Anomaly Exercises.

1. (a) The z-score table indicates that the % of data value below $z = -3$ is 0.00135. Thus, the # of outliers is $1000000 \times 0.00135 \times 2 = \boxed{2700}$

(b) We know: sd (uniform distr.) $= \sqrt{\frac{(b-a)^2}{12}}$ and mean $= \frac{a+b}{2}$

∴ the observation that are not outliers are

$$N \times \frac{1}{b-a} \cdot b\sqrt{\frac{(b-a)^2}{12}} = \frac{\sqrt{3}(b-a)N}{(b-a)} = \sqrt{3}N > N.$$

This makes no sense for uniform distr. since there are no data exceeding magnitude of three sd away from mean.

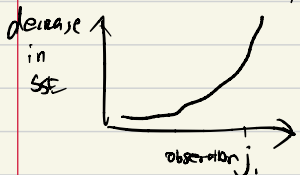
2. num_anomalies: $0.01 \times 0.99 = 0.0099$.
total_anomalies: $1 - 0.99 = 0.01$
false_anomalies: $0.99 \cdot 0.01 = 0.0099$
classified_as_anomalies: $0.99 \cdot 0.01 + 0.01 \cdot 0.99 = 0.0198$
detection_rate: $\frac{0.0099}{0.01} = 0.99$
false_alarm_rate
$= \frac{0.0099}{0.0198} = 0.5$

3. It does not provide new info since observations are uniformly distributed in space. Therefor, "outlier" that is intended to detect observations that is sparsely distributed is not meaningful.

4. Find the mean of the n observations. Then order from small to largest for the n observations. Find the ones furthest from the mean.

Q5 We should use hierarchical clustering with single linkage. We want to apply this procedure to build the full cluster. The observations that cannot be clustered together are candidate outliers. When k=5, we traverse the cluster to assign scores on nodes in the following procedure: add 1 if obj. i is not grouped with at least another k observations for each level.

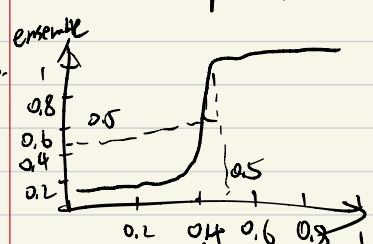
Q6 We use the elbow method. If we increase j, we will receive lower SSE. We want to graph the drop of SSE against each observations, the graph is shown below. As decrease in SSE become more quicker, all points on the RHS are outliers.



Final Exercises.

3. Bootstrap sampling ends up in OOB samples by sampling the original sample with replacement. A new distribution based on the variation of the original sample is displayed. The new added samples are viewed as OOB samples. This assists in error estimation of the bootstrap distribution. as the new distribution assures the result is not disputable. Thus, it helps out in error estimation procedure.

4.



5. I worked with my family members on zoom to pick from post cards. I drew with replacement from full 50 post cards of different types. I eventually got 32 out of the 50 cards, which is about 60% ($\frac{2}{3}$) of the full data set.

7. NBA player data to predict and inference on player's performance and future value.