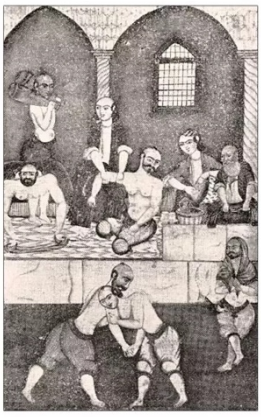
Probability 2 – term bonus project

Fall 2022

Submission details:

- **Due: end of Thursday, February 2nd, 2023**
- **Send me your report via email at ab.safari.w@gmail.com**
- **Presentation day: Saturday, February 5th, 2023**



Wrestling is an ancient sport with origins from 20 centuries ago (from cave drawings!). Each match consists of two 3-minute periods (freestyle) and the winner is declared by the total points at the end of both periods (or a technical superiority is reached during the match). You can find more details about the rules of this sport on United World Wrestling (UWW) website. Predicting a wrestler score prior to a match is a very challenging problem and depends on many factors. Here we aim to tackle a simpler problem: obtaining the statistical distribution (i.e., modelling!) of the average score of a professional wrestler in an international tournament given some assumptions (to further simplify the problem).

Here are the scores of Hassan-Aliazam Yazdani-Charati (aka Hassan Yazdani with UWW rank of 4) in all of his international matches prior to 2022 tournaments since 2015:

3, 9, 6, 6, 16, 6, 8, 10, 9, 12, 10, 9, 13, 6, 10, 7, 10, 4, 11, 11, 10, 4, 10, 10, 12, 10, 13, 12, 10, 12, 11, 10, 11, 10, 10, 6, 16, 11, 11, 10, 10, 11, 5, 11, 5, 9, 10, 10, 10, 10, 3, 7, 12, 11, 6, 8, 10, 12

These scores build our "belief" about Hassan's score during a match in an international tournament.

Step 1:

Let's try to model these data under a simplified scenario:

a) Let $X_1, \dots, X_{58}$ be Hassan's scores in all of his international matches prior to 2022. The above data are the observations for these variables.

b) Further assume these X_i's are IID following a normal distribution:

$$X_1, \dots, X_{58} \overset{iid}{\sim} N(\mu, \sigma^2)$$

c) By using the above data, estimate the parameters (μ and σ^2) of the normal distribution given in part b.
d) Generate 1000 random numbers from this normal distribution with the estimated parameters in part c.
e) Compare the distribution of the generated sample with the distribution of the original data (by using, e.g., a histogram plot – can you think of a better visualization tool?). Comment on whether the assumption of normality for the Hassan's scores was a good choice.

Step 2:

It is already end of 2022 and Hassan's scores for the 2022 tournaments are available as well:

11, 12, 12, 1, 10, 11, 10

Let's aim to "*update*" our "*prior belief*" about Hassan's score with his scores in 2022:

a. Assume the following *hierarchical model* for the available data:
$\underline{Y} = (Y_1, \dots, Y_7)$: 2022 observations

$New\ observation\ given\ prior\ information$:

$$Y_1, \dots, Y_7 | \theta, \lambda \overset{iid}{\sim} N(\theta, \lambda)$$

$Prior\ information$:

$$\theta \sim N(\mu, \sigma^2)$$
$$\lambda = 8.4$$

Recall the main question we are trying to answer here: "*model average score of Hassan in an international tournament*". In the above model, θ is the average score of Hassan in 2022. As you might already noticed, unlike what we have seen so far, θ (the mean parameter of a normal distribution) is a random variable itself following another normal distribution! This is a simple scenario in a *Bayesian framework*. The main idea is:

i. we are interested in a random quantity (here: average score of Hassan in a tournament, θ)
ii. we have some prior information about that quantity (here: $\theta \sim N(\mu, \sigma^2)$)
iii. we have new information/observations for the random quantity (here: $Y_1, \dots, Y_7 | \theta, \lambda \sim N(\theta, \lambda)$)
iv. finally, we want to update our prior belief given the new available information.

Think why it is called *Bayesian*!

b. Find the distribution of θ given the new observations, that is, $\theta | \underline{Y}$
c. By using the estimates you obtained in part c of Step 1, estimates the parameters of the conditional distribution of $\theta | \underline{Y}$
d. Repeat part d of Step 1 for the conditional distribution of $\theta | \underline{Y}$

e. Compare distribution of:
   i. generated random numbers in part d of Step 1
   ii. generated random numbers in part d here
   iii. all the available scores of Hassan (since 2015 until end of 2022)
f. Can you see any improvement on the estimated distribution obtained here than that of in Step 1?

Step 3:

Hassan Yazdani, like every other champion, is most of the time well-prepared for a match and sometimes not as prepared. His level of strength may depend on many (mostly latent) factors. Based on the field experts, Hassan's fitness in match i can be expressed as a binary random variable:

$$Z_i = \begin{cases} 1, well - prepared \\ 0, otherwise \end{cases}$$

Clearly, as you may expect, his score in a match heavily depends on his fitness status (e.g., we expect him to score better when he is well-prepared comparing to when he is not). This is a *mixture* of two normal distributions. Let's reconstruct our previous models by incorporating this new piece of information:

a. Given the binary nature of new random variable, Z, it is reasonable to assume a Bernoulli distribution:

$$Z_i \overset{iid}{\sim} B(1, p)$$

b. The hierarchical model presented in part a of Step 2 can be rewritten as follows:

$$Y_i|Z_i, \underline{\theta}, \lambda \sim N(\theta_{Z_i}, \lambda)$$
$$\theta_{Z_i}|Z_i, \sim N(\mu_{Z_i}, \sigma^2)$$
$$Z_i \sim B(1, p)$$
$$\lambda = 8.4$$

c. Repeat part b of Step 2, but assuming the Hassan's fitness, Z_i, is known (i.e., add Z_i's to the condition). Remember, since we have two means now (μ_1, μ_2), you need to find the joint distribution of these two means (i.e., distribution of $(\theta_0, \theta_1)|\underline{Y}, \underline{Z}$). You can assume that the two means are independent given the observed data.
d. In part c, we assumed Hassan's fitness is known. However, that was not a realistic assumption as this information is mainly latent to almost everyone! Now, repeat part c, but this time drop $\underline{Z}$ from the condition (i.e., find the distribution of $(\theta_0, \theta_1)|\underline{Y}$). How much more complex is this new conditional distribution compared to part c?
e. Given all the available data from Hassan's scores (from 2015 until end of 2022), what is the distribution of his score in a tournament in 2023? In other words, obtain an expression for $f_{Y_{2023}|\underline{Y}(y)}$.

Step 4:

From Step 3, you probably realized that the updated distribution is quite complex, and therefore, it is not easy to generate random numbers from it (so, we can have a better understanding about how the distribution looks like). Here, we propose an algorithm that can generates random numbers from such complex distribution called *Gibbs Sampling*.

a. Since Hassan's fitness is not being observed, we cannot really have a very precise information for the distribution of Z_i. But we can always approximate it! Here is one way of doing it. Looking at the results of Hassan's matches, we see Hassan won almost 90% of his matches during these years. That is, we can assume $p \approx 0.9$.
b. Further, we need to estimate μ_1 and μ_0 (average prior belief of Hassan's scores when he is well-prepared and when he is not, respectively). As we do not know when he was well-prepared and when he was not, we cannot precisely estimate these two averages. Again, we need to make some assumptions and use some approximations. We expect to see higher scores (on average) from matches where Hassan was well-prepared than those in which Hassan was not as ready. Given our approximation for p in part a, let's estimate μ_1 by the top 90% of the scores in our prior belief data and use the lower 10% of data to estimate μ_0. As we assume equal variation under two conditions (hence using σ^2 for both $Z_i = 0$ and $Z_i = 1$ in the hierarchical model), we can use all of the belief data to estimate σ^2.
c. Obtain the following two conditional distributions:

$$\underline{Z}|\underline{Y}, \underline{\theta}, \lambda$$
$$\underline{\theta}|\underline{Y}, \underline{Z}, \lambda$$

d. Initialize $\underline{\theta} = (\theta_0, \theta_1)$ by some random values (e.g., by estimated μ_0 and μ_1 from part b).
e. Repeat the following two items 1000 times and for every item use the most updated generated numbers you have for $\underline{Z}$ and $\underline{\theta}$ (you need to use the estimated values for the parameters from parts a and b):
    i. Generate $\underline{Z}$ from the conditional distribution
    ii. Generate $\underline{\theta}$ from the conditional distribution
f. Repeat parts e and f from Step 2.