# CASE STUDY 1: Random Effects Multinomial Model for predicting ATM withdrawal amounts

Suppose we have data on a random sample of 5000 bank accounts making ATM withdrawals. We have data on the amount and time of each withdrawal from each bank account. Note that multiple withdrawals can be made from the same account over time. Our goal is to build a statistical model to estimate

* The most likely withdrawal amount in the population
* Future withdrawal behavior of a randomly sampled customer.

Q: Why may such a predictive analysis be of practical importance?

Q: What are the benefits of including random effects in the statistical analysis?

The data has the following format

|  |  |  |  |
| --- | --- | --- | --- |
| **Account Number** | **Withdrawal Amount** | **Date** | **Time** |
| 2098765 | $200 | 1/07/2017 | 9:00 |
| 2098765 | $50 | 15/07/2017 | 12:15 |
| 2098765 | $120 | 28/07/2017 | 14:38 |
| 3055488 | $500 | 13/07/2017 | 10:23 |
| 3055488 | $500 | 13/07/2017 | 11:31 |
| 3055488 | $500 | 13/07/2017 | 12:45 |
| 3055488 | $500 | 13/07/2017 | 13:58 |
| 3055488 | $500 | 14/07/2014 | 15:03 |
| 3055488 | $400 | 2/08/2017 | 18:14 |
| 3055488 | $50 | 2/08/2017 | 18:20 |
|  |  |  |  |

Q: Let’s first fit a simple model, without any random effects or adjustments for serial dependence. Describe the statistical model you would fit to achieve this goal. Discuss the limitations of this model.

Q: Modify your above model to include random effects.

# CASE STUDY 2: A Dirichlet Multinomial Mixture Model-based approach for short text clustering

Q: What is short text?

Q: What is short text clustering and why do you think short text clustering has become an increasingly important task?

Q: What are some of the challenges with short text clustering (versus standard document classification)?

Suppose we have a data set of `D’ queries/short comments. Call these documents. We want to classify the documents into one of `K’ classes. Let `z’ denote the cluster label. So we want to compute p(z=k|d) = [p(z=k)p(d|z=k)]/p(d) (Bayes’ rule)

Q: What data do we have to compute p(d| z=k)? Write out a mathematical expression for p(d|z=k).

Q: We also need to specify the density p(z=k). What could we assume?

Q: What are the key assumptions of the above model?