Contents

#+title Questions

1 Imputation is bad because

- Doesnt account of feature correlation
- Reduces variance of the data, increase bias
- Less accurate model, narrower confidence interval due to smaller variance

2 Outlier

- Finding an outlier by Z-Score : z-score of a datapoint with is +/-3 is an outlier
  - Z score can't be used with small datasets
  - Data has to be normally distributed
- Q1 - 1.5(IQR) or Q3 + 1.5(IQR)
- DBScan clustering, Isolation Forest
- Anomaly detection algorithms

3 Inlier

- Identifying them is hard: Needs external data to identify them eg: A simple example of an inlier might be a value in a record reported in the wrong units (F instead of C - temperature)

4 [Missing data]

- Delete data if there is a lot of observations
- Mean, median, mode imputation
- Assigning a unique value for data missing
- Predict the missing values
- use RF

5 [Call centre duration]

- Durations follow a Log Normal distribution
- Use a QQ plot to confirm it

```
# create a random normal distribution; create the log-normal distribution of the same c
```

6 [Administrative data]

- Used by governments for non statistical purposes
- Could have human errors, missing values, wrong formats etc.
- Large and cost efficient

7 [80-20 rule]

- 80% of effects come from 20% of the causes

8 [Lift]

- Measure of the performance of a targeting model against a random choice targeting model. How much better your model is compared to having no model.

9 [Causation]

- Why a cause occured
- Correlation measured by Pearson's correlation. corr(A, B) = corr(B, A)
- Causation can be tested by Hypothesis testing

10 [Law of Large Numbers]

- According to the law, the average of the results obtained from a large number of trials should be close to the expected value and tends to become closer to the expected value as more trials are performed.

11 [Number of Samples Needed]

- Margin of Error, ME $= t * \frac{s}{\sqrt{n}} = z * \frac{\sigma}{\sqrt{n}}$

12 [Reduce Sampling Error Methods]

- Randomize the population so that every sample is drawn with equal probability (Random Sampling)

13 [Confounding Variable]

- Variable that influences both independent and dependent variable causing a spurious association; two or more variables are associated but not causally related

14 Infection rates at a hospital above a 1 infection per 100 person-days at risk are considered high.

A hospital had 10 infections over the last 1787 person-days at risk. Give the p-value of the correct one-sided test of whether the hospital is below the standard.

- Since we looking at the number of events (# of infections) occurring within a given timeframe, this is a Poisson distribution question.
- Null (H0): 1 infection per person-days
- Alternative (H1): >1 infection per person-days

Probability of k events occuring in an interval = $\frac{\lambda^k e^{-\lambda}}{k!}$