Interview Questions and Summaries

D. Jinrui K. Mikdad L. Xiang S. Jia

On December 19, 2023, we interviewed Dr. Marcos Malosetti. He has a PhD in Statistics and Plant Breeding. He is the lead scientist in statistics at Nunhems, a seasoned researcher.

Based on his experiences, we organized our interview questions into four parts. They are working experiences, data processing, modeling, and academia versus industry. The questions and summaries of answers to each part are as follows. Note that FQ means follow-up questions. In the parenthesis are the names of the person who proposed the question.

Working experiences

Q 1: What does a typical day look like for a researcher? How do you divide tasks? (Jia)

A: A researcher's role centers on effective communication with team members who develop data analytic products and their users. Key responsibilities involve explaining how to use data analytics, ensuring efficient analysis deployment, and scouting for improvements and innovations. Internal decision-making processes and team management, which include motivating team members and ensuring their professional growth, are also crucial. Overall, researchers' role requires a balance between maintaining current operations and actively seeking opportunities for enhancement and innovation.

Q 2: Could you share an instance where a statistical finding positively impacted the plant breeding business? (Jia)

A: There are lots of examples. For instance, to test numerous potential new plant hybrids, we could use statistics to find efficient experimental layouts. Given the limited time and resources, insights from data help allocate treatments in layout and reduce the number of replications, thus reducing the whole experiment.

Data processing

Q 3: How much data do you work with? (Jinrui)

A: Large. We have fifty or more ongoing projects with hundreds of experiments in different locations. We have twenty-four crops. They could be fruits, roots, or leafy vegetables, all having distinct traits. To deal with the highly diverse data, we create data pipelines that are robust to all the diversity so we don't have to take care of every data set separately.

FQ 3.1: We have established that efficiency is crucial. How does a data analysis pipeline make data analysis more efficient? (Jinrui)

A: Data analysis is a decision-making process in which we can make many options that are usually not right or wrong. What we should do is to make them transparent. That is the essence of these pipelines. The steps of the pipeline must be coherent, efficient, and robust, a thing that works. Knowing that data will not be perfect, we need to anticipate the problems the data might have. We figure out how to diagnose them and find the decisions that help fix them on the fly. Once we have 80% of the analysis in that way, we can live with 20% that we have to dig in a little bit more detail and troubleshoot in more detail.

Modeling

Q 4: Which part is the most complex or time-consuming in modeling? (Xiang)

A: I would bring in three challenges. The first one is random effects and modeling correlation. Data from genetics automatically brings random effects, and cooperating information from parent tests requires modeling correlations. The mixed model can solve this challenge. In addition, different types of response variables are in the models: continuous, dummy, categorical, and counting. We must build models for each of the response. Finally, we have to be fast and make the solution as robust as possible, which makes it difficult to trade off simple and elaborate models.

Q 5: What is the primary purpose of models in your work? (Xiang)

A: The primary purpose is prediction and prediction with small data. Predicting which material or plant hybrid can be a commercially potential product helps us narrow the number of experiments to a small amount and save time.

FQ 5.1: What do you mean by small data? (Jinrui)

A: The number of experiments you can run is small because time is limited.

Academia vs industry

Q 6: Were there any noticeable differences between academia and industry? (Mikdad)

A: I already had experience on the applied side of statistics in academia, so there wasn't much change. On the other hand, there was indeed a slight difference concerning priorities and developments.

FQ 6.1: I imagine you explain things to other researchers in academia. Was it more challenging to communicate with others in the industry? (Mikdad)

A: That is true! One of the challenges in the industry is to communicate with others. In academia, you should explain with more detail, which is different from explaining things to end-users. Yet, the communication is still there!

FQ 6.2: Is visualization techniques essential for explaining ideas to other people? (Mikdad)

A: Yes, we statisticians do not necessarily have all the skills to do development, which is why we need to work with other people who have the right skills to build efficient, attractive, and user-friendly dashboards.

Q 7: What would be your advice to students of master statistics and data science? (Mikdad)

A: The most important thing is to like the area and enjoy what you are doing. Moreover, you should understand the context.