SML Exam 2022

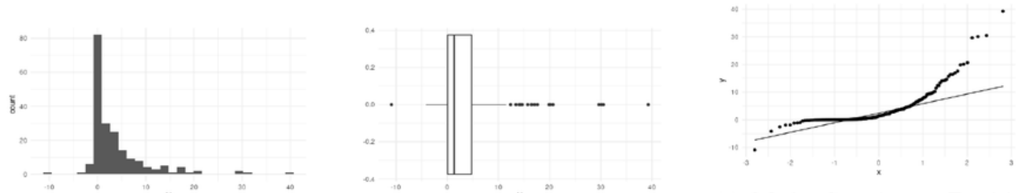
Task 1 Mean-square error of an estimate can be expressed as:

- $bias + var^2$
- $var^2 + bias^2$
- $var + bias^2$

Task 2 Metropolis-Hastings algorithm:

- Gives biased estimates for non-symmetric proposal distribution
- Converges to the stationary distribution which is true posterior
- Can be used to sample posterior for hierarchical Bayesian models

Task 3 Based on the figures, indicate correct answers:



(a) Histogram

(b) Boxplot

(c) QQplot Gaussian vs. Empirical

- Plot suggest that the data are non-Gaussian
- The sample mean is positive
- The sample median is larger than the mean

Task 4 The data set $X_1, \ldots, X_n$ was clustered into k groups using different techniques.

- The hierarchical clustering requires to specify the number of clusters at the beginning
- In the k-means algorithm, the centres of clusters $C_i \in X_1, \ldots, X_j)$
- The k-medoid method does not depend on the initial clustering

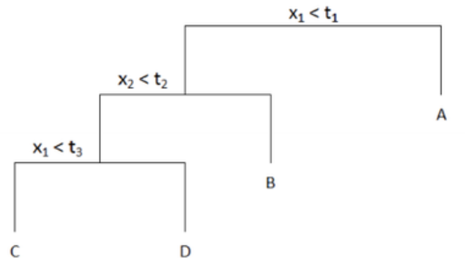
Task 5 The laboratory of prof. Unreal investigates whether alcohol has impact on reaction time. Indicate the correct procedure. Let T_a denote mean reaction time after drinking and T_b before drinking.

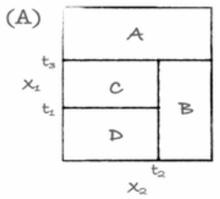
- To verify that alcohol increases reaction time, they perform a test with $H_0 : T_a = T_B$ vs $H_1 : T_a > T_b$.
- To verify that alcohol changes reaction time, they perform a test with $H_0 : T_a = T_B$ vs $H_1 : T_a < T_b$.
- To verify that alcohol increases reaction time, they perform a test with $H_0 : T_a = T_B$ vs $H_1 : T_a \leq T_b$.

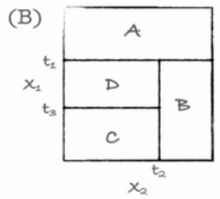
Task 6 The laboratory found 50 potential candidates for genes responsible for faster or slower growth of a cancer tumor. To verify them, the experiments were performed and results of each experiment were concluded by performing a statistical test (H_0: gene has no impact vs H_1: gene is important). Indicate correct statements.

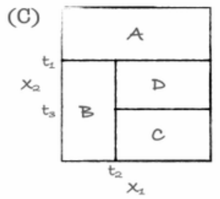
- We can say that genes with p-values $<$ are significant, but such a procedure does not control a fraction of false discoveries
- To control family-wise error rate on level 0.1 we can choose genes with p-value < 0.005
- To control FDR we can apply Benjamini-Hochberg procedure

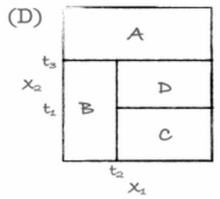
Task 7 Consider the decision tree:











- Tree corresponds to (B)
- if $x_1 > t_1$ the prediction is A
- if $x_2 > t_2$ we can not predict C or D

Task 8 Consider data set D, with n observations and p features. We use a Random Forest model with hyperparameters: m (number of sampled features) and T (number of trees).

- There exists m such that for every data D we have the smallest training error
- There exists m that for every data D we have the smallest test error
- For given data D there exists m and T which minimizes training error

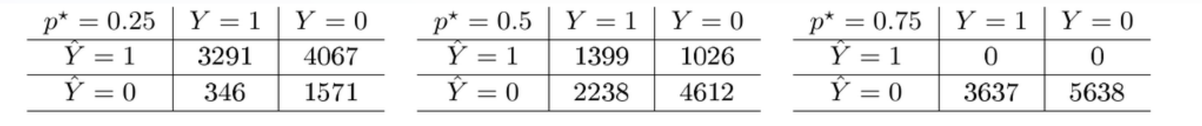
Task 9 Consider maximum likelihood estimator in regular family.

- This estimator is unbiased
- This estimator is consistent
- This estimator is asymptotic notmal, but we cannot say anything about its asymptotic variance

Task 10 Consider the linear model (A) $Y = XB + e$ and smaller model (B) with removed r columns from X.

- The RSS in the model (A) is larger that in model (B)
- The R^2 coefficient in model (A) is larger than in model (B)
- If model (B) is true, the test error in model (B) will be always smaller than in model (A)

Task 11 Based on the confusion matrix:



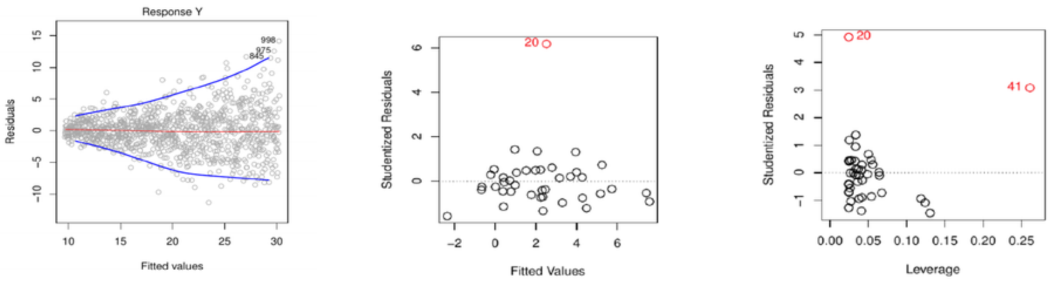
$p^\star = 0.25$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	3291	4067
$\hat{Y} = 0$	346	1571

$p^\star = 0.5$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	1399	1026
$\hat{Y} = 0$	2238	4612

$p^\star = 0.75$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	0
$\hat{Y} = 0$	3637	5638

- Accuracy for $p* = 0.5$ is higher than for $p* = 0.25$
- We have not enough information to compute FDR for $p* = 0.8$
- The sensitivity for $p* = 0.75$ is higher than for $p* = 0.25$

Task 12 Based on diagnostic plots, indicate true sentences:



(a) (b) (c)

- Observation 41 is an outlier but not high leverage
- The variance of noise depends on the value of the response
- Observation 20 is an outlier but not high leverage