevin McGregor

Question 1: A researcher wants to study whether an existing drug causes a particular kind of skin reaction. She collects data from an existing medical database; she samples individuals who showed the skin reaction as well as individuals without a reaction. She then examines how many individuals had taken the drug of interest.

(a) [**2 points**] What kind of study is this? Be as specific as possible, and explain your answer.

Answer: This is a **retrospective** study, because the researcher is using existing data and looking back into the past.

(b) [**2 points**] Suppose the researcher finds that those who took the drug had a higher chance of developing a skin reaction. Do you believe this result? Can you think of other specific factors that might affect the conclusion? Explain your answers.

Answer: Since the study is observational, we cannot say for certain whether the drug caused the skin reaction. In an observational study there is a possibility of confounders. In this example, the prior health of the patient could be related to both the chance of being prescribed the drug AND the chance of developing a skin reaction.

Question 2: [**2 points**] A researcher runs a randomized experiment where study participants are randomized to be given either a drug or a placebo. Another researcher wants to perform an additional study on this same study group. He asks participants whether they get more or less than 3 hours of exercise per week. He concludes that individuals with more than 3 hours of exercise per week have lower blood pressure than those who get less than 3 hours, on average. Are you worried about confounders in this conclusion? Explain.
Answer: There is a possibility of confounders. While the original study was randomized, it was the drug itself that was randomized, not exercise. There could still be a confounder such as age or prior health in the exercise/blood pressure relationship. In fact, if exercise and blood pressure were measured after the drug was administered, then the drug itself could be a confounder, because it could affect exercise habits as well as blood pressure.

Question 3: We want to study the genetics of colour distribution in a population of unicorns. Suppose that in unicorns there is a single gene that determines colour; there are two variants of the gene, one called A and the other called a. Each unicorn has two copies of the gene (one from each parent). The combination of the gene variants for copy 1 and copy 2 of the gene determines colour as follows:

Copy 1	Copy 2	Colour
A	A	Red
A	a	Pink
a	A	Pink
a	a	White

We want to determine if “random mating” is happening in this unicorn population. This would mean that the probability that a newly born unicorn inherits variant A or a with probability equal to the prevalence of each variant in the overall population.

(a) [**3 points**] Let p be the proportion of the A variant in the population, and q be the proportion of the variant a, so that $q = 1-p$. Under random mating, each newborn unicorn inherits A with probability p and a with probability q and the two copies inherited in each unicorn are independent of one another. Calculate the expected proportions of unicorn colours under random mating.

Answer: Since the two copies of the gene are inherited independently, we can calculate colour probabilities as follows:

$$P(\text{Red}) = P(AA) = P(A) \cdot P(A) = p^2$$
$$P(\text{Pink}) = P(Aa) + P(aA) = 2P(A)P(a) = 2pq$$
$$P(\text{White}) = P(aa) = P(a) \cdot P(a) = q^2$$

(b) [**5 points**] Suppose you know that $p = 0.75$ and $q = 0.25$, and that you collect the following data from a sample of unicorns:

Colour	Number of unicorns
Red	45
Pink	49
White	12

Use the appropriate statistical test to determine whether the random mating assumption holds in this population. **Write out your calculations for this part; using R for this part will not result in credit**.

Answer: Calculating expected proportions using part (a) we have:

$$P(\text{Red}) = 0.75^2 = 0.5625$$
$$P(\text{Pink}) = 2(0.75)(0.25) = 0.375$$
$$P(\text{White}) = 0.25^2 = 0.0625$$

The total number of counts is $n = 45 + 49 + 12 = 106$. Therefore, the expected counts of the three colours is:

Colour	Expected count
Red	106(0.5625)=59.625
Pink	106(0.375)=39.75
White	106(0.0625)=6.625

Since we're comparing the counts of a single categorical variable to hypothesized proportions, we're doing a chi-squared test for goodness of fit. The test statistic is:

$$\begin{aligned}X^2 &= \frac{(45-59.625)^2}{59.625} + \frac{(49-39.75)^2}{39.75} + \frac{(12-6.625)^2}{6.625} \\ &= 10.101\end{aligned}$$

There are three categories, so we have $3-1=2$ degrees of freedom. The chi-squared upper tail for level $\alpha = 0.05$ is 0.0064. Therefore, we reject the null hypothesis that the observed counts follow the hypothesized distribution. Therefore, we do **not** have evidence of random mating.