CAS Exam MAS-2
Practice Exam #1

These practice exams should be used during the month prior to your exam.

This practice exam contains **42 questions**, of equal value,
corresponding to about a **4 hour** exam.

Each problem is similar to a problem in my study guides, sold separately.
Solutions to problems are at the end of each practice exam.

prepared by
Howard C. Mahler, FCAS
Copyright ©2026 by Howard C. Mahler.

Howard Mahler
hmahler@mac.com
www.howardmahler.com/Teaching

CAS Exam MAS-2, Practice Exam #1

1. Use the following information:
- A neural network has one input node, two hidden layers each with two units, and one output node.
- $w_{1,0}^{(1)} = -1,\ w_{1,1}^{(1)} = 5,\ w_{2,0}^{(1)} = 4,\ w_{2,1}^{(1)} = -2.$
- $w_{1,0}^{(2)} = 3,\ w_{1,1}^{(2)} = -4,\ w_{1,2}^{(2)} = 2,\ w_{2,0}^{(2)} = -2,\ w_{2,1}^{(2)} = 6,\ w_{2,2}^{(2)} = 3.$
- $\beta_0 = 10,\ \beta_1 = 2,\ \beta_2 = 1.$
- The neural network uses the ReLU activation function.
- The output is quantitative.

Determine the output for X = 0.5.

[ ] (round to the nearest integer)

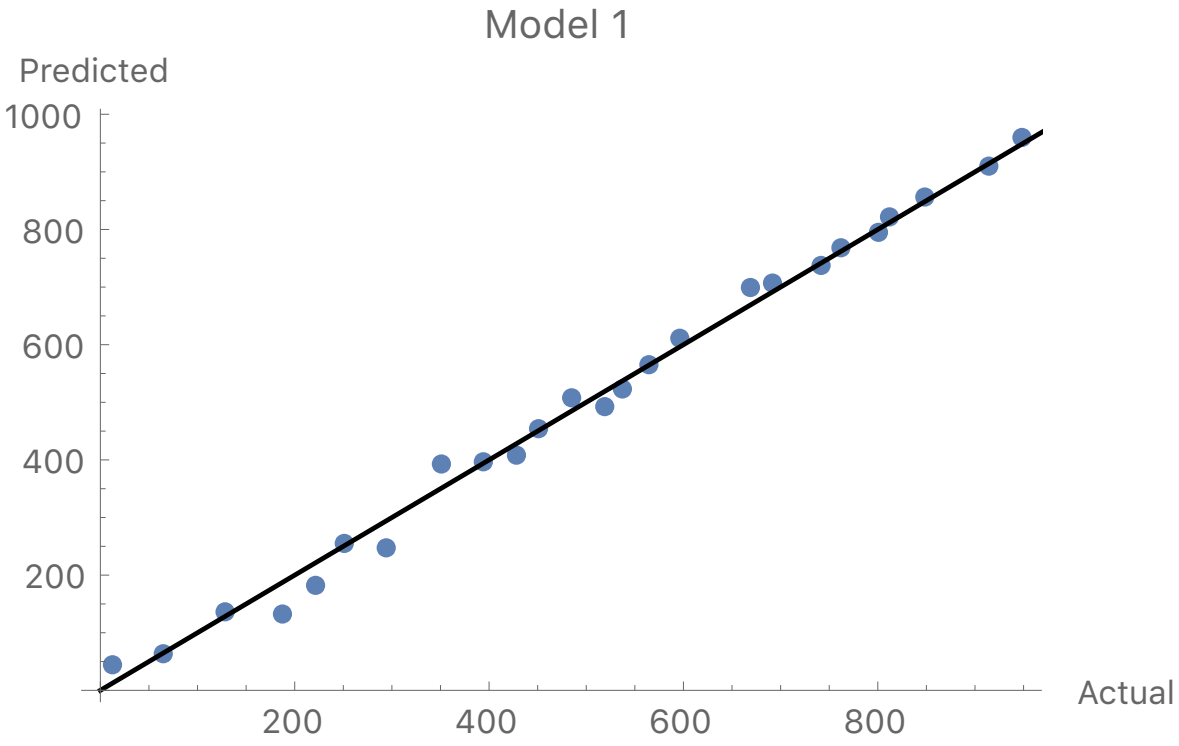
2. You are given:
- The number of mistakes that any particular cashier makes per hour follows a Poisson distribution with mean λ.
- The prior distribution of λ is assumed to follow a Gamma Distribution with $\alpha = 0.8$ and $\theta = 1/40$.
- A particular cashier, Marvelous Marv, is observed for 100 hours and makes 6 errors.

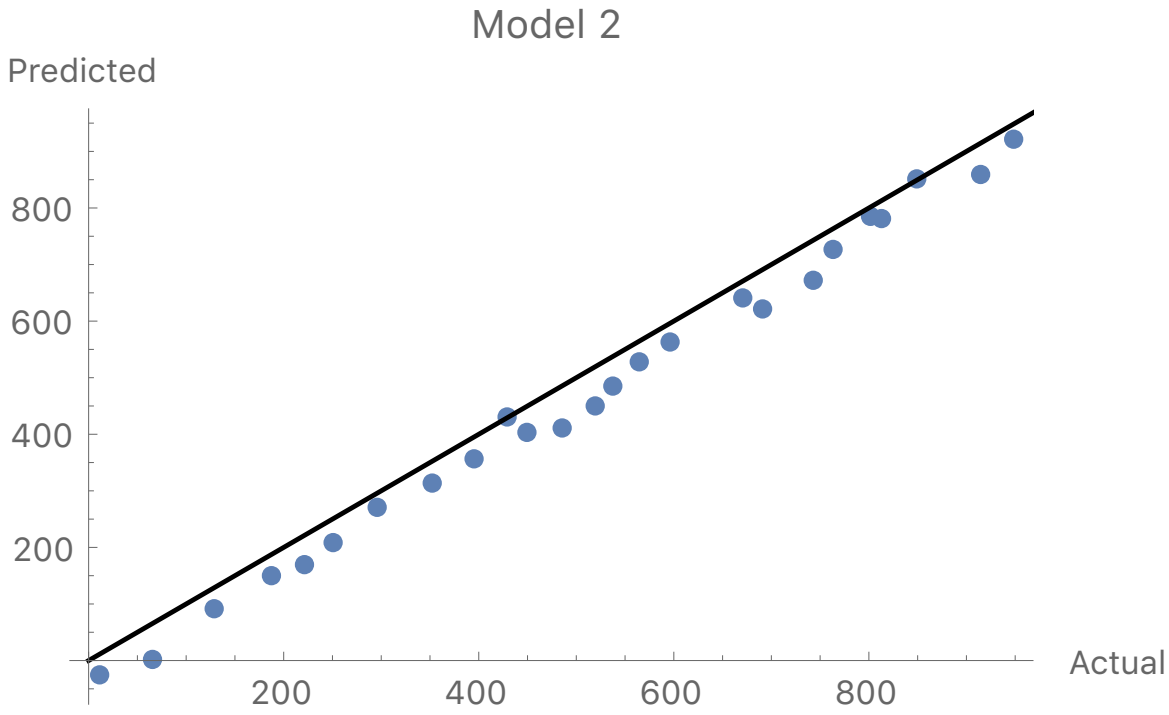
Determine the expected number of errors that Marv will make in his next 100 hours.
A. Less than 3.0
B. At least 3.0, but less than 3.5
C. At least 3.5, but less than 4.0
D. At least 4.0, but less than 4.5
E. At least 4.5

3. You are given the following information:
- The time series $\{X_t\}$ is stationary in mean and variance.
- The first 8 observed values of x are as follow: 23, 25, 22, 26, 27, 26, 30, 29.

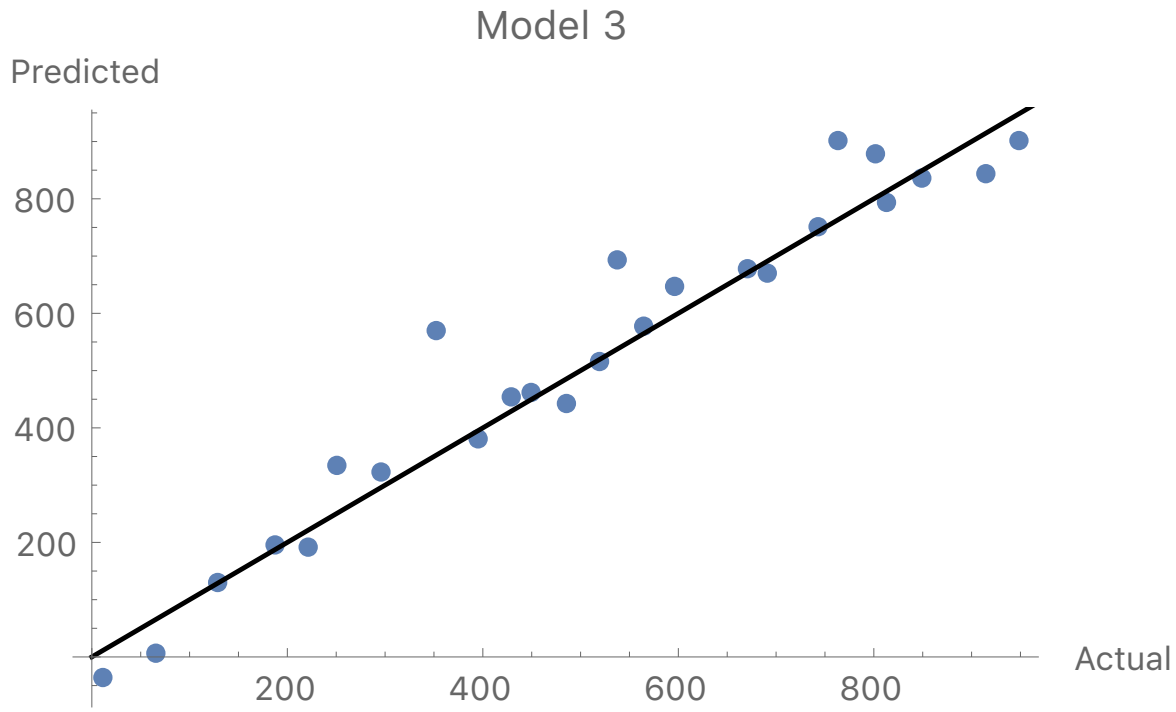
Calculate the sample auto covariance function, c_k, for lag k = 2.
A. 1.5 B. 1.6 C. 1.7 D. 1.8 E. 1.9

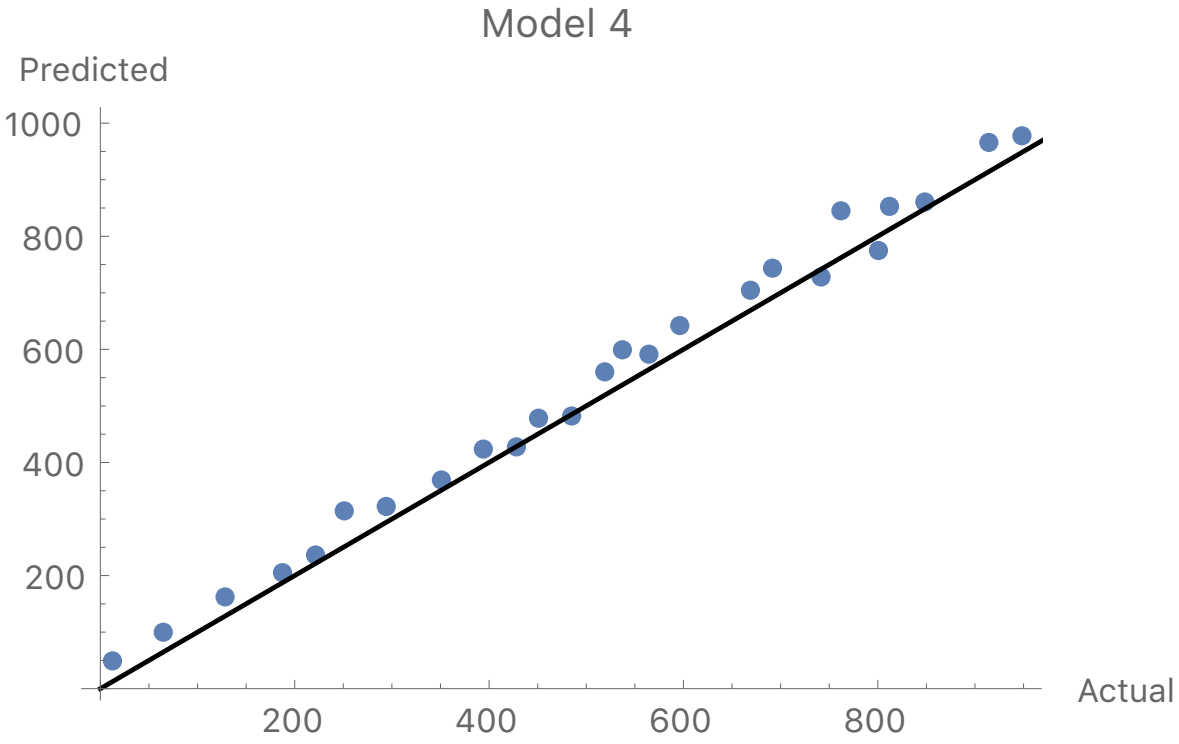
4. For four GLMs you are given the following graphs based on holdout data:





QUESTION CONTINUED ON THE NEXT PAGE

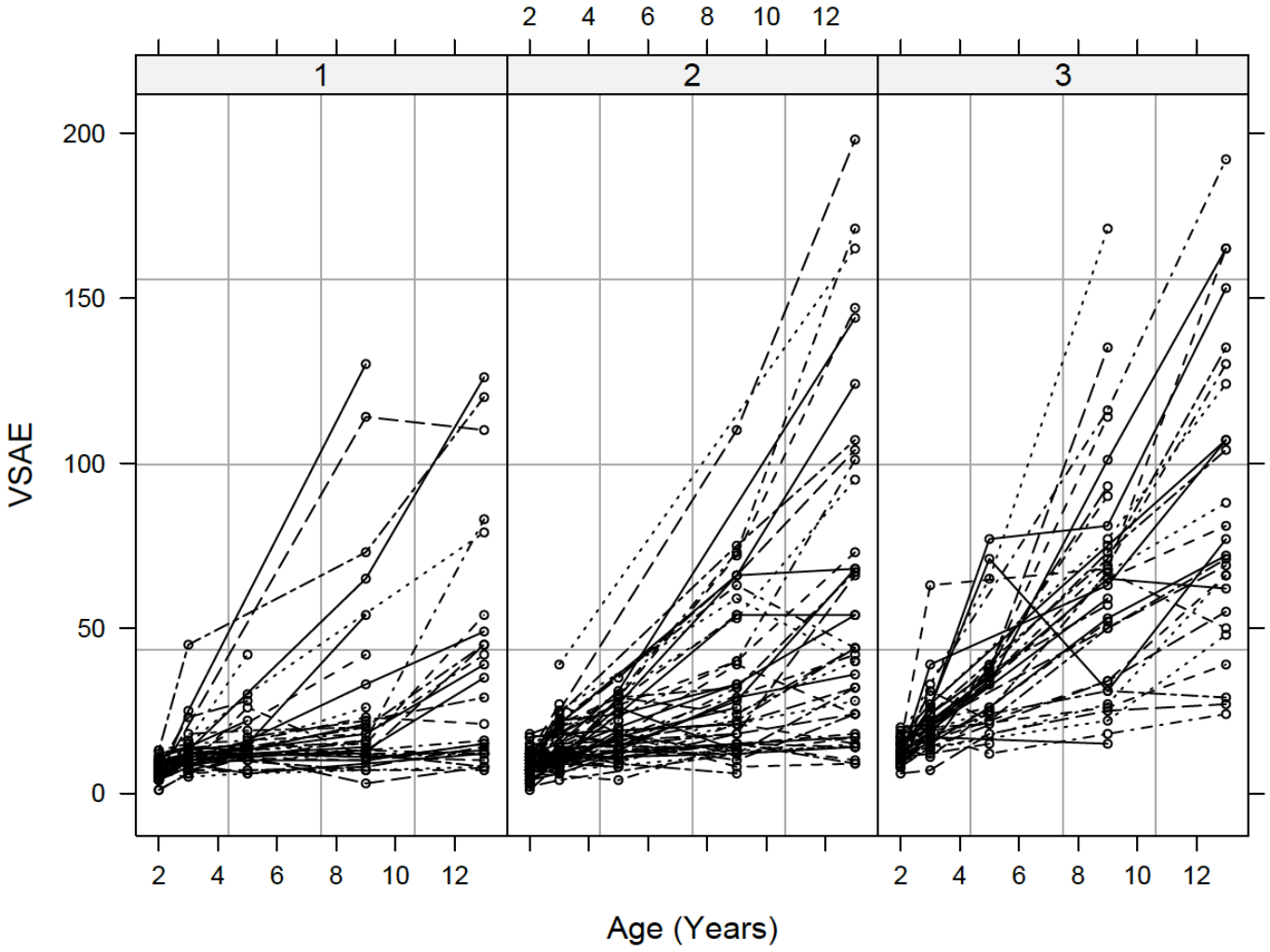




Which model do you prefer?

Model # ☐

5. Data was collected on the development of children at ages 2, 3, 5, 9, and 13 years.
A subset of 158 autistic or pervasive developmental disorder (PDD) children was considered.
A model was developed of the influence of language proficiency at age 2 on the developmental trajectories of the socialization of these children. Initial language development was assessed using the Sequenced Inventory of Communication Development (SICD) scale.
Data for individual children by SICD group:



Which of the following statements is false?
A. For each SICD group, there is substantial variation from child to child.
B. For some children their VSAE scores increased with age, while for others their scores did not.
C. For each SICD group there is not much variability in the initial values of VSAE at age 2.
D. For each SICD group, at each successive year of age there is increasing between-child variability in the VSAE scores.
E. None of A, B, C, or D is false.

6. You are constructing a classification tree using Entropy.
Each of 100 observations is classified as Yes or No.
You are considering various splits into two regions. Which one do you select?
A. {45 Yeses, 15 Nos} and {15 Yeses and 25 Nos}
B. {40 Yeses, 10 Nos} and {20 Yeses and 30 Nos}
C. {43 Yeses, 11 Nos} and {17 Yeses and 29 Nos}
D. {50 Yeses, 17 Nos} and {10 Yeses and 23 Nos}
E. {35 Yeses, 8 Nos} and {25 Yeses and 32 Nos}

7. A time series is modeled using the function below:

$$x_t = \alpha_0 + \alpha_1 t + z_t$$

- z_t is a white noise series
- $z_1 = -5.1$
- $z_3 = 7.2$
- The first order difference at time t = 2 is $\nabla x_2 = -11.4$
- The first order difference at time t = 3 is $\nabla x_3 = 2.5$
- $x_3 = 27.7$

Calculate the forecast value of x_6.
A. Less than -20
B. At least -20, but less than -10
C. At least -10, but less than 0
D. At least 0, but less than 10
E. At least 10

8. A data set has two predictors X_1 and X_2 and two classes T and F.

$Pr[Y = F \mid X_1 = x_1 \text{ and } X_2 = x_2] = \exp[-(x_1^2 + x_2^2)/100]$.
You are given the following test data:

i	x_1	x_2	y
1	2	-6	T
2	5	0	F
3	7	6	T
4	13	-10	F
5	-2	9	T
6	-5	-4	T

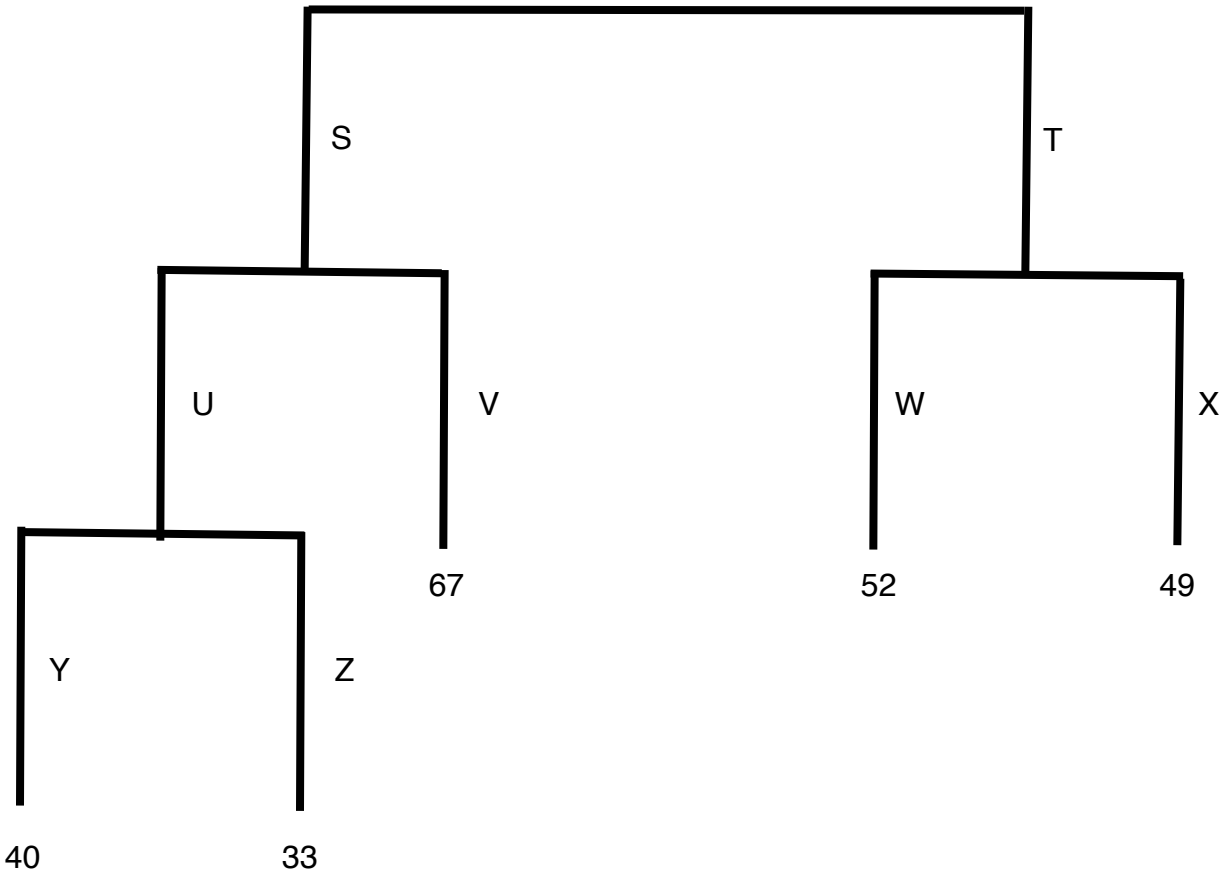
Calculate the Bayes error rate on the test data.
A. 26% B. 28% C. 30% D. 32% E. 34%

9. You are given the following:
Prior to observing any data, you assume that the claim frequency rate per exposure has mean = 0.08 and variance = 0.12.
A full credibility standard is devised that requires the observed sample frequency rate per exposure to be within 10% of the expected population frequency rate per exposure 99% of the time.
You observe 112 claims on 1,000 exposures.
Estimate the number of claims you expect for these 1000 exposures next year.
A. 89 B. 91 C. 93 D. 95 E. 97

10. Determine which of the following statements about Recurrent Neural Networks are true.
I. In Seq2Seq learning, two tracks of hidden-layer activations are maintained.
II. In LSTM, both the input sequence and the target sequence are represented by a structure similar to a simple recurrent neural network, and they share the hidden units.
III. When using a Recurrent Neural Network to classify a document, one can represent each word in a much lower-dimensional embedding space.
(A) I only (B) II only (C) III only (D) I, II, and III
(E) The correct answer is not given by (A), (B), (C), or (D)

11. You are given:
(i) X_i is the claim count observed for insured i for one year.
(ii) X_i has a Negative Binomial Distribution with parameters $\beta = 0.7$ and r_i.
(iii) The r_i's have a Gamma Distribution with parameters α and θ.
Determine the Buhlmann credibility parameter, K.
(A) 1.19α (B) $1.19\alpha/\theta$ (C) 2.43α (D) $2.43\alpha/\theta$ (E) None of A, B, C, or D

12. You are given the following unpruned regression tree:



The values at each terminal node are the contribution to the residual sums of squares (RSS) at that node. The table below gives gives the contribution to the RSS at nodes S, T, and U if the tree was pruned at those nodes:

Node	RSS
S	193
T	114
U	91

The RSS for the null model is 345. You use the cost complexity pruning algorithm with the tuning parameter, α, equal to 20 in order to evaluate the following pruning strategies. Determine which of the following pruning strategy is selected.

A. No nodes pruned
B. Prune both nodes Y and Z
C. Prune both nodes W and X
D. Prune both nodes U and V
E. Prune both nodes S and T

13. You are given the following information:

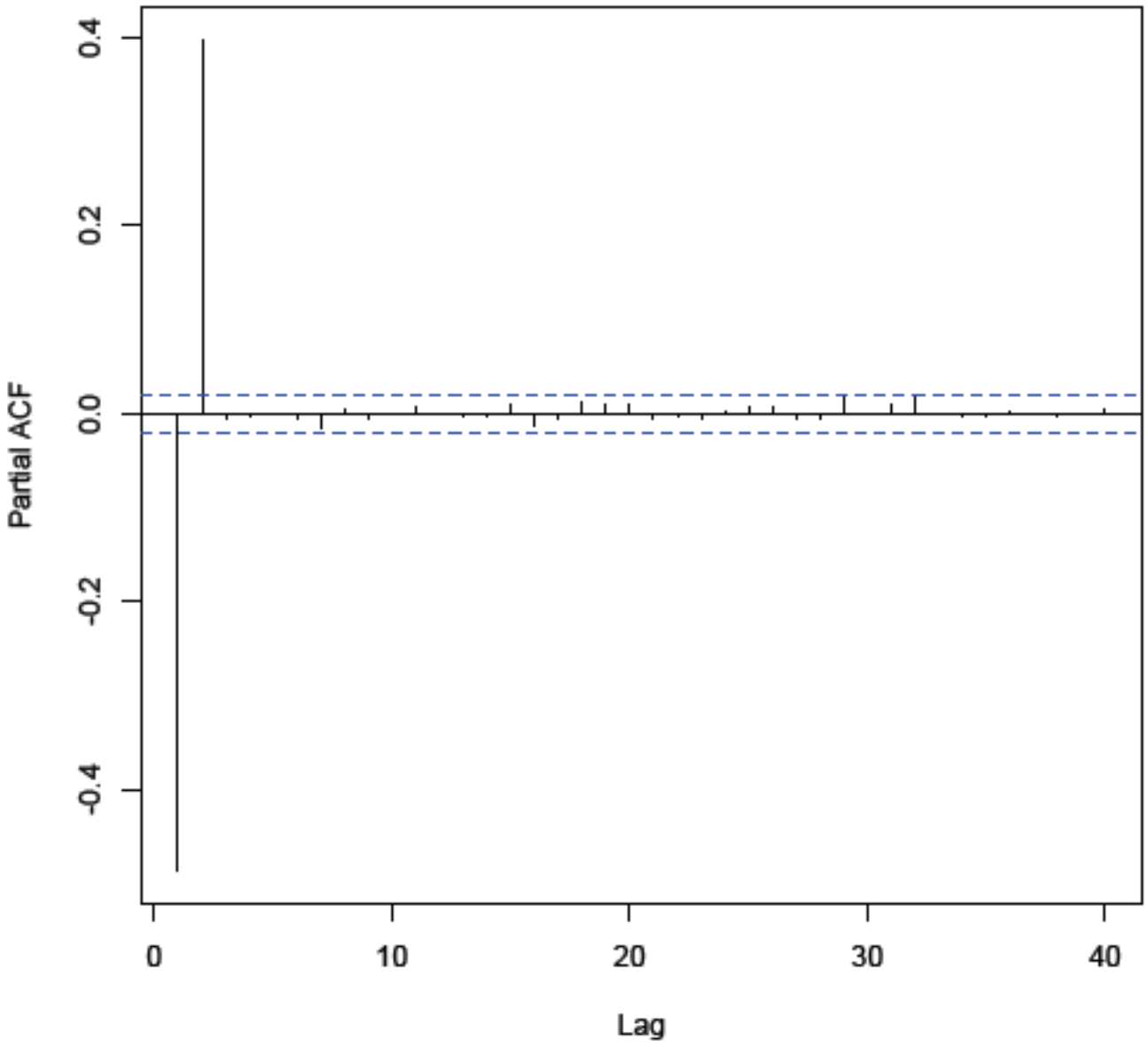
- A standard for full credibility of 5,000 claims has been selected so that the actual pure premium would be within 5% of the expected pure premium 98% of the time.
- The number of claims follows a Poisson distribution, and is independent of the severity distribution.

Using the concepts of classical credibility, determine the coefficient of variation of the severity distribution underlying the full credibility standard.

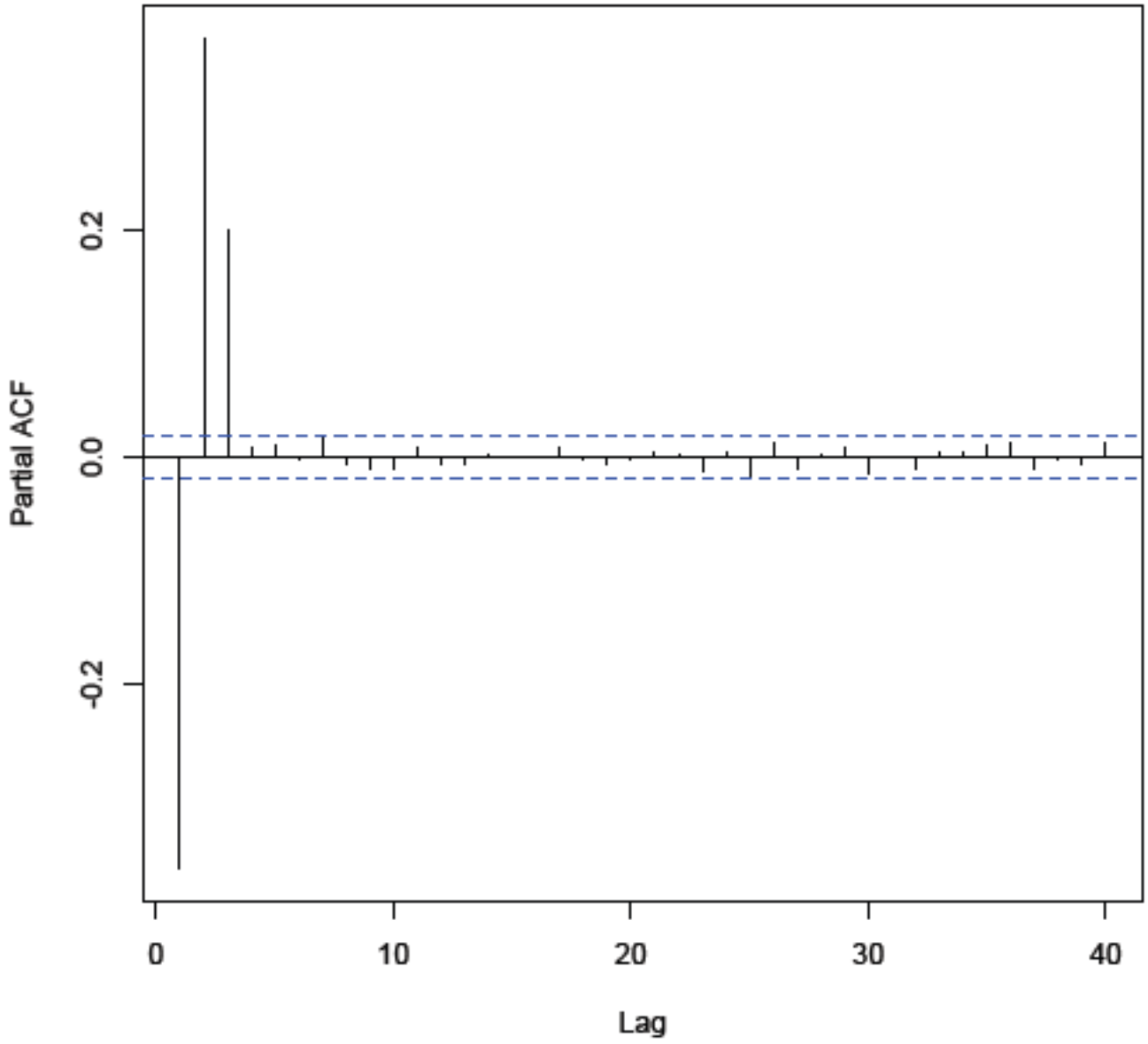
A. Less than 1.20
B. At least 1.20 but less than 1.25
C. At least 1.25 but less than 1.30
D. At least 1.30 but less than 1.35
E. At least 1.35

14. Which of the following graphs of sample partial autocorrelations is from an AR(4) model?

A.

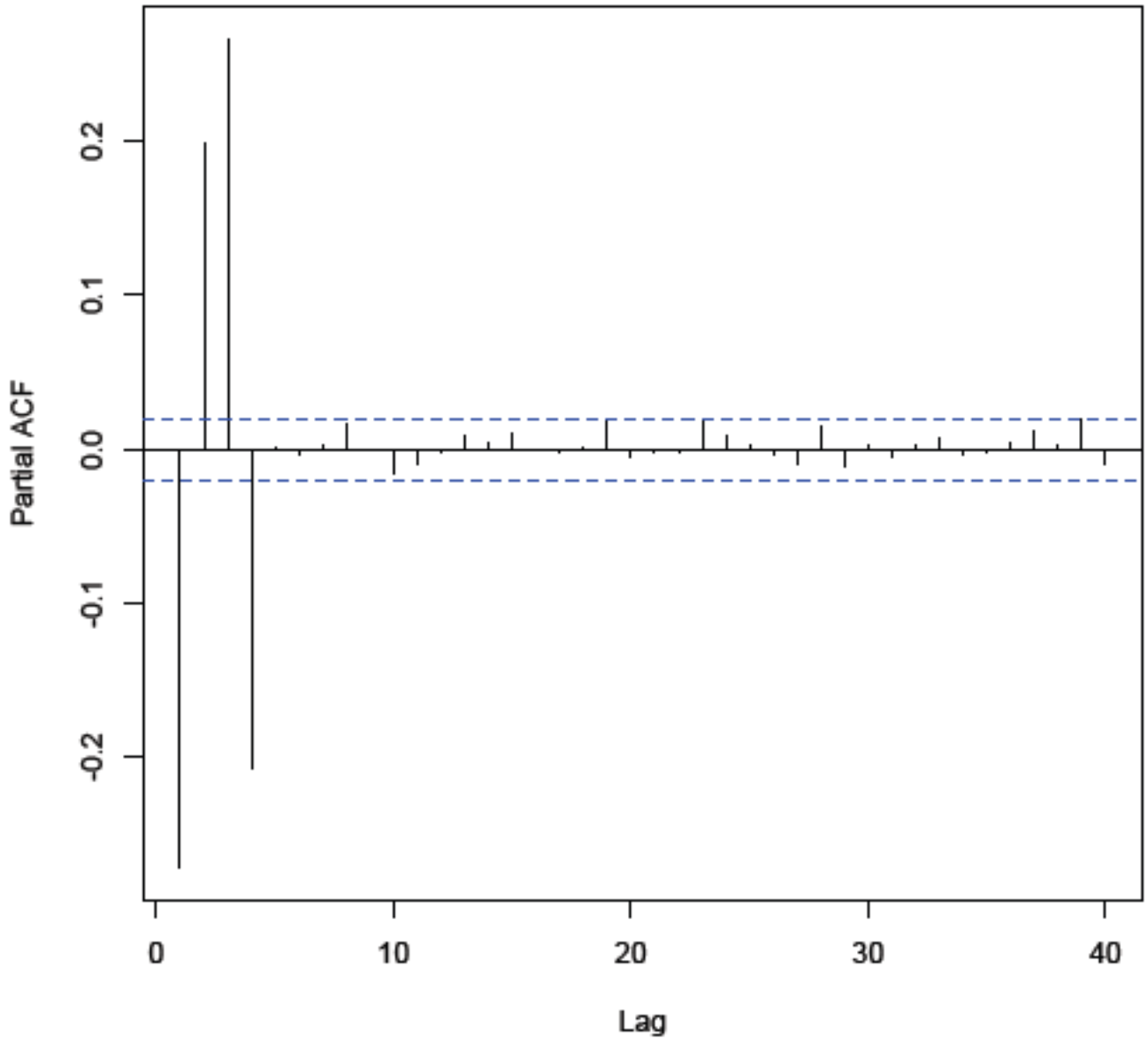


QUESTION CONTINUED ON THE NEXT PAGE



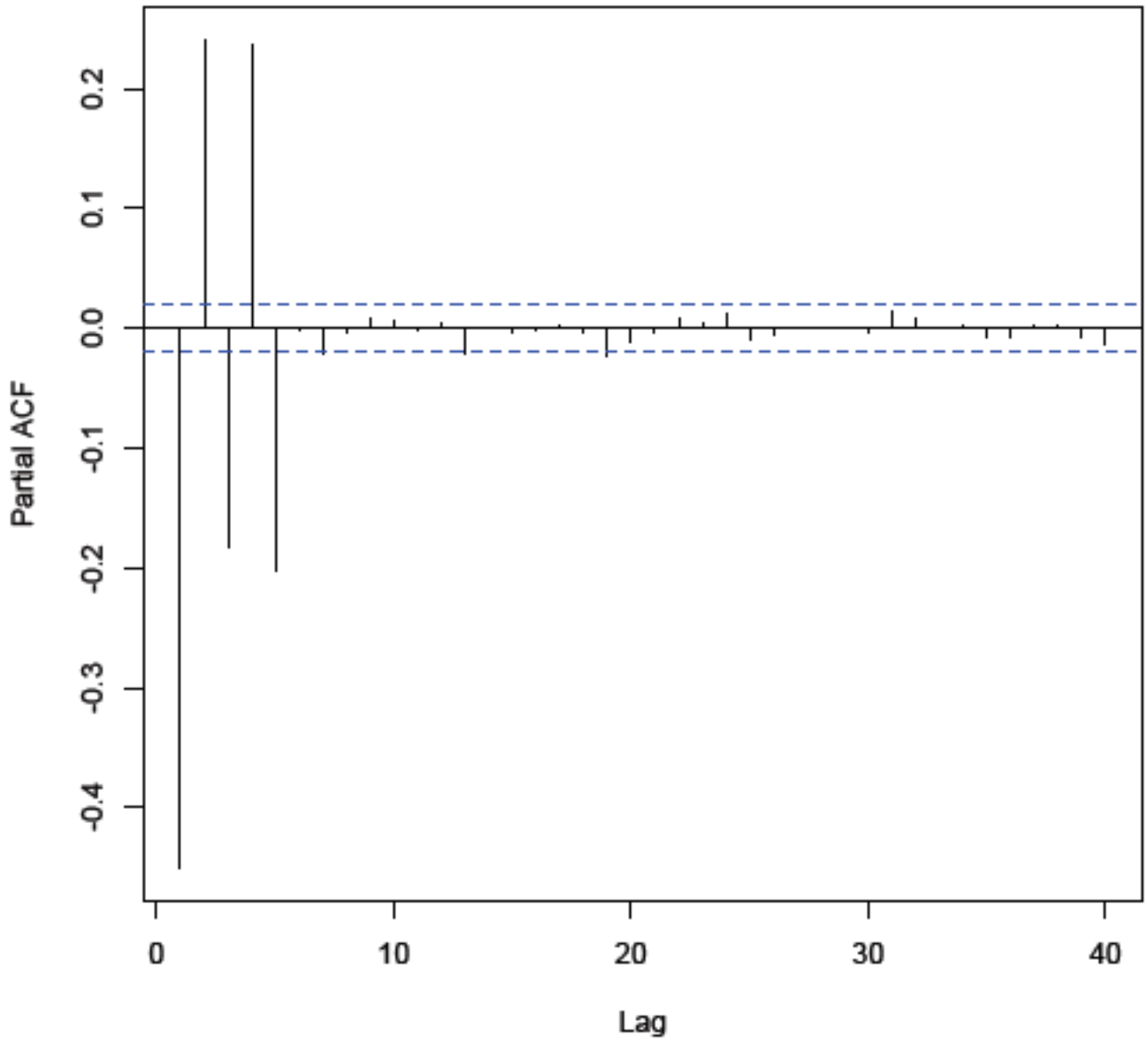
B.

QUESTION CONTINUED ON THE NEXT PAGE



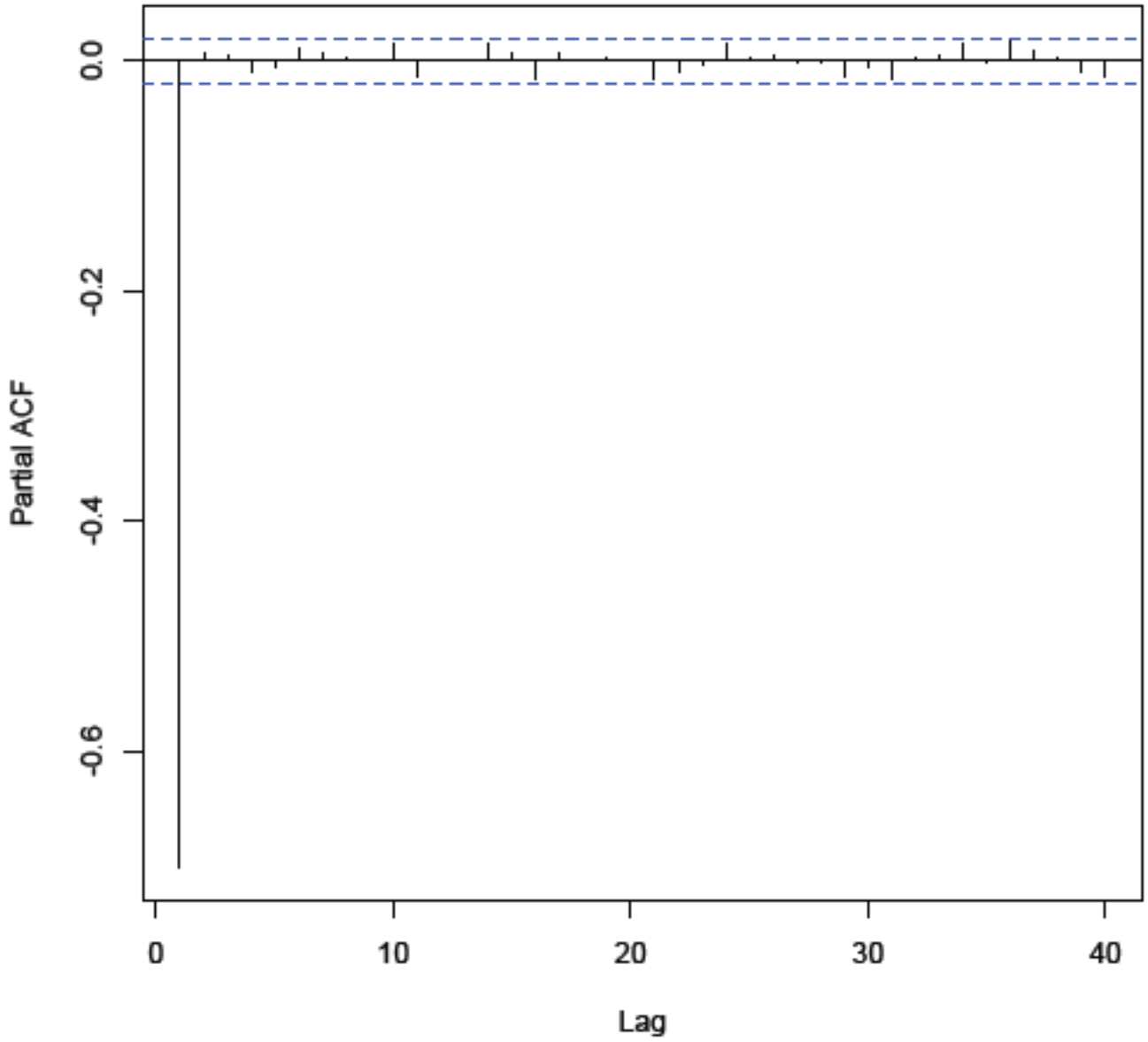
C.

QUESTION CONTINUED ON THE NEXT PAGE



D.

QUESTION CONTINUED ON THE NEXT PAGE



E.

15. The number of claims that a particular policyholder makes in a year is Bernoulli with mean q.
The q values of the portfolio of policyholders follow a Beta Distribution as per the MAS-2 Tables, with a = 3, b = 14, and $\theta = 1$.
An insured has 5 claims over 20 years.
What is the posterior probability density function for this insured's Bernoulli parameter q?
A. $308{,}864{,}160\ q^7\ (1-q)^{29}$
B. $1{,}467{,}104{,}760\ q^8\ (1-q)^{29}$
C. $242{,}082{,}720\ q^7\ (1-q)^{28}$
D. $1{,}119{,}632{,}580\ q^8\ (1-q)^{28}$
E. None of A, B, C, or D

16. You are given the following joint distribution:

	Θ	
X	1	2
100	0.3	0.1
200	0.2	0.1
500	0.1	0.2

For a given value of Θ and a sample of size 3 for X:

$$\sum_{i=1}^{3} x_i = 900.$$

Determine the Bühlmann credibility premium, in other words the estimated future value of X using Buhlmann Credibility.
(A) 260 (B) 265 (C) 270 (D) 275 (E) 280

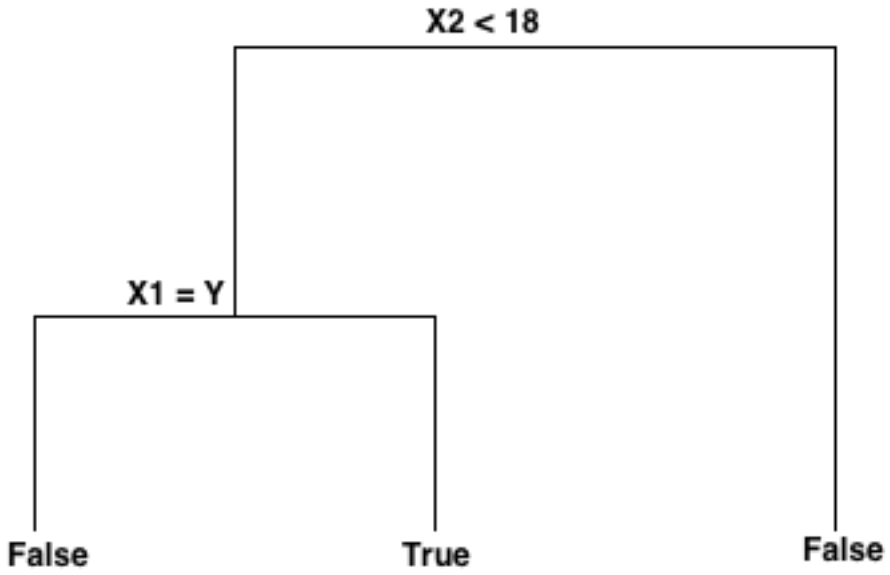
17. Determine which of the statements I, II, and III are true.
I. Boosting has three tuning parameters.
II. Boosting makes predictions from an average of regression trees,
each of which is built using a random sample of data and predictors.
III. In boosting, using large trees often works well.
A. I only B. II only C. III only D. I, II, and III
E. The answer is not given by (A), (B), (C), or (D)

18. Each insured's claim frequency follows a Poisson Distribution.
There are two types of insureds as follows:

Type	A Priori Probability	Mean Annual Claim Frequency (Poisson Parameter)
A	70%	10%
B	30%	20%

You observe 2 claims by an individual in a single year.
Use Bayesian Analysis to predict that individual's future claim frequency.
A. 15% B. 16% C. 17% D. 18% E. 19%

19. You are given the following classification decision tree and data set, with two predictors X_1 and X_2 and the response Z.



Observation	X_1	X_2	Z
1	Y	37	T
2	N	27	F
3	Y	17	F
4	N	38	T
5	N	25	T
6	Y	12	F
7	Y	4	F
8	N	10	T
9	Y	22	F
10	N	16	F

● The Entropy is calculated using the natural log.
● G = a weighted average of the Gini Indices for every terminal node.
● D = a weighted average of the Entropies for every terminal node.
Determine D - G.
A. less than 0.10
B. at least 0.10 but less than 0.12
C. at least 0.12 but less than 0.14
D. at least 0.14 but less than 0.16
E. at least 0.16

20. The three models AR(1), AR(2) and ARMA(2,1) are fitted to a time series of length 100. The results using R are as follows:

AR(1)		ar1	intercept
		0.3275	-0.1156
	Std Err	0.0950	0.1701
σ^2 = 1.321	log likelihood = -155.85		

AR(2)		ar1	ar2	intercept
		0.4548	-0.3827	-0.1127
	Std Err	0.0931	0.0922	0.1147
σ^2 = 1.124	log likelihood = -147.96			

ARMA(2,1)		ar1	ar2	ma1	intercept
		0.2092	-0.2932	0.2825	-0.1133
	Std Err	0.2986	0.1530	0.3193	0.1254
σ^2 = 1.121	log likelihood = -147.84				

For the model with the best AIC, what is the upper end of a 95% confidence interval for α_1?

(A) Less than 0.4
(B) At least 0.4, but less than 0.5
(C) At least 0.5, but less than 0.6
(D) At least 0.6, but less than 0.7
(E) At least 0.7

21. A Linear Mixed Model with longitudinal observations is being built.
The observations within a subject exhibit positive auto correlation.
The variance of the residuals is 0.800.
Determine which of the following covariance matrices of the residuals exhibits the first-order autoregressive structure.

A. $\begin{pmatrix} 0.800 & 0.600 & 0.600 & 0.600 \\ 0.600 & 0.800 & 0.600 & 0.600 \\ 0.600 & 0.600 & 0.800 & 0.600 \\ 0.600 & 0.600 & 0.600 & 0.800 \end{pmatrix}$

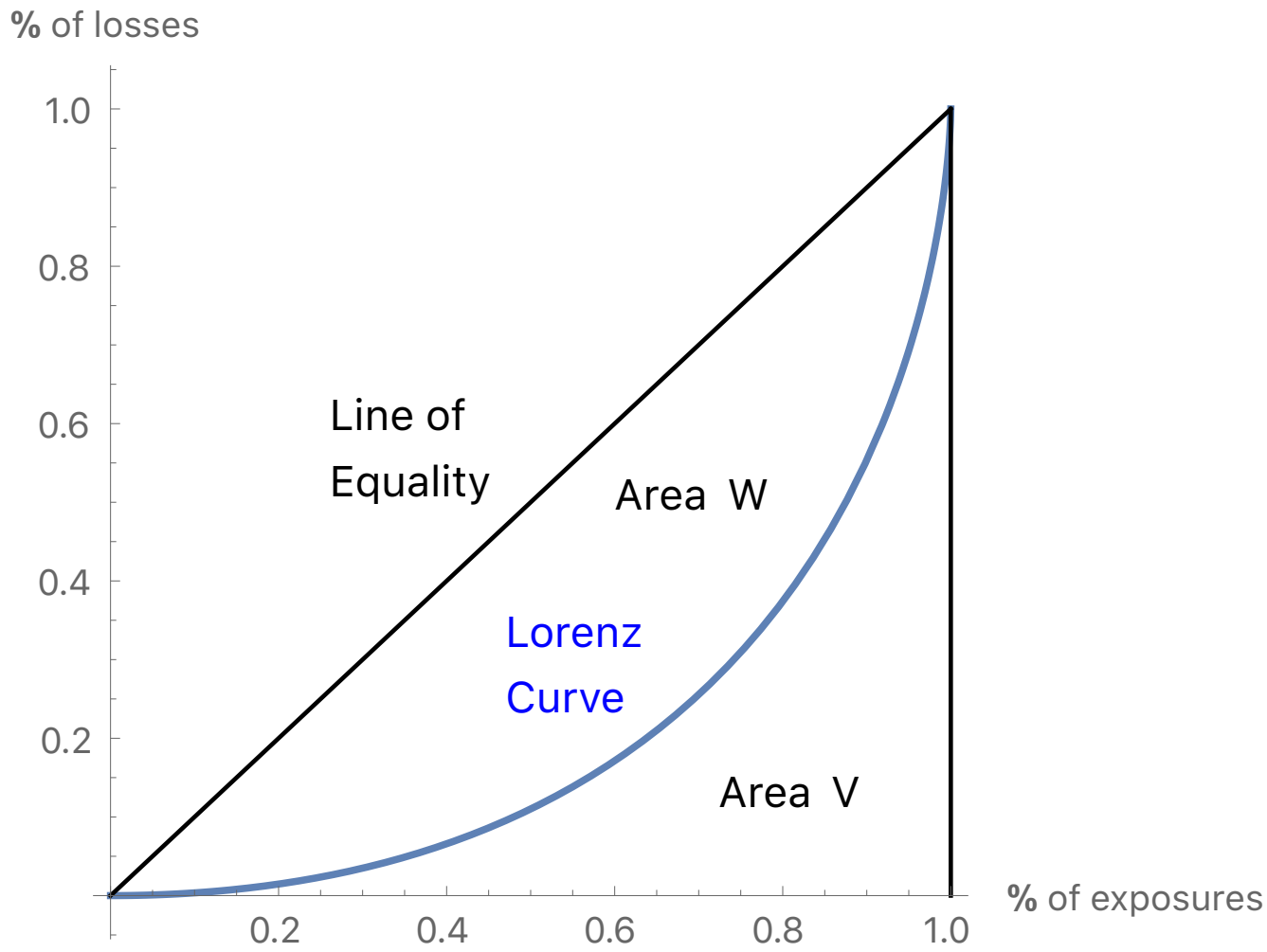
B. $\begin{pmatrix} 0.800 & 0.600 & 0.400 & 0.200 \\ 0.600 & 0.800 & 0.600 & 0.400 \\ 0.400 & 0.600 & 0.800 & 0.600 \\ 0.200 & 0.400 & 0.600 & 0.800 \end{pmatrix}$

C. $\begin{pmatrix} 0.800 & 0.600 & 0.000 & 0.000 \\ 0.600 & 0.800 & 0.600 & 0.000 \\ 0.000 & 0.600 & 0.800 & 0.600 \\ 0.000 & 0.000 & 0.600 & 0.800 \end{pmatrix}$

D. $\begin{pmatrix} 0.800 & 0.400 & 0.200 & 0.100 \\ 0.400 & 0.800 & 0.400 & 0.200 \\ 0.200 & 0.400 & 0.800 & 0.400 \\ 0.100 & 0.200 & 0.400 & 0.800 \end{pmatrix}$

E. $\begin{pmatrix} 0.800 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.700 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.600 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.500 \end{pmatrix}$

22. Areas have been labeled in the following graph of a Lorenz Curve.



Determine which of the following are equal to the Gini index. (Select all that apply.)

- ☐ A. V
- ☐ B. W
- ☐ C. V/W
- ☐ D. 2W
- ☐ E. W/(V+W)

23. You are provided with the following normalized and scaled data set:

i	X_1	X_2	X_3
1	-0.5	0.5	1.225
2	-0.5	-2	-0.817
3	-0.5	0.5	-0.817
4	-0.5	0.5	1.225
5	2	0.5	-0.817

The first principal component loading vector of the data set is (-0.4438, 0.4438, 0.7785).
Calculate the proportion of variance explained by the first principal component.
A. Less than 42%
B. At least 42% but less than 44%
C. At least 44% but less than 46%
D. At least 46% but less than 48%
E. At least 48%

24. An actuary uses four separate models to fit a time series. All models have mean $\mu_x = 0$.

- Model 1: A random walk model with no drift
- Model 2: A stationary autoregressive process of order 1 with a root of the characteristic equation of the backwards shift operator equal to 1.5
- Model 3: A stationary autoregressive process of order 1 with a root of the characteristic equation of the backwards shift operator equal to 4
- Model 4: A non-stationary autoregressive process of order 1 with a root of the characteristic equation of the backwards shift operator greater than 0

The most recent values of x at time t, x_t, are given in the table below:

t	x_t
4	22
5	11
6	14

Determine which model will result in the largest predicted values of x_7.
A. Model 1
B. Model 2
C. Model 3
D. Model 4
E. There is not enough information given to determine the correct answer.

25. You are given:

	Group	Year 1	Year 2	Year 3	Total
Total Losses	1		160,000	114,000	274,000
Number in Group			200	190	390
Average			800	600	702.56
Total Losses	2	75,000	96,000		171,000
Number in Group		150	160		310
Average		500	600		551.61
Total Losses					445,000
Number in Group					700
Average					635.71

You are also given that the estimate of the Variance of the Hypothetical Means is 4631.
Use the nonparametric empirical Bayes method to estimate the credibility factor for Group 2.
(A) 0.38 (B) 0.40 (C) 0.42 (D) 0.44 (E) 0.46

26. Determine which of the following statements are true.
I. Bagging is a general-purpose procedure for reducing the variance of a statistical learning method.
II. In bagging classification trees, the importance of each predictor can be measured by the amount that splits over that predictor reduce the Gini index, averaged over the different trees.
III. In bagging, one takes many training sets from the population, builds a separate prediction model using each training set, and averages the resulting predictions.
A. I, II only B. I, III only C. II, III only D. I, II, and III
E. The answer is not given by (A), (B), (C), or (D)

27. You are given:
(i) Conditionally, given λ, an individual loss X follows the exponential distribution with probability density function: $f(x \mid \lambda) = \lambda \exp(-\lambda x)$, $0 < x < \infty$.
(ii) The prior distribution of λ is Gamma with probability density function:
$\pi(\lambda) = \exp(-\lambda 100)\ \lambda^2 / 500{,}000$, $\lambda > 0$.
(iii) An insured has 4 losses that total 140.
What is the probability that the next loss from this insured will be greater than 60?
A. 17% B. 19% C. 21% D. 23% E. 25%

28. A company implements a new strategy for managing its warehouses. It is implemented under the direction of several regional managers who are each responsible for different states. Jeff builds a Linear Mixed Model to model the increase in efficiency over the next year after this new strategy is implemented.
Below is the model output from this Linear Mixed Model.

Solution for Fixed Effects		
Effect	Location	Estimate
Intercept		2.81
Location	Rural	1.03
Location	Suburban	0.74
Location	Urban	0.00
Warehouse Age		-0.25
Warehouse Size		0.12

Solution for Selected Random Effects			
Effect	Manager	State	Estimate
Intercept	Connie		1.33
Intercept	Connie	S1	1.60
Intercept	Connie	S2	-0.77
Intercept	Miller		-0.43
Intercept	Miller	S6	-1.08
Intercept	Miller	S7	2.16

Consider a warehouse with the following characteristics:

Location	Age	Size	Director	State
Rural	7	4	Miller	S6

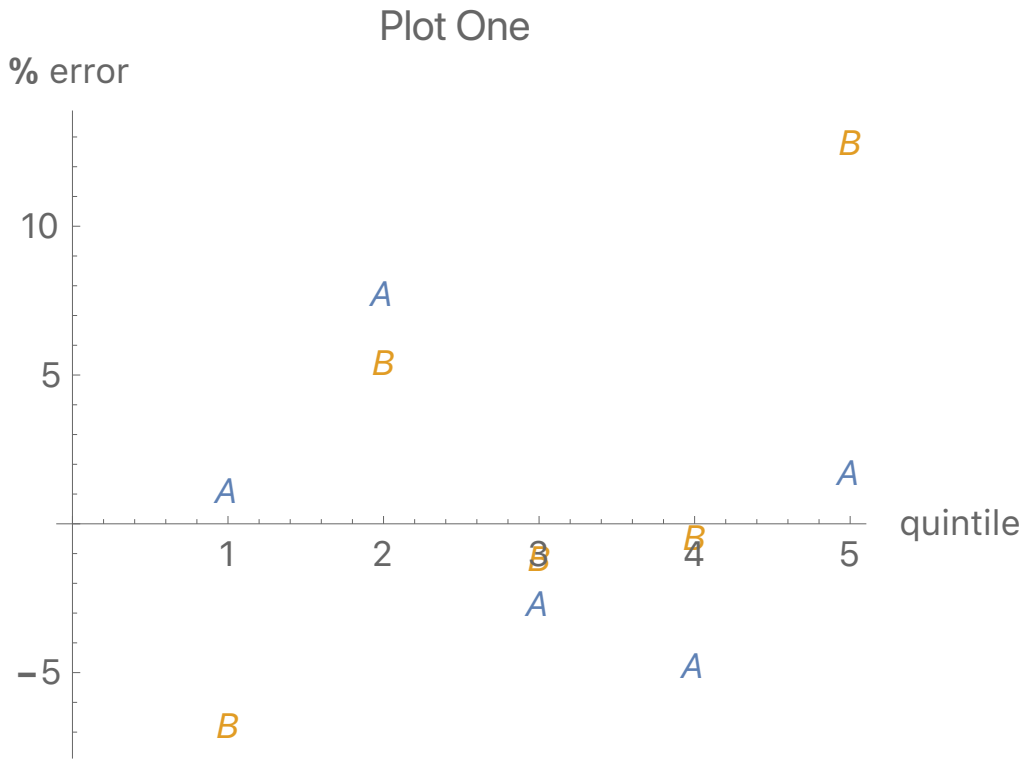
Calculate the expected increase in efficiency over this time period according to this Linear Mixed Model.

[ ] Round your answer to two decimal places.

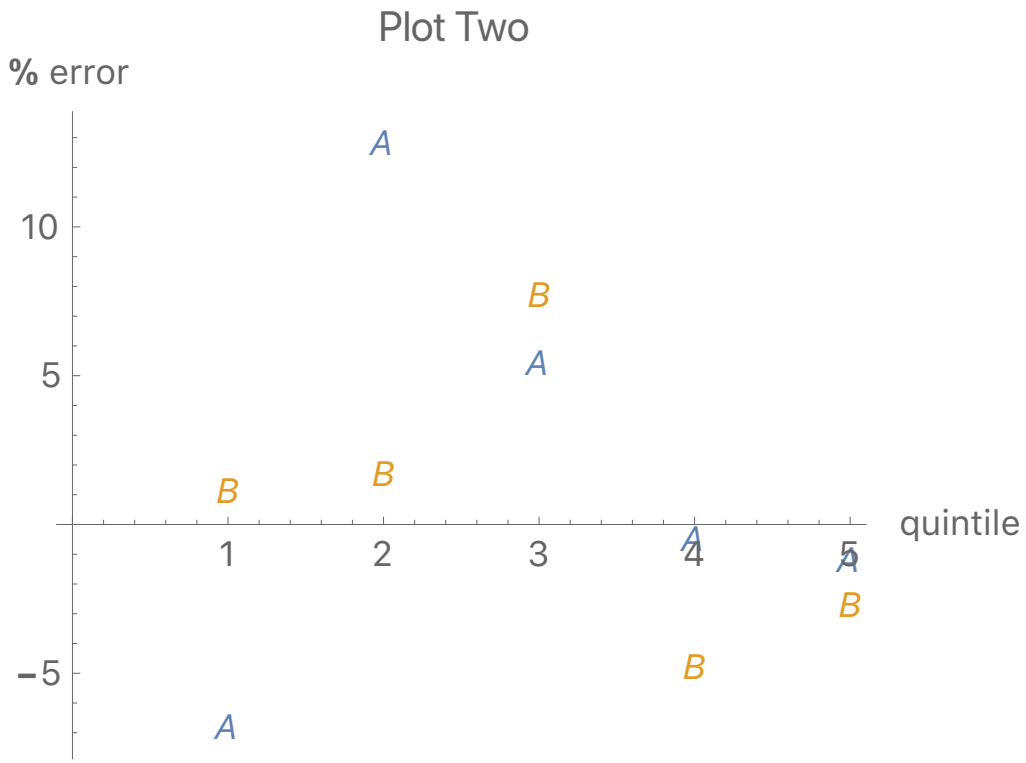
29. Two generalized linear models to predict loss costs have been built.
Output for each model are shown below:

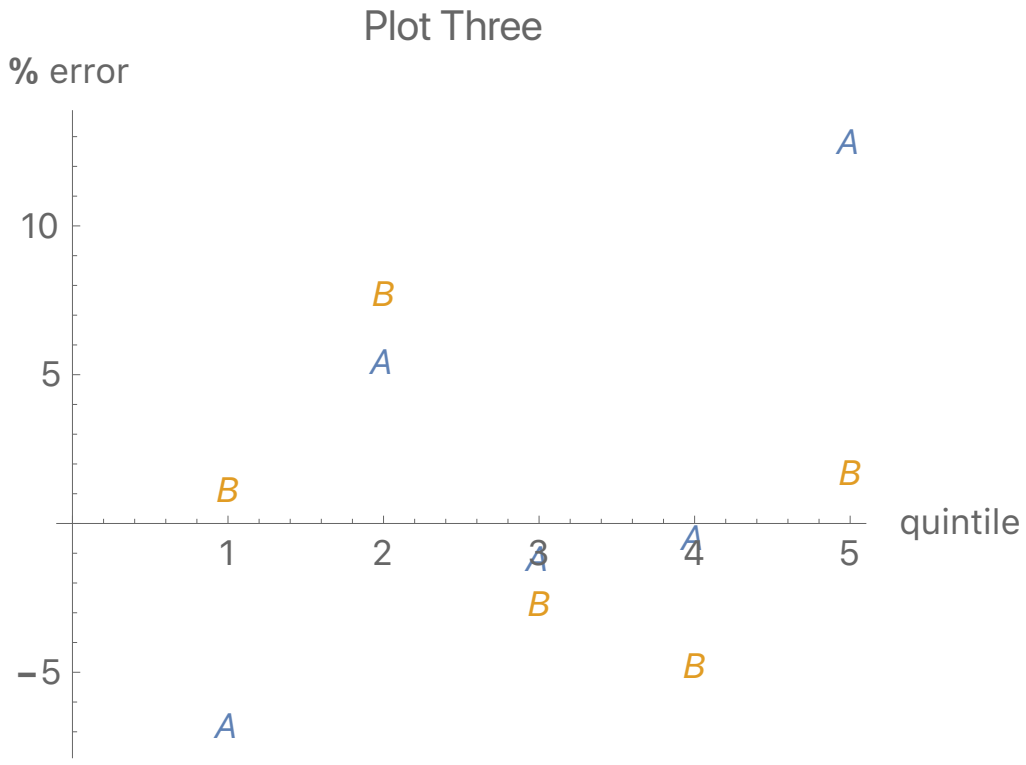
Observation	Actual Pure Premium	Model A Pure Premium	Model B Pure Premium	Ratio of Model A over Model B
1	500	530	510	1.039
2	600	620	630	0.984
3	700	750	770	0.974
4	800	940	850	1.106
5	1000	1090	980	1.112
6	1200	1140	1150	0.991
7	1400	1430	1380	1.036
8	1600	1560	1490	1.047
9	1800	1710	1780	0.961
10	2000	1830	2060	0.888

For this data, which of the following five plots is a double lift chart using quintiles? Plot # ☐

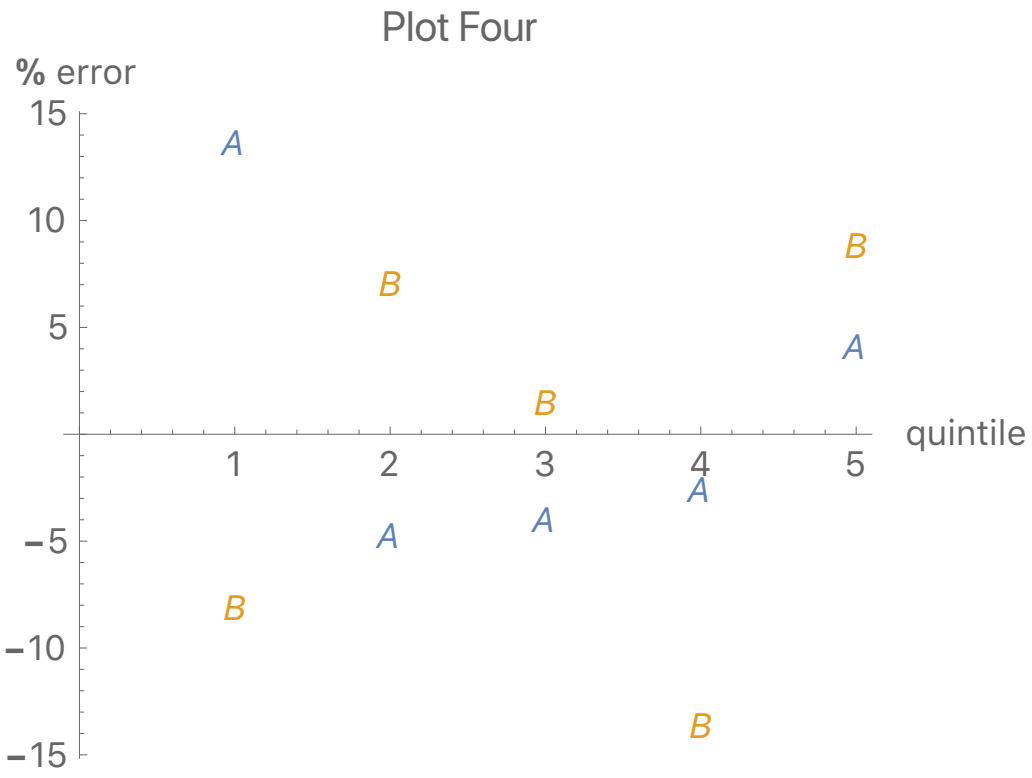


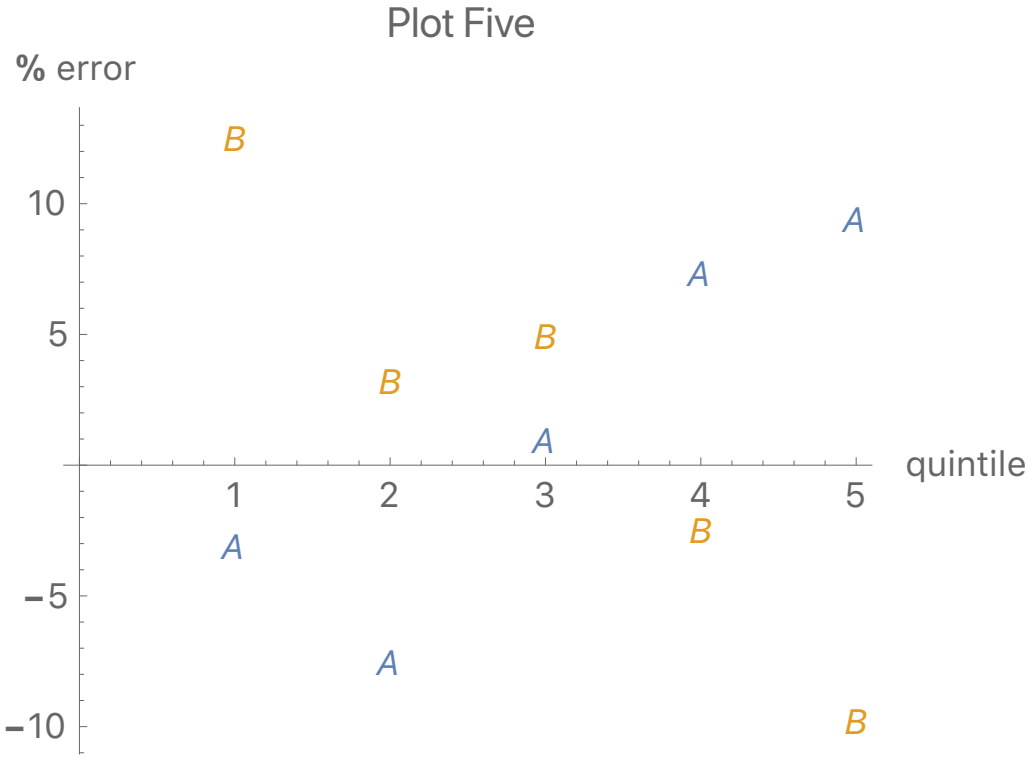
QUESTION CONTINUED ON THE NEXT PAGE





QUESTION CONTINUED ON THE NEXT PAGE


Plot Four
% error
15
10
5
−5
−10
−15
1
2
3
4
5
quintile
A
B
B
A
B
A
A
B
B

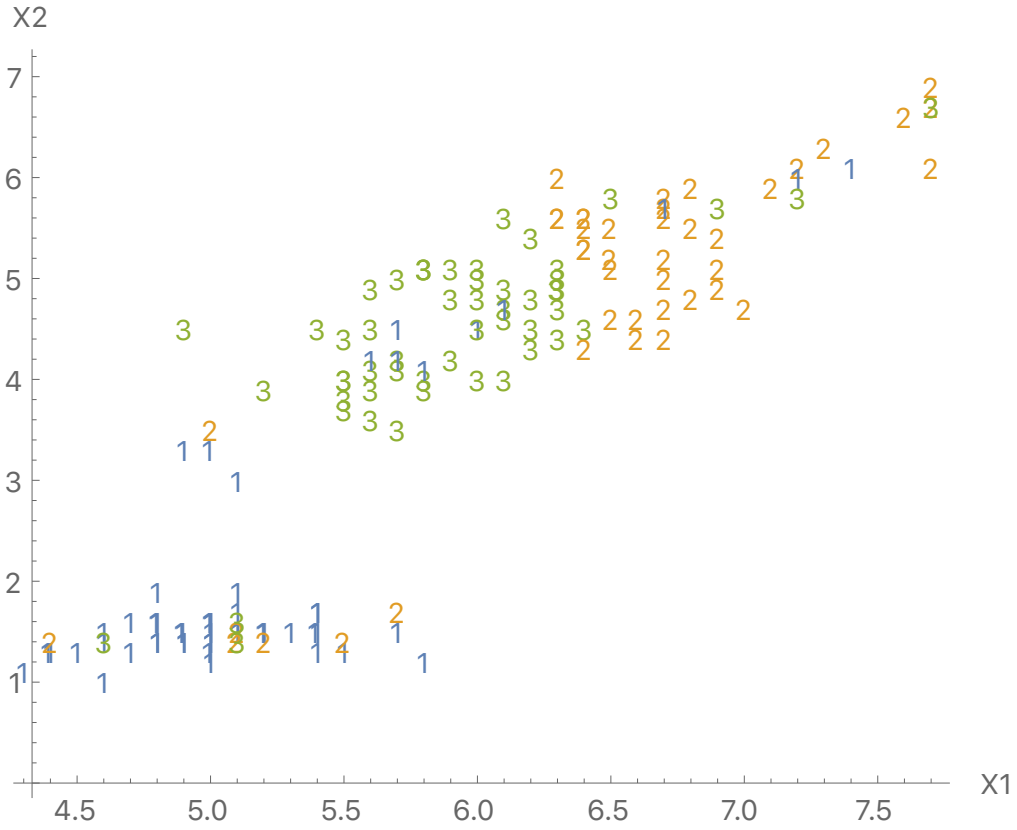

Plot Five
% error
10
5
−5
−10
1
2
3
4
5
quintile
B
A
A
B
B
A
A
B
B

30. Determine which of the following statements about Principal Component Analysis is false.
A. PCA finds a low dimension representation of a dataset that contains as much variation as possible.
B. PCA serves as a tool for data visualization.
C. It is recommend that one not individually scale the variables prior to performing PCA.
D. PCA provide low-dimensional linear surfaces that are closest to the observations.
E. A scree plot provides a method for determining the number of principal components to use.

31. For a group of policies, you are given:
(i) The annual loss on an individual policy follows a Pareto distribution with parameters $\alpha = 6$ and θ.
(ii) The prior distribution of θ has a mean of 80,000.
(iii) A randomly selected policy had losses of:
11,000 in Year 1, 18,000 in Year 2, and 9000 in Year 3.
(iv) Loss data for Year 4 was misfiled and unavailable.
(v) Based on the data in (iii), the Bühlmann credibility estimate of the loss on the selected policy in Year 6 is 15,000.
(vi) After the estimate in (v) was calculated, the data for Year 4 was located.
The loss on the selected policy in Year 4 was 7000.
Calculate the Bühlmann credibility estimate of the loss on the selected policy in Year 6 based on the data for Years 1, 2, 3, and 4.
(A) 13,900 (B) 14,100 (C) 14,300 (D) 14,500 (E) 14,700

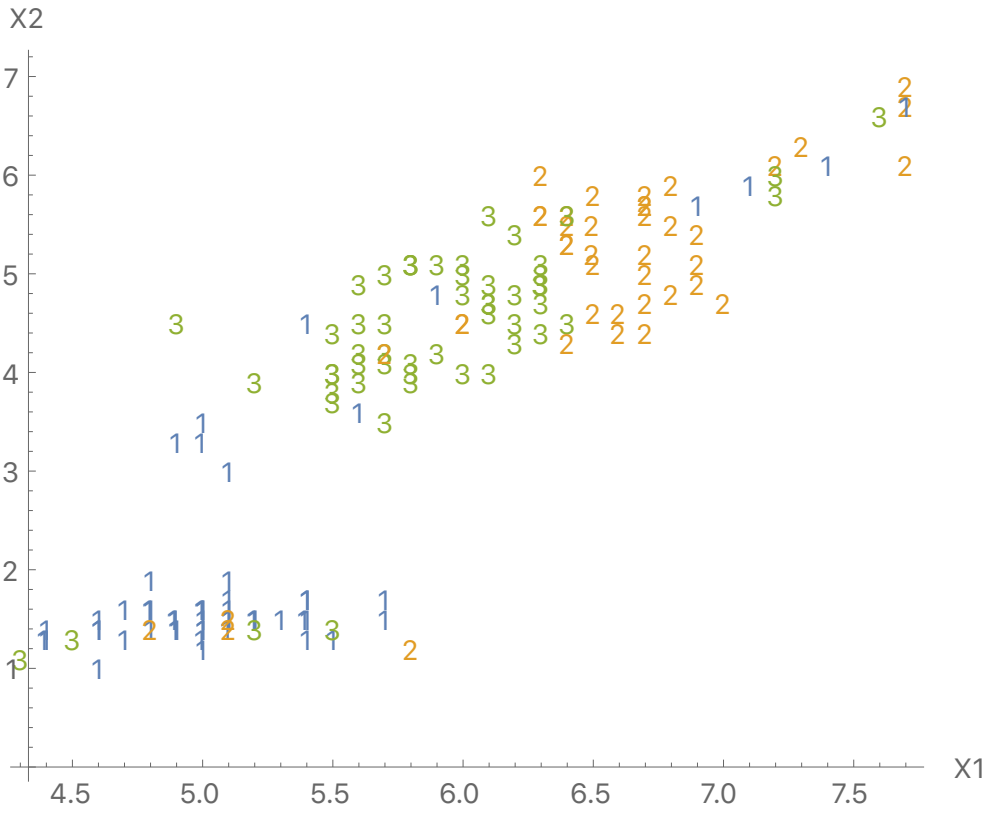
32. There are two predictors X_1 and X_2.
Data has been clustered into three clusters.
Which of the following clusterings is best?

A.

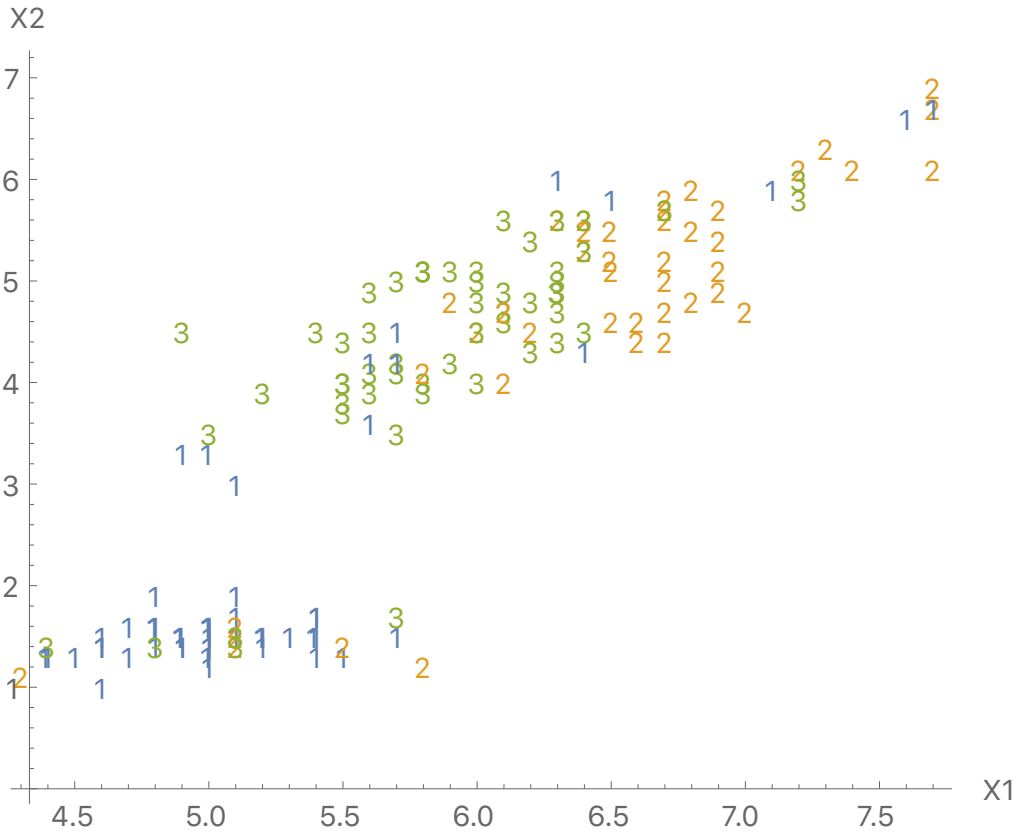


QUESTION CONTINUED ON THE NEXT PAGE

B.

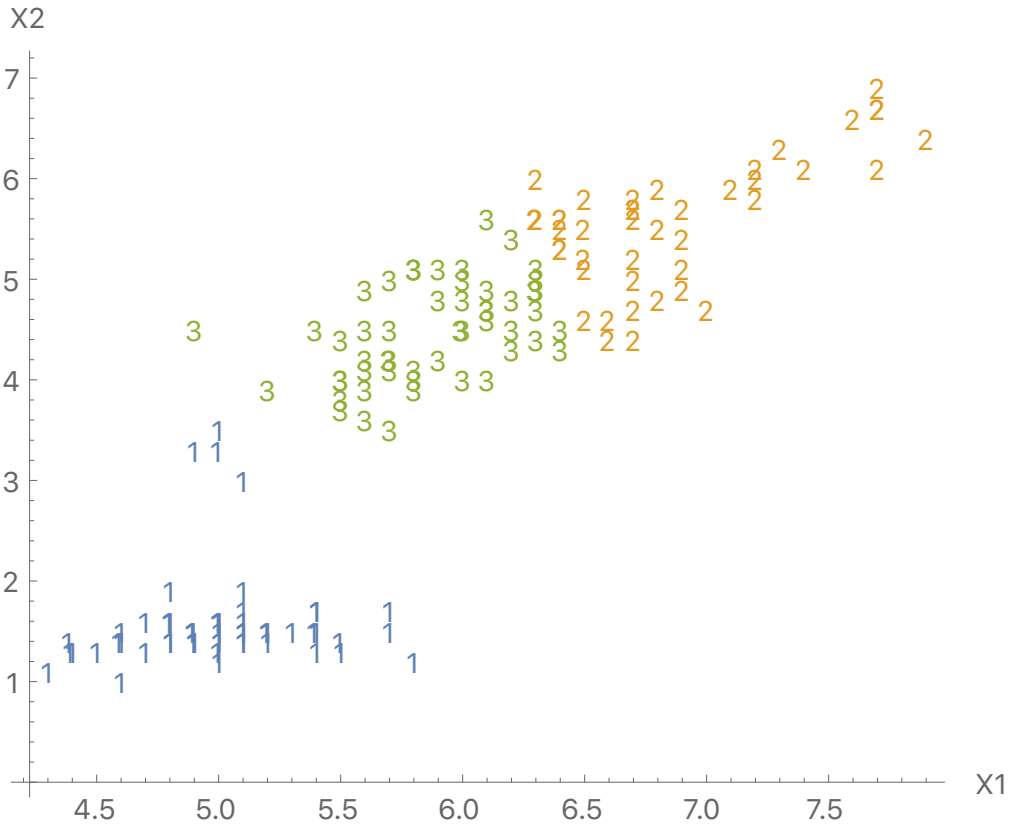


C.

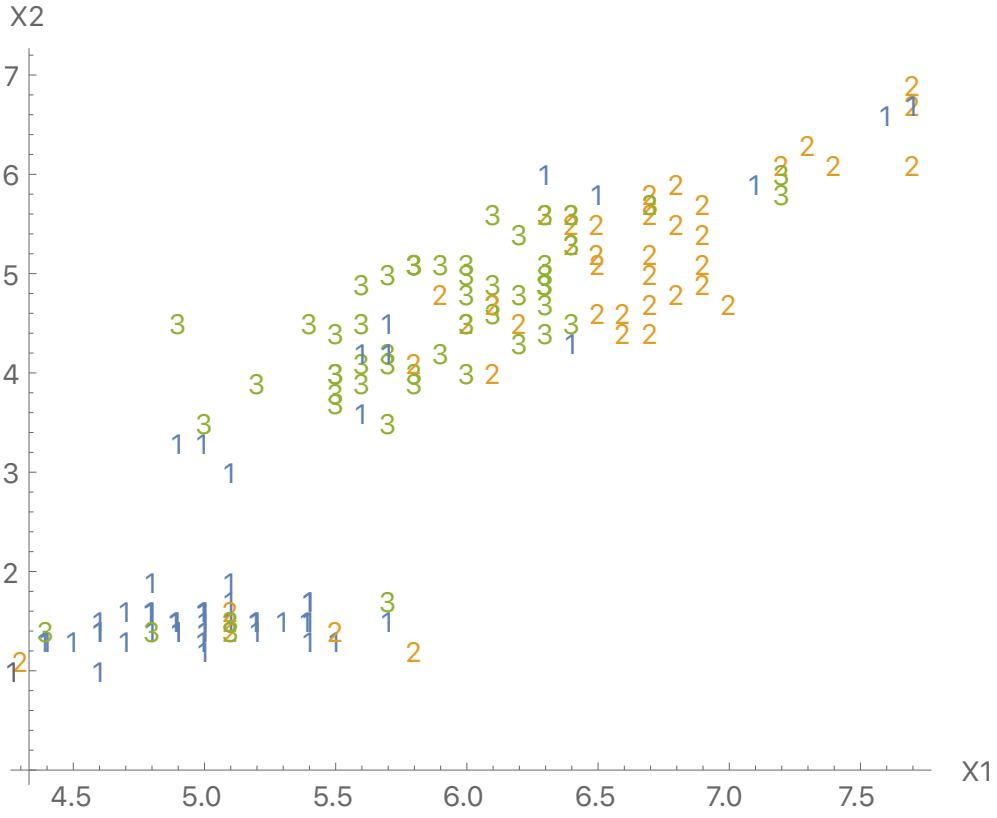


QUESTION CONTINUED ON THE NEXT PAGE

D.


X2
7
6
5
4
3
2
1
4.5
5.0
5.5
6.0
6.5
7.0
7.5
X1

E.


X2
7
6
5
4
3
2
1
4.5
5.0
5.5
6.0
6.5
7.0
7.5
X1

33. You are given:
(i) Claim counts follow a Poisson distribution with mean λ.
(ii) Claim sizes follow a lognormal distribution with parameters μ and σ.
(iii) Claim counts and claim sizes are independent.
(iv) The prior distribution has joint probability density function:

$$f(\lambda, \mu, \sigma) = \frac{0.016\,\sigma}{\lambda^2}\ , 0.03 < \lambda < 0.08,\ 7 < \mu < 9,\ 1 < \sigma < 2.$$

Calculate K, the Bühlmann Credibility Parameter for aggregate losses.
(A) Less than 400
(B) At least 400, but less than 700
(C) At least 700, but less than 1000
(D) At least 1000, but less than 1300
(E) At least 1300

34. You are provided the following data set with a single variable X:
5, 26, 38, 64, 88.
A dendrogram is built from this data set using agglomerative hierarchical clustering with complete linkage and Euclidean distance as the dissimilarity measure.
Calculate the tree height at which the observation of 5 fuses.
A. Less than 20
B. At least 20, but less than 25
C. At least 25, but less than 30
D. At least 30, but less than 35
E. At least 35

35. You are given the following three statements regarding parameter estimates of Linear Mixed Models using maximum likelihood (ML) estimation and residual maximum likelihood (REML) estimation.
I. ML takes into account the loss of degrees of freedom that results from estimating the fixed effects.
II. ML and REML estimates of the fixed effects parameters differ.
III. REML estimates of the covariance parameters are unbiased.
Determine which of the preceding statements are true.
A. None are true B. I and II only C. I and III only D. II and III only
E. The answer is not given by (A), (B), (C), or (D)

36. You are given two random walk models.
These models are identical in every respect, except that one includes a known non-zero drift parameter and the other does not include a drift parameter.
Which of the following statements are true? (select all that apply).

☐ A. For the random walk without drift, all forecasted values from time T are equal.

☐ B. For the random walk without drift, $Cov[x_6, x_{20}] = 14\ \sigma^2$.

☐ C. For the random walk with drift, the standard error of the forecast from time T increases or decreases, depending on the drift parameter, as the forecast horizon increases.

☐ D. For the random walk with drift, $Cov[x_3, x_{11}] = 3\ \sigma^2$.

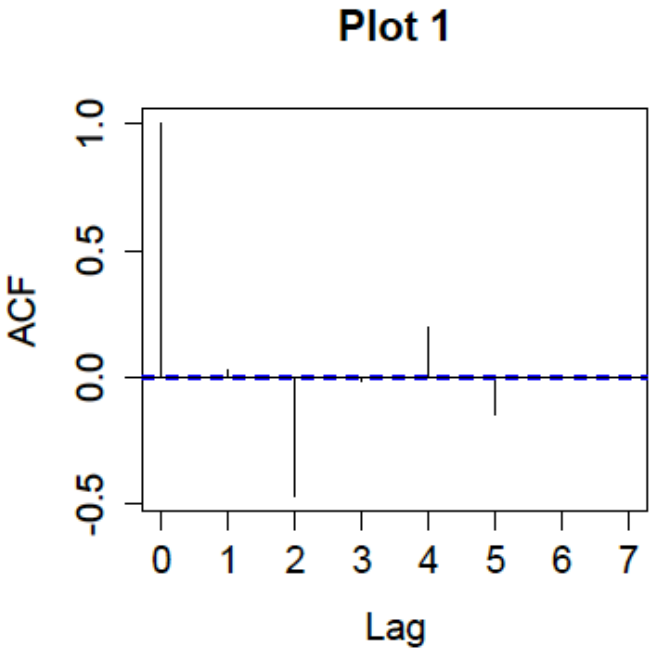
☐ E. For a random walk with drift, the first difference is a white noise series.

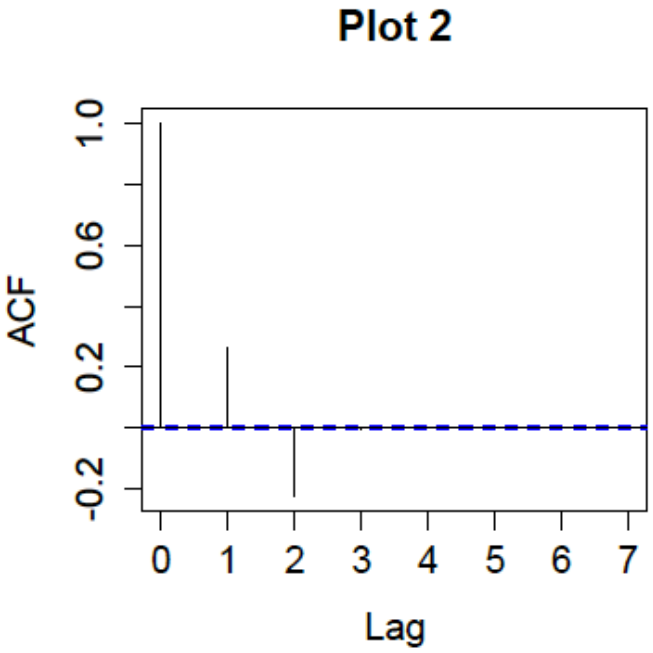
37. You are given the following information about two classes of risks:

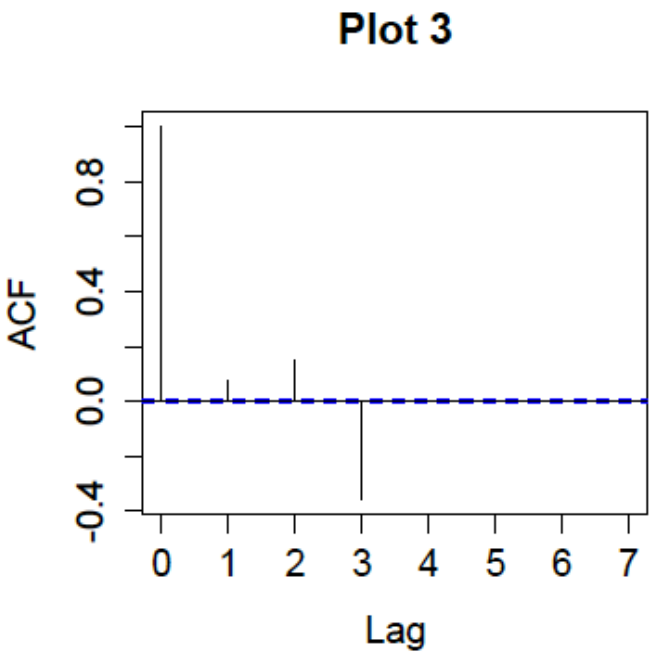
- Risks in Class A have a Binomial annual claim count distribution, with m = 5 and q = 0.3.
- Risks in Class B have a Binomial annual claim count distribution, with m = 5 and q = 0.6.
- Risks in Class A have a Single Parameter Pareto severity distribution, with $\alpha = 2$ and $\theta = 10$.
- Risks in Class B have a Single Parameter Pareto severity distribution, with $\alpha = 3$ and $\theta = 10$.
- Class A has twice the number of risks in Class B.
- Within each class, severities and claim counts are independent.
- A risk is randomly selected and observed to have two claims during one year.
- The first claim was of size 31 and the second claim was of size 17.

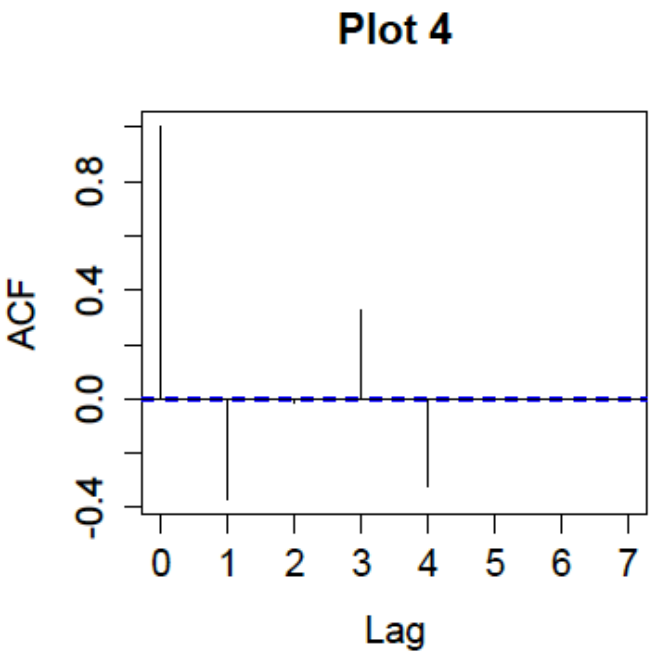
Calculate the posterior expected value of the pure premium for this risk.
(A) 31 (B) 32 (C) 33 (D) 34 (E) 35

38. You are given the following graphs of sample autocorrelations.









If these are moving average models, what order are they?

Plot 1 MA order ☐

Plot 2 MA order ☐

Plot 3 MA order ☐

Plot 4 MA order ☐

39. N has a Poisson distribution with mean 6.
Let $A = X_1 + ... + X_N$, where $Prob(X_i = 2) = 70\%$ and $Prob(X_i = 5) = 30\%$, for all i,
where the X_i's are independent.
Let $B = Y_1 + ... + Y_N$, where $Prob(Y_i = 3) = 50\%$ and $Prob(Y_i = 7) = 50\%$, for all i,
where the Y_i's are independent.
The X_i's are independent of the Y_i's.
Calculate the correlation coefficient between A and B.
(A) 0.82 (B) 0.84 (C) 0.86 (D) 0.88 (E) 0.90

40. Data was collected from a large sample of law schools.
SALARY is the starting salary of graduate i of law school j.
TUITION is the annual cost of attending law school j.
PUBLIC is whether law school j is public (1) as opposed to private (0).
LSAT is the score on the Law School Admission Test of graduate i of law school j.
AGE is the age at graduation of graduate i of law school j.
Determine which of the following equations specifies the Linear Mixed Model for a graduate i of law school j with the greatest number of random effects.

A. $\ln[SALARY_{ij}] = \beta_0 + \beta_1 LSAT_{ij} + \beta_2 AGE_{ij} + \beta_3 TUITION_j + \beta_4 PUBLIC_j + u_{0j} + \varepsilon_{ij}$

B. $\ln[SALARY_{ij}] = \beta_0 + \beta_1 LSAT_{ij} + \beta_2 AGE_{ij} + \beta_3 TUITION_j + \beta_4 PUBLIC_j + u_{0j} + u_{1j} LSAT_{ij} + \varepsilon_{ij}$

C. $\ln[SALARY_{ij}] = \beta_0 + \beta_1 LSAT_{ij} + \beta_2 AGE_{ij} + \beta_3 TUITION_j + \beta_4 PUBLIC_j + u_{0j} + u_{1j} AGE_{ij} + \varepsilon_{ij}$

D. $\ln[SALARY_{ij}] = \beta_0 + \beta_1 LSAT_{ij} + \beta_2 AGE_{ij} + \beta_3 TUITION_j + \beta_4 PUBLIC_j + u_{0j} + u_{1j} LSAT_{ij} + u_{2j} AGE_{ij} + \varepsilon_{ij}$

E. $\ln[SALARY_{ij}] = \beta_0 + \beta_1 LSAT_{ij} + \beta_2 AGE_{ij} + \beta_3 TUITION_j + \beta_4 PUBLIC_j + u_{0j} + u_{1j} LSAT_{ij} + u_{2j} AGE_{ij} + u_{3j} LSAT_{ij} \times PUBLIC_j + \varepsilon_{ij}$

41. Use the following information:

- Frequency for an individual is a 90-10 mixture of two Poissons with means λ and 4λ.
- The prior distribution of λ is Gamma with $\alpha = 5$ and $\theta = 0.01$.

An insured is chosen at random and observed to have one claim in the first year.
Use Bayes Theorem in order to estimate the expected number of claims next year for the same insured.
A. Less than 0.08
B. At least 0.08, but less than 0.09
C. At least 0.09, but less than 0.10
D. At least 0.10, but less than 0.11
E. At least 0.11

42. You are fitting a Linear Mixed Model and considering the following two models.

Model Specification I:

$Y_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + u_{0j} + u_{1j} X_{ij}^{(2)} + \varepsilon_{ij}$

u_{0j} and u_{1j} follow a Bivariate Normal Distribution with means of 0 and covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$.

ε_{ij} follows a normal distribution with mean 0 and variance σ^2

All ε_{ij} are independent

Model Specification II:

$Y_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + u_j + \varepsilon_{ij}$

u_j follows a normal distribution with mean 0 and variance σ_1^2

ε_{ij} follows a normal distribution with mean 0 and variance σ^2

All ε_{ij} are independent

	Restricted Maximum Likelihood	Maximum Likelihood
Model I	-569.398	-559.839
Model II	-572.413	-564.163

What is the result of a likelihood ratio test comparing the two models?

A. Reject H_0 at 0.005.

B. Do not reject H_0 at 0.005. Reject H_0 at 0.010.

C. Do not reject H_0 at 0.010. Reject H_0 at 0.025.

D. Do not reject H_0 at 0.025. Reject H_0 at 0.050.

E. Do not reject H_0 at 0.050.

END OF PRACTICE EXAM

This page intentionally left blank.

Solutions:

1. $A_1^{(1)} = (-1 + 5x)_+$. $A_2^{(1)} = (4 - 2x)_+$.

$A_1^{(2)} = (3 - 4 A_1^{(1)} + 2 A_2^{(1)})_+$. $A_2^{(2)} = (-2 + 6 A_1^{(1)} + 3 A_2^{(1)})_+$.

$Y = 10 + 2 A_1^{(2)} + A_2^{(2)}$.

For X = 0.5: $A_1^{(1)} = (1.5)_+ = 1.5$, and $A_2^{(1)} = (3)_+ = 3$.

$A_1^{(2)} = (3)_+ = 3$, and $A_2^{(2)} = (16)_+ = 16$.

Y = 10 + (2)(3) + 16 = **32**.
Comment: Similar to Q.14.1 in "Mahler's Guide to Advanced Statistical Learning."

2. E. The parameters of the posterior Gamma are $\alpha' = \alpha + C = 0.8 + 6 = 6.8$,
and $1/\theta' = 1/\theta + E = 40 + 100 = 140$.
The posterior mean is: 6.8/140 = 0.0486.
Therefore, the expected number of mistakes that Marv will make in his next 100 hours is:
(0.0486)(100) = **4.86**.
Alternately, $K = 1/\theta = 40$. $Z = 100 / (100 + 40) = 71.4\%$.
Prior mean is 0.02. Observed mean frequency is: 6/100 = 0.06.
The estimated future frequency = (71.4%)(0.06) + (28.6%)(0.02) = 0.0486.
(0.0486)(100) = **4.86**.
Comment: Similar to Q. 4.131 (4B, 5/99, Q. 24) in "Mahler's Guide to Conjugate Priors."

3. A. Mean is: 208/8 = 26.

$$c_2 = \frac{(23-26)(22-26)+(25-26)(26-26)+(22-26)(27-26)+(26-26)(26-26)+(27-26)(30-26)+(26-26)(29-26)}{8}$$

= **1.5**.
Comment: Similar to Q.3.16 (CAS S, 5/16, Q.43) in "Mahler's Guide to Time Series."

4. The predictions in Model 1 are closer to the actual than in Model 3.
Model 2 is biased; the overwhelming majority of predictions are less than the actuals.
Model 4 is biased; the overwhelming majority of predictions are more than the actuals.
I prefer **Model 1.**
Comment: Similar to Q.2.1 in "Mahler's Guide to Advanced GLMs".
See Figure 21 in Generalized Linear Models for Insurance Rating.

5. E. All of the statements are true.
Comment: Similar to Q.12.5 in "Mahler's Guide to Linear Mixed Models."
See Figure 6.1 in Linear Mixed Models.
SICD group #1 displays less variability between children than the other two groups.

6. C. We choose the split with the smallest Entropy.
A. Contribution from the first region: - {(45/60)ln(45/60) + (15/60)ln(15/60)} = 0.562335.
Contribution from the second region: - {(15/40)ln(15/40) + (25/40)ln(25/40)} = 0.661563.
Weighting together the contributions: (60/100)(0.562335) + (40/100)(0.661563) = 0.602026.
B. Contribution from the first region: - {(40/50)ln(40/50) + (10/50)ln(10/50)} = 0.500402.
Contribution from the second region: - {(20/50)ln(20/50) + (30/50)ln(30/50)} = 0.673012.
Weighting together the contributions: (50/100)(0.500402) + (50/100)(0.673012) = 0.586707.
C. Contribution from the first region: - {(43/54)ln(43/54) + (11/54)ln(11/54)} = 0.505494.
Contribution from the second region: - {(17/46)ln(17/46) + (29/46)ln(29/46)} = 0.658724.
Weighting together the contributions: (54/100)(0.505494) + (46/100)(0.658724) = **0.57598**.
D. Contribution from the first region: - {(50/67)ln(50/67) + (17/67)ln(17/67)} = 0.566397.
Contribution from the second region: - {(10/33)ln(10/33) + (23/33)ln(23/33)} = 0.61341.
Weighting together the contributions: (67/100)(0.566397) + (33/100)(0.61341) = 0.581911.
E. Contribution from the first region: - {(35/43)ln(35/43) + (8/43)ln(8/43)} = 0.480439.
Contribution from the second region: - {(25/57)ln(25/57) + (32/57)ln(32/57)} = 0.685587
Weighting together the contributions: (43/100)(0.480439) + (57/100)(0.685587) = 0.597373.
Comment: Similar to Q.4.9 in "Mahler's Guide to Advanced Statistical Learning."
The Entropy for the m^{th} region is: $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk})$.
I have used the natural log; which base you use for the log should not change your final answer.

7. B. $\nabla x_3 = 2.5. \Rightarrow x_3 - x_2 = 2.5. \Rightarrow x_2 = x_3 - 2.5 = 27.7 - 2.5 = 25.2.$

$\nabla x_2 = -11.4. \Rightarrow x_2 - x_1 = -11.4. \Rightarrow x_1 = x_2 + 11.4 = 25.2 + 11.4 = 36.6.$

$x_1 = \alpha_0 + \alpha_1 + z_1. \Rightarrow 36.6 = \alpha_0 + \alpha_1 - 5.1. \Rightarrow 41.7 = \alpha_0 + \alpha_1.$

$x_3 = \alpha_0 + 3\alpha_1 + z_3. \Rightarrow 27.7 = \alpha_0 + 3\alpha_1 + 7.2. \Rightarrow 20.5 = \alpha_0 + 3\alpha_1.$

Subtracting the two equations: $-21.2 = 2\alpha_1. \Rightarrow \alpha_1 = -10.6. \Rightarrow \alpha_0 = 52.3.$

$\hat{x}_6 = \alpha_0 + 6\alpha_1 = 52.3 + (6)(-10.6) =$ **-11.3**.

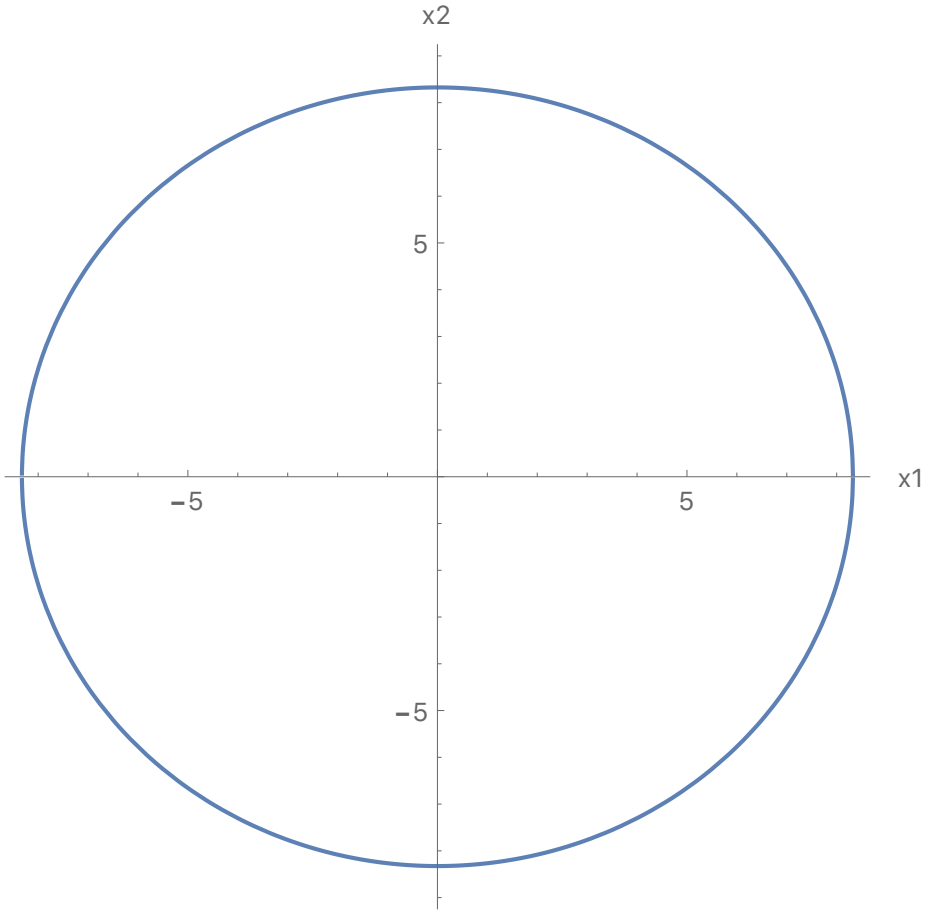
Comment: Similar to Q.16.22 (CAS S, 5/17, Q.45) in "Mahler's Guide to Time Series."
The difference operator is defined as: $\nabla x_t = x_t - x_{t-1}$.
z_t is a white noise series. $\Rightarrow$

z_6 is independent of the previous observed values of z_t, and $E[z_6] = 0$.

8. C.

i	x_1	x_2	$\exp[-(x_1^2 + x_2^2)/100]$
1	2	-6	0.6703
2	5	0	0.7788
3	7	6	0.4274
4	13	-10	0.0679
5	-2	9	0.4274
6	-5	-4	0.6637

Using the Bayes Classifier we assign the observation to whatever class is most likely.
Therefore, in each case the chance of an error is: 1 - Max[Pr[Y = F], Pr[Y = T]].
The Bayes error rate is the expected value of an error using the Bayes Classifier:
{(1 - 0.6703) + (1 - 0.7788) + 0.4274 + 0.0679 + 0.4274 + (1 - 0.6637)} / 6 = **0.302**.
Comment: Similar to Q.2.1 in "Mahler's Guide to Advanced Statistical Learning."
We make no use of the observed values in the test data of true/false.
Along the following circle, Prob[Y = F] = 0.5. Inside the circle, Prob[Y = F] > 0.5.
Outside the circle, Prob[Y = F] < 0.5. The circle is the Bayes Decision Boundary:



9. A. P = 99%. Therefore, y = 2.576, since $\Phi(2.576) = 0.995 = (1+P)/2$. k = 0.10.

Standard For Full Credibility is: $(y / k)^2 (\sigma_f^2/\mu_f) = (2.576/0.1)^2(0.12/0.08) = 995$ claims,

or 995/0.08 = 12,438 exposures.

$Z = \sqrt{1000/12,438} = 28.4\%$.

Estimated future frequency is: (28.4%)(112/1000) + (71.6%)(.08) = 8.91%.
Expected number of future claims is: (1000)(8.91%) = **89**.
Alternately, using the expected number of claims of (0.08)(1000) = 80, and the standard for full credibility in terms of expected claims: $Z = \sqrt{80/995} = 28.4\%$. Proceed as before.

Comment: Similar to Q. 6.7 in "Mahler's Guide to Classical Credibility."
When available one generally uses the number of exposures (1000) or the expected number of claims (80) in the square root rule, rather than the observed number of claims (112), since the observed number of claims is subject to random fluctuation.

10. C. In Seq2Seq learning, both the input sequence and the target sequence are represented by a structure similar to a simple recurrent neural network, and they share the hidden units.
Statement I is not true.
In long term and short term memory (LSTM), two tracks of hidden-layer activations are maintained, so that when the activation is computed, it gets input from hidden units both further back in time, and closer in time. Statement II is not true.
Statement III is true.
Comment: Similar to Q.17.1 in "Mahler's Guide to Advanced Statistical Learning."
The positions of words in the embedding space preserve semantic meaning; in other words, synonyms should appear near each other.

11. E. The process variance for an individual insured is: $r\beta(1+\beta) = (0.7)(1.7)r = 1.19r$.

EPV = $E[1.19r] = 1.19E[r] = 1.19\alpha\theta$.

The mean frequency for an individual insured is: $r\beta = 0.7r$.

VHM = $Var[0.7r] = 0.7^2 Var[r] = 0.49\alpha\theta^2$.

K = EPV/VHM = $1.19\alpha\theta / (0.49\alpha\theta^2) =$ **2.43/**θ.
Comment: Similar to Q. 10.65 (4, 5/00, Q.37) in "Mahler's Guide to Buhlmann Credibility."

12. C. The RSS for various amounts of pruning:
No nodes pruned: 40 + 33 + 67 + 52 + 49 = 241.
Prune both nodes Y and Z: 91 + 67 + 52 + 49 = 259.
Prune both nodes W and X: 40 + 33 + 67 + 114 = 254.
Prune both nodes U and V: 193 + 52 + 49 = 294
Prune both nodes S and T (resulting in the null model): 345.
We score the original tree and each possible subtree using:
RSS + $\alpha\ |T|$ = RSS + 20 $|T|$, where $|T|$ is the number of terminal nodes.

Nodes Pruned	RSS	Terminal Nodes	Tree Score
None	241	5	341
Y and Z	259	4	339
W and X	254	4	**334**
U and V	294	3	354
S and T	345	1	365

The **third** strategy produces the lowest tree score and thus is selected.
Comment: Similar to Q.3.10 in "Mahler's Guide to Advanced Statistical Learning."
The term $\alpha\ |T|$ in the tree score penalizes a tree for being complex.

In general, a larger value of α would lead to simpler trees being selected.

13. A. We have y = 2.326 since $\Phi(2.326) = 0.99$.
Therefore, $n_0 = y^2/k^2 = (2.326/0.05)^2 = 2164$.
standard for full credibility = $n_F = n_0\ (1+CV^2)$.
Therefore CV = $\sqrt{(n_F / n_0) - 1}\ = \sqrt{(5000 / 2164) - 1}\ =$ **1.14**.
Comment: Similar to Q. 5.4 in "Mahler's Guide to Classical Credibility."

14. C. For an AR(4), the 4th partial autocorrelation is non-zero, but partial autocorrelations of order 5 or greater are zero.
Graph C has a 4th sample partial autocorrelation significantly different than zero, while the 5th and greater sample partial autocorrelations are not significantly different than zero
Comment: Similar to Q.7.14 in "Mahler's Guide to Time Series."

15. C. The posterior distribution is a Beta Distribution with
a' = a + number of claims = 3 + 5 = 8,
b' = b + number of years - number of claims = 14 + 20 - 5 = 29, and $\theta = 1$.
This is: $\frac{\Gamma(8+29)}{\Gamma(8)\ \Gamma(29)}\ q^{8-1}(1-q)^{29-1} = \frac{36!}{(7!)\ (28!)}\ q^7(1-q)^{28} =$ **242,082,720 $q^7(1-q)^{28}$**.
Comment: Similar to Q. 6.8 in "Mahler's Guide to Conjugate Priors."

16. B. Adding the probabilities, there is a 60% a priori probability of $\Theta = 1$, risk type A, and a 40% a priori probability of $\Theta = 2$, risk type B.
For risk type A, the distribution of X is:
100 @ 0.3/0.6 = 1/2, 200 @ 0.2/0.6 = 1/3, and 500 @ 0.1/0.6 = 1/6.
The mean for risk type A is: $E[X \mid \Theta = 1] = (100)(1/2) + (200)(1/3) + (500)(1/6) = 200$.
The 2nd moment for risk type A is: $E[X^2 \mid \Theta = 1] = (100^2)(1/2) + (200^2)(1/3) + (500^2)(1/6) = 60{,}000$. Process Variance for risk type A is: $Var[X \mid \Theta = 1] = 60{,}000 - 200^2 = 20{,}000$.
Similarly, risk type B has mean 325, second moment 137,500, and process variance 31,875.

Risk Type	A Priori Chance	Mean	Square of Mean	Process Variance
A	0.6	200	40,000	20,000
B	0.4	325	105,625	31,875
Average		250	66,250	24,750

Variance of the hypothetical means = $66{,}250 - 250^2 = 3750$.
K = EPV/VHM = 24,750/3750 = 6.6.
Z = 3/(3 + K) = 31.3%. Observed mean is: 900/3 = 300. Prior mean is 250.
Estimate is: (0.313)(300) + (0.687)(250) = **266**.
Comment: Similar to Q. 9.73 (4, 11/02, Q.29)
in "Mahler's Guide to Buhlmann Credibility."

17. A. The three tuning parameters are: B the number of trees, λ the shrinkage parameter, and d the splits in each tree. Statement I is true.
Random Forests makes predictions from an average of regression trees, each of which is built using a random sample of data and predictors. Statement II is not true.
In boosting, often using trees with a single split works well. Statement III is not true.
Comment: Similar to Q.7.2 in "Mahler's Guide to Advanced Statistical Learning."
In boosting the trees are grown sequentially; each tree is grown using information from previously grown trees.

18. B. Chance of observing 2 claims is $\lambda^2 e^{-\lambda} / 2$.

Type	A Priori Probability	Poisson Parameter	Chance of 2 Claims	Probability Weights	Posterior Probability	Mean
A	0.7	0.100	0.00452	0.00317	0.3920	0.100
B	0.3	0.200	0.01637	0.00491	0.6080	0.200
Sum	1	0.13		0.00808	1.0000	**0.161**

Comment: Similar to Q. 2.3 in "Mahler's Guide to Conjugate Priors."

19. C. We determine $\hat{p}_{mk}$, the proportion of training observations in the m^{th} region that are from the k^{th} class.
Of the 3 observations with $X_2 < 18$ and $X_1 = Y$, all 3 have Z = F.
Of the 2 observations with $X_2 < 18$ and $X_1 = N$, one has Z = T and one has Z = F.
Of the 5 observations with $X_2 \geq 18$, 3 have Z = T and two has Z = F.

	Observations	Gini Index	Entropy
$X_2 < 18$ and $X_1 = Y$	3	(1)(1 - 1) + (0)(1 - 0)	0
$X_2 < 18$ and $X_1 = N$	2	(1/2)(1 - 1/2) + (1/2)(1 - 1/2)	-(1/2)ln(1/2) - (1/2)ln(1/2)
$X_2 \geq 18$	5	(3/5)(1 - 3/5) + (2/5)(1 - 2/5)	-(0.6)ln(0.6) - (0.4)ln(0.4)

The number of observations act as the weights.
G = a weighted average of the Gini Indices for every terminal node =
(3/10)(0) + (2/10)(1/2) + (5/10)(0.48) = 0.34.
D = a weighted average of the Entropies for every terminal node =
(3/10)(0) + (2/10)(0.6931) + (5/10)(0.6730) = 0.4751.
D - G = 0.4751 - 0.34 = **0.1351**.
Comment: Similar to Q.4.6 in "Mahler's Guide to Advanced Statistical Learning."

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}).$$

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log(\hat{p}_{mk}).$$ If $\hat{p}_{mk} = 0$, then the contribution to the entropy is taken as zero.

20. D. AIC = (-2) (loglikelihood) + (2) (number of parameters).

Model	Number of parameters	Loglikelihood	AIC
AR(1)	3	-155.85	317.70
AR(2)	4	-147.96	303.92
ARMA(2,1)	5	-147.84	305.68

Best AIC is smallest; the AR(2) model is best.
A 95% confidence interval for α_1 is: 0.4548 ± (1.96)(0.0931) = (0.27, **0.64**).
Comment: Similar to Q.11.8 in "Mahler's Guide to Time Series."
I have assumed an intercept and have included sigma in the fitted parameters;
as long as one is consistent, this will not affect which model has the best AIC.

21. D. The first-order autoregressive structure has the variances along the main diagonal and the covariances decline geometrically as they get further from the main diagonal.
In Matrix D, the covariances decline by a factor of 0.5.
Comment: Similar to Q.4.20 (MAS-2, 11/18, Q.7) in "Mahler's Guide to Linear Mixed Models."

22. D & E. Gini Index = $\frac{\text{Area W}}{\text{Area W + Area V}}$.

However, Area W + Area V = 1/2. Thus the Gini Index = 2 W.
Comment: Similar to Q.6.1 in "Mahler's Guide to Advanced GLMs".
The Gini Index is also: 1 - 2V.

23. E. The scores for the first principal component are:
(-0.4438)(-0.5) + (0.4438)(0.5) + (0.7785)(1.225) = 1.397.
(-0.4438)(-0.5) + (0.4438)(-2) + (0.7785)(-0.817) = -1.302.
(-0.4438)(-0.5) + (0.4438)(0.5) + (0.7785)(-0.817) = -0.192.
(-0.4438)(-0.5) + (0.4438)(0.5) + (0.7785)(1.225) = 1.397.
(-0.4438)(2) + (0.4438)(0.5) + (0.7785)(-0.817) = -1.302.
Z_1 = (1.397, -1.302, -0.192, 1.397, -1.302).
Since Z_1 has a mean of zero (subject to rounding):
$\text{Var}[Z_1] = \{1.397^2 + (-1.302)^2 + (-0.192)^2 + 1.397^2 + (-1.302)^2\}/5 = 1.466$.
The vectors X_1, X_2, X_3 have each been scaled so that each has a variance of 1.
Thus the proportion of variance explained by the first principal component is:
1.466 / (1 + 1 + 1) = **48.9%**.
Comment: Similar to Q.9.1 in "Mahler's Guide to Advanced Statistical Learning."
Alternately, if one instead computes the unbiased variances by using 4 in the denominator rather than 5, one would obtain the same final answer.
$\{1.397^2 + (-1.302)^2 + (-0.192)^2 + 1.397^2 + (-1.302)^2\}/4 = 1.833$.
$\{(-0.5)^2 + (-0.5)^2 + (-0.5)^2 + (-0.5)^2 + 2^2\}/4 = 1.25$.
$\{0.5^2 + (-2)^2 + 0.5^2 + 0.5^2 + 0.5^2\}/4 = 1.25$.
$\{1.225^2 + (-0.817)^2 + (-0.817)^2 + 1.225^2 + (-0.817)^2\}/4 = 1.251$.
1.833/(1.25 + 1.25 + 1.251) = 48.9%.

24. E. Model 1: $\hat{x}_7 = x_6 = 14$.

Model 2: $\alpha = 1/1.5$. $\hat{x}_7 = (1/1.5)\, x_6 = 14/1.5 = 9.333$.

Model 3: $\alpha = 1/4$. $\hat{x}_7 = (1/4)\, x_6 = 14/4 = 3.5$.

Model 4: $\alpha \geq 1$. $\hat{x}_7 = \alpha\, x_6 \geq 14$.

The largest predicted value is for either Model 4 or is a tie between Models 1 and 4.
Therefore, E. Is the best answer.
Comment: Similar to Q.8.18 (MAS-1, 11/19, Q.42) in "Mahler's Guide to Time Series."
For an AR(1), the characteristic equation is: $1 - \alpha B = 0$, with root $1/\alpha$.
An autoregressive process is stationary if and only if all of the roots of its characteristic equation exceed one.
Thus a non-stationary autoregressive process of order one has $|\alpha| \geq 1$.

For a random walk, the predicted future value is equal to the last observed value.

25. A. X_{it} = pure premiums = (total losses) / (number in group).

Group	Pure Premium		$XBar_i$	v_i
1	800	600	702.56	3,897,436
2	500	600	551.61	774,194
			Average	2,335,815
	Exposures		Total	
1	200	190	390	
2	150	160	310	

$v_i = \frac{1}{Y - 1} \sum_{t=1}^{Y} m_{it} (X_{it} - \bar{X}_i)^2$ = estimated process variance for group i.

Estimated EPV = (1/C) $\sum v_i$ = (3,897,436 + 774,194)/2 = 2,335,815.

The estimated VHM is given as 4631. K = EPV/VHM = 2,335,815/4631 = 504.

Group 2 has 310 exposures, so its data is given credibility of: 310 / (310 + 504) = **38.1%**.

Comment: Similar to Q. 4.38 (4, 5/05, Q.25) in "Mahler's Guide to Nonparametric Credibility."

Even though there are missing years of data, each group has two years of data.

Therefore, there is no need to use the formulas involving differing numbers of years of data.

Let $\Pi = m - \sum m_i^2 / m = 700 - (390^2 + 310^2)/700 = 345.43$.

$\bar{X}$ = overall average loss per exposure = 445,000/700 = 635.71.

Estimated VHM = $\frac{\sum_{i=1}^{C} m_i (\bar{X}_i - \bar{X})^2 - EPV (C - 1)}{\Pi}$ =

$\{390(702.56 - 635.71)^2 + 310(551.61 - 635.71)^2 - (2-1)(2,335,815)\} / 345.43 = 4631$, as given.

26. A. Statement III is not practical because we generally do not have access to multiple training sets. Instead, we can bootstrap, by taking repeated samples from the (single) training data set.

Comment: Similar to Q.5.1 in "Mahler's Guide to Advanced Statistical Learning."

In bagging regression trees, the importance of each predictor can be measured by the amount that splits over that predictor reduce the RSS, averaged over the different trees.

27. C. This is an Gamma-Exponential with $\alpha = 3$ and $\theta = 1/100$.

The posterior Inverse Gamma has parameters:

$\alpha' = \alpha + C = 3 + 4 = 7$. $1/\theta' = 1/\theta + L = 100 + 140 = 240$.

The predictive distribution is a Pareto with $\alpha = 7$ and $\theta = 240$.

$F(x) = 1 - \{\theta/(\theta + x)\}^{\alpha}$. $S(60) = (240/300)^7 =$ **0.2097**.

Comment: Similar to Q. 8.15 in "Mahler's Guide to Conjugate Priors".

28. 2.81 + 1.03 + (7)(-0.25) + (4)(0.12) - 0.43 - 1.08 = **1.06**
Comment: Similar to Q.10.9 (MAS-2, 5/19, Q.16) in "Mahler's Guide to Linear Mixed Models."

29. 1. For each observation, calculate Sort Ratio = $\frac{\text{Model A Predicted Pure Premium}}{\text{Model B Predicted Pure Premium}}$.

2. Sort the dataset based on the Sort Ratio, from smallest to largest.
3. Group the data in quintiles.

4. For each group, calculate the percent error for each model: $\frac{\text{Predicted Pure Premium}}{\text{Actual Pure Premium}} - 1$.

Observation	Actual Pure Premium	Model A Pure Premium	Model B Pure Premium	Ratio of Model A over Model B	Quintile
1	500	530	510	1.039	4
2	600	620	630	0.984	2
3	700	750	770	0.974	2
4	800	940	850	1.106	5
5	1000	1090	980	1.112	5
6	1200	1140	1150	0.991	3
7	1400	1430	1380	1.036	3
8	1600	1560	1490	1.047	4
9	1800	1710	1780	0.961	1
10	2000	1830	2060	0.888	1

The quintiles are observations: 10 and 9, 3 and 2, 6 and 7, 1 and 8, 4 and 5.

The actual pure premiums by quintile are: (2000 + 1800)/2 = 1900, (700 + 600)/2 = 650, (1200 + 1400)/2 = 1300, (500 + 1600)/2 = 1050, (800 + 1000)/2 = 900.

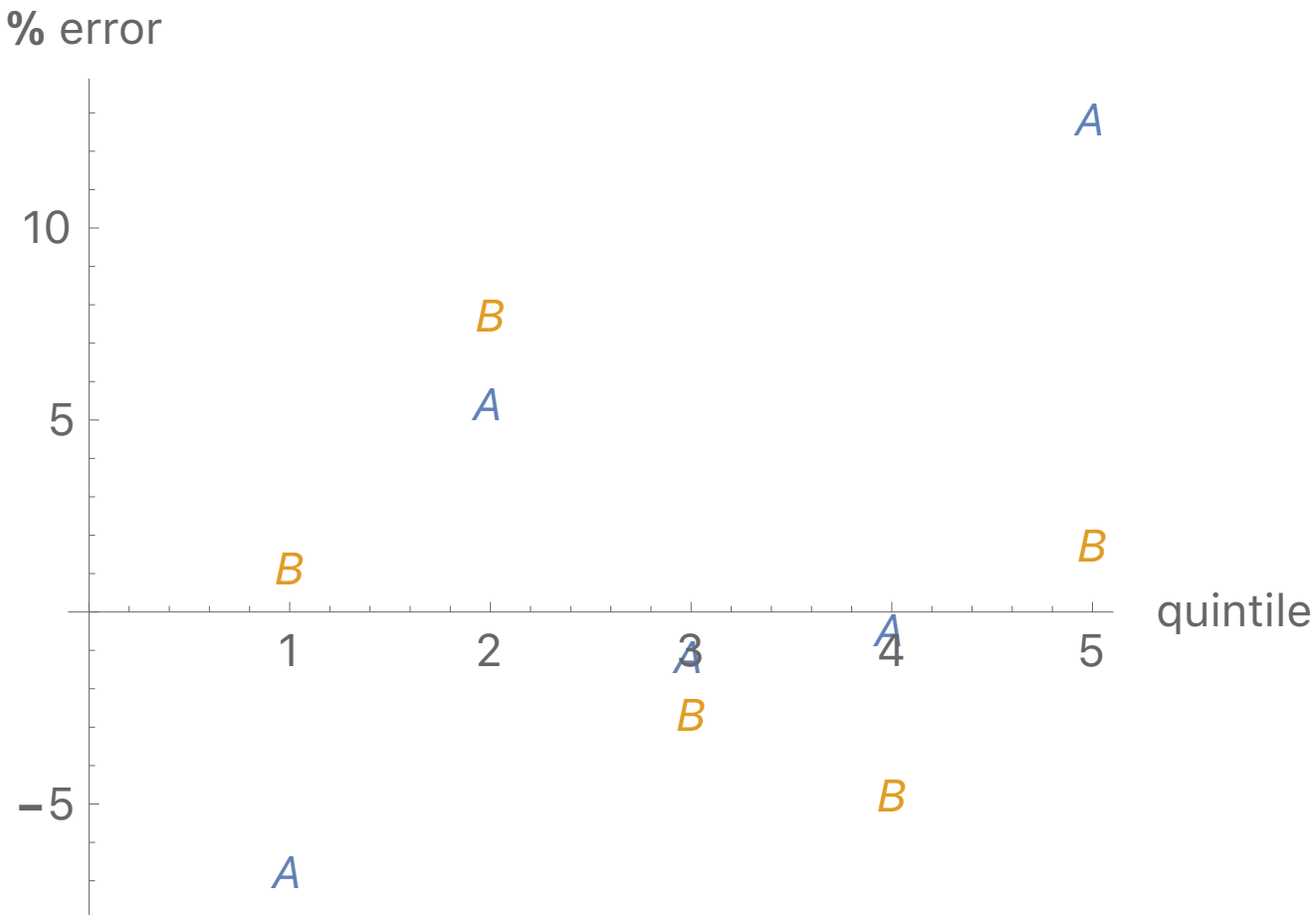
The Model A pure premiums by quintile are: (1830 + 1710)/2 = 1770, (750 + 620)/2 = 685, (1140 + 1430)/2 = 1285, (530 + 1560)/2 = 1045, (940 + 1090)/2 = 1015.

The Model B pure premiums by quintile are: (2060 + 1780)/2 = 1920, (770 + 630)/2 = 700, (1150 + 1380)/2 = 1265, (510 + 1490)/2 = 1000, (850 + 980)/2 = 915.

The percent errors for Model A: 1770/1900 - 1 = -6.8%, 685/650 - 1 = 5.4%, 1285/1300 - 1 = -1.2%, 1045/1050 - 1 = -0.5%, 1015/900 - 1 = 12.8%.

The percent errors for Model B: 1920/1900 - 1 = 1.1%, 700/650 - 1 = 7.7%, 1265/1300 - 1 = -2.7%, 1000/1050 - 1 = -4.8%, 915/900 - 1 = 1.7%.

The double lift chart is **Plot 3**:



Comment: Similar to Q.4.9 in "Mahler's Guide to Advanced GLMs".
See Section 7.2.2 in Generalized Linear Models for Insurance Rating.
This is the alternate representation of a double lift chart showing the percent errors.
"As an alternate representation of a double lift chart, one can plot two curves: the percent error for the model predictions and the percent error for the current loss costs, where percent error is calculated as: $\frac{\text{Predicted Loss Cost}}{\text{Actual Loss Cost}}$ - 1. In this case, the winning model is the one with the flatter line centered at y = 0, indicating that its predictions more closely match actual pure premium."

30. C. "The results obtained when we perform PCA will also depend on whether the variables have been individually scaled." "Because it is undesirable for the principal components obtained to depend on an arbitrary choice of scaling, we typically scale each variable to have standard deviation one before we perform PCA." Statement C is false.
Comment: Similar to Q.9.30 (SOA Exam SRM, Sample Q.5)
in "Mahler's Guide to Advanced Statistical Learning."

31. C. The a priori mean is: $E[\theta / (\alpha-1)] = E[\theta / 5] = E[\theta]/5 = 80{,}000 / 5 = 16{,}000$.
The mean for observed Years 1, 2, and 3 is: $(11{,}000 + 18{,}000 + 9000)/3 = 12{,}667$.
Therefore, $Z12{,}667 + (1 - Z)(16{,}000) = 15{,}000. \Rightarrow Z = 0.3. \Rightarrow 3/(3+K) = 0.3 \Rightarrow K = 7$.
For four years of data, the observed mean is: $(11{,}000 + 18{,}000 + 9000 + 7000)/4 = 11{,}250$,
and $Z = 4/(4 + 7) = 4/11$.
Revised estimate of Year 6 is: $(4/11)(11{,}250) + (7/11)(16{,}000) =$ **14,273**.
Comment: Similar to Q. 8.63 (4, 11/06, Q.6)
in "Mahler's Guide to Buhlmann Credibility."
For example, in 2015 you could be trying to estimate 2016; at first you only have data from years 2011 to 2013, but later locate the data from 2014 and use all four years of data to estimate 2016.
We are estimating one future year.
Since there is no inflation or differing exposures by year in this question, it does not matter which future year; the estimate for year 6 is the same as the estimate for year 5.

32. D.
Cluster D has the most homogeneous clusters; it has the smallest within cluster variation.
Comment: Similar to Q.11.7 in "Mahler's Guide to Advanced Statistical Learning."

33. E. For fixed parameters, the mean aggregate is: $\lambda \exp[\mu + \sigma^2/2] = \lambda \exp[\mu] \exp[\sigma^2/2]$.
Variance of aggregate is: λ(2nd moment of severity) $= \lambda \exp[2\mu + 2\sigma^2] = \lambda \exp[2\mu] \exp[2\sigma^2]$.

$$\text{EPV} = \int_{\lambda=0.03}^{0.08} \int_{\mu=7}^{9} \int_{\sigma=1}^{2} \lambda \exp[2\mu] \exp[2\sigma^2] \, 0.016\sigma/\lambda^2 \, d\sigma \, d\mu \, d\lambda$$

$$= \int_{0.03}^{0.08} 0.016/\lambda \, d\lambda \int_{7}^{9} \exp[2\mu] \, d\mu \int_{1}^{2} \sigma \exp[2\sigma^2] \, d\sigma$$

$$= 0.016 \ln[\lambda]\Big]_{\lambda=0.03}^{\lambda=0.08} \{\exp[2\mu]/2\}\Big]_{\mu=7}^{\mu=9} \{\exp[2\sigma^2]/4\}\Big]_{\sigma=1}^{\sigma=2}$$

$= (0.002)(0.980829)(64{,}457{,}365)(2973.57) = 375{,}987{,}885.$

$$\text{Overall mean} = \int_{\lambda=0.03}^{0.08} \int_{\mu=7}^{9} \int_{\sigma=1}^{2} \lambda \exp[\mu] \exp[\sigma^2/2] \, 0.016\sigma/\lambda^2 \, d\sigma \, d\mu \, d\lambda$$

$$= 0.016 \ln[\lambda]\Big]_{\lambda=0.03}^{\lambda=0.08} \exp[\mu]\Big]_{\mu=7}^{\mu=9} \exp[\sigma^2/2]\Big]_{\sigma=1}^{\sigma=2}$$

$= (0.016)(0.980829)(7006.45)(5.740335) = 631.17.$

$$\text{2nd moment of hypothetical means} = \int_{\lambda=0.03}^{0.08} \int_{\mu=7}^{9} \int_{\sigma=1}^{2} \lambda^2 \exp[2\mu] \exp[\sigma^2] \, 0.016\sigma/\lambda^2 \, d\sigma \, d\mu \, d\lambda$$

$$= 0.016 \lambda\Big]_{\lambda=0.03}^{\lambda=0.08} \{\exp[2\mu]/2\}\Big]_{\mu=7}^{\mu=9} \{\exp[\sigma^2]/2\}\Big]_{\sigma=1}^{\sigma=2}$$

$= (0.016)(0.05)(64{,}457{,}365/2)(51.8799/2) = 668{,}808.$

VHM $= 668{,}808 - 631.17^2 = 270{,}432.$

K = EPV/VHM = 375,987,885 / 270,432 = **1390**.

Comment: Similar to Q.10.72 (4, 11/04, Q.29) in "Mahler's Guide to Buhlmann Credibility."

The derivative of $\exp[2\sigma^2]$ is: $(2)(2\sigma) \exp[2\sigma^2]$. Thus the integral of $\sigma \exp[2\sigma^2]$ is $\exp[2\sigma^2]/4$.

If $y = 2\sigma^2$, then $\exp[2\sigma^2] \, 4\sigma \, d\sigma = \exp[y] \, dy$.

34. D. At the first step we group the two closest points: 26 and 38.
For each of the remaining points we calculate the maximum of the distances from this group:
{5} to {26,38} is Max[26 - 5, 38 - 5] = 33.
{64} to {26,38} is Max[64 - 26, 64 - 38] = 38.
{88} to {26,38} is Max[88 - 26, 88 - 38] = 62.
64 to 88 is 22.
So next we group 64 and 88.
{64, 88} to {26,38} is 62.
So we next group {5} with {26,38}.
The distance using the complete linkage is Max[26 - 5, 38 - 5] = **33**.
Comment: Similar to Q.12.23 (MAS-2, 11/18, Q.42)
in "Mahler's Guide to Advanced Statistical Learning."

35. D. REML takes into account the loss of degrees of freedom that results from estimating the fixed effects. Statement I is not true.
Using Restricted Maximum Likelihood uses the same formulas to compute estimates of the fixed effects as using Maximum Likelihood and will result in similar estimates of the fixed effects as using Maximum Likelihood; however, there will be some difference in the estimated fixed effects due to using different estimates of the covariance matrix.
Statement II is true.
ML estimates of the covariance parameters are biased, while those of REML are not.
Thus Statement III is true.
Comment: Similar to Q.5.11 (MAS-2, 11/18, Q.11) in "Mahler's Guide to Linear Mixed Models."

36. A., D.

For a random walk without drift, Covariance[x_t, x_{t+k}] = t σ^2. Thus B is not true.

The forecast error increases with the forecast horizon, so C is not true.

For the random walk with drift: $x_t = x_{t-1} + \delta + w_t$. Thus $x_t = \sum_{i=1}^{t} w_i + \delta t$.

Since δ is a known constant, for a given t, δt is a constant, and thus the variances, autocovariances, and autocorrelations are the same as for the random walk without drift.
Thus D is true.

$x_t - x_{t-1} = \sum_{i=1}^{t} w_i + \delta t - \{\sum_{i=1}^{t-1} w_i + \delta(t-1)\} = w_t + \delta$. Thus E is not true.

Comment: Similar to Q 6.7 (4, 11/00, Q.40) in "Mahler's Guide to Time Series."
For a random walk without drift, the first difference is a white noise series.

37. B. The density at 2 of a Binomial with m = 5 is: $10q^2(1-q)^3$.
The density at x of a Single Parameter Pareto Distribution, with $\theta = 10$ is: $\alpha 10^\alpha / x^{\alpha+1}$.
If the risk is from Class A, then the chance of the observation is:
$\{(10)(0.3^2)(0.7^3)\}\{(2)(10^2/31^3)\}\{(2)(10^2/17^3)\} = 0.00008437$.
If the risk is from Class B, then the chance of the observation is:
$\{(10)(0.6^2)(0.4^3)\}\{(3)(10^3/31^4)\}\{(3)(10^3/17^4)\} = 0.00002688$.
The mean of a Single Parameter Pareto Distribution is: $\alpha\theta/(\alpha-1)$.
The mean pure premium for Class A is: $\{(5)(0.3)\}\{(2)(10)/(2-1)\} = 30$.
The mean pure premium for Class B is: $\{(5)(0.6)\}\{(3)(10)/(3-1)\} = 45$.

Class	A Priori Chance of This Class	Chance of the Observation	Prob. Weight = Product of Columns B & C	Posterior Chance of This Class	Mean Pure Premium
A	0.6667	0.00008437	0.00005625	0.8626	30
B	0.3333	0.00002688	0.00000896	0.1374	45
Overall			0.00006521	1.0000	**32.06**

Comment: Similar to Q. 5.45 in "Mahler's Guide to Buhlmann Credibility."

38. Plot 1 has the lag 5 autocorrelation significantly different from zero, and none of the subsequent lags, thus it is MA(5).
Plot 2 has the lag 2 autocorrelation significantly different from zero, and none of the subsequent lags, thus it is MA(2).
Plot 3 has the lag 3 autocorrelation significantly different from zero, and none of the subsequent lags, thus it is MA(3).
Plot 4 has the lag 4 autocorrelation significantly different from zero, and none of the subsequent lags, thus it is MA(4).
Comment: Similar to Q 9.10 in "Mahler's Guide to Time Series."
For white noise, none of the autocorrelations would be significantly different from zero.

39. B. $E[X] = (0.7)(2) + (0.3)(5) = 2.9$. $E[X^2] = (0.7)(2^2) + (0.3)(5^2) = 10.3$.
$E[A] = (6)(2.9) = 17.4$. $Var[A] = (6)(10.3) = 61.8$.
$E[Y] = (0.5)(3) + (0.5)(7) = 5$. $E[Y^2] = (0.5)(3^2) + (0.5)(7^2) = 29$.
$E[B] = (6)(5) = 30$. $Var[B] = (6)(29) = 174$.
Given a certain number of claims, n, has occurred,
$AB = (X_1 + ... + X_n)(Y_1 + ... + Y_n) = X_1Y_1 + X_1Y_2 + ... + X_nY_n$.
$E[AB \mid n] = E[X_1Y_1 + X_1Y_2 + ... + X_nY_n] = n^2E[XY] = n^2E[X]E[Y] = n^2(2.9)(5) = 14.5n^2$.
$E[AB] = E_n[E[AB \mid n]] = E_n[14.5n^2] = 14.5\ E_n[n^2] =$
(14.5)(2nd moment of the Poisson) $= (14.5)(6 + 6^2) = 609$.
$Cov[A, B] = E[AB] - E[A]E[B] = 609 - (17.4)(30) = 87$.
$Corr[A, B] = Cov[A, B] / \sqrt{Var[A]Var[B]} = 87 / \sqrt{(61.8)(174)} =$ **0.839**.

Comment: Similar to Q. 3.45 (4, 11/01, Q.29) in "Mahler's Guide to Buhlmann Credibility."
X and Y are chosen from separate severity distributions in an independent manner.
Therefore, $E[XY] = E[X]E[Y]$.
If for example N = 3, then we have $3^2 = 9$ terms: $AB = (X_1 + X_2 + X_3)(Y_1 + Y_2 + Y_3) =$
$X_1Y_1 + X_1Y_2 + X_1Y_3 + X_2Y_1 + X_2Y_2 + X_2Y_3 + X_3Y_1 + X_3Y_2 + X_3Y_3 =$
$E[AB \mid N = 3] = E[X_1Y_1 + X_1Y_2 + X_1Y_3 + X_2Y_1 + X_2Y_2 + X_2Y_3 + X_3Y_1 + X_3Y_2 + X_3Y_3] =$
$E[X_1]E[Y_1] + E[X_1]E[Y_2] + E[X_1]E[Y_3] + E[X_2]E[Y_1] + E[X_2]E[Y_2] + E[X_2]E[Y_3] + E[X_3]E[Y_1]$
$+ E[X_3]E[Y_2] + E[X_3]E[Y_3] = 9\ E[X]\ E[Y]$.
If instead you have paired samples, $(X_1, Y_1), ..., (X_n, Y_n)$, for example the heights of husbands and wives, such that X_i and Y_j are independent for $i \neq j$, then:

$$Cov[\bar{X}, \bar{Y}] = E[\bar{X}\bar{Y}] - E[\bar{X}]E[\bar{Y}] = E[\frac{\sum_{i=1}^{N} X_i}{N} \frac{\sum_{j=1}^{N} Y_j}{N}] - E[X]\ E[Y] = \sum_{i=1}^{N}\sum_{j=1}^{N} E[X_iY_j] / N^2 - E[X]\ E[Y] =$$

$$\sum_{i \neq j}\sum_{j=1}^{N} E[X_iY_j] / N^2 + \sum_{i=1}^{N} E[X_iY_i] / N^2 - E[X]\ E[Y] =$$

$(N)(N-1)E[X]E[Y] + N\ (Cov[X, Y] + E[X]E[Y]) / N^2 - E[X]\ E[Y] = Cov[X,Y] / N$,
where Cov[X, Y] is the covariance of the elements of the paired samples.

40. D. Model D has more random effects than Models A, B, or C.
Model E has more random effects than Model D, but Model E suffers from aliasing.
Model E for P = 1:
$\ln[SALARY_{ij}] = \beta_0 + \beta_1\ LSAT_{ij} + \beta_2\ AGE_{ij} + \beta_3\ TUITION_j + \beta_4$
$+ u_{0j} + u_{1j}\ LSAT_{ij} + u_{2j}\ AGE_{ij} + u_{3j}\ LSAT_{ij} + \varepsilon_{ij}$
u_{ij} and u_{3j} are performing the same role; they are each adjusting the slope of LSAT for law school j. Model E is overdetermined!
Comment: Similar to Q.8.21 (MAS-2, 11/18, Q.12) in "Mahler's Guide to Linear Mixed Models."

41. A. $\pi(\lambda)$ is proportional to: $\lambda^4 e^{-100\lambda}$, $\lambda > \infty$.
The chance of one claims given λ is: $0.9\lambda e^{-\lambda} + 0.1\ (4\lambda)e^{-4\lambda}$.
Thus the posterior distribution is proportional to:
$\lambda^4 e^{-100\lambda}\ (9\lambda e^{-\lambda} + 4\lambda e^{-4\lambda}) = 9\lambda^5 e^{-101\lambda} + 4\lambda^5 e^{-104\lambda}$.
The mean frequency given λ is: $0.9\ \lambda + (0.1)(4\lambda) = 1.3\lambda$.
Thus the posterior mean frequency is:

$$1.3\ \frac{9\int_0^\infty \lambda^6 e^{-101\lambda}\ d\lambda\ + 4\int_0^\infty \lambda^6 e^{-104\lambda}\ d\lambda}{9\int_0^\infty \lambda^5 e^{-101\lambda}\ d\lambda\ + 4\int_0^\infty \lambda^5 e^{-104\lambda}\ d\lambda} = 1.3\ \frac{(9)\ (6!\ /\ 101^7) + (4)\ (6!\ /\ 104^7)}{(9)\ (5!\ /\ 101^6) + (4)\ (5!\ /\ 104^6)} = \mathbf{0.0766}.$$

Comment: Similar to Q. 6.46 in "Mahler's Guide to Buhlmann Credibility."

For Gamma type integrals: $\int_0^\infty t^n e^{-ct}\ dt = n!\ /\ c^{n+1}$.

One can not use Gamma-Poisson shortcuts to answer this question which has a mixed distribution of Poissons rather than a single Poisson Distribution.

42. D. The null hypothesis has one non-zero variance while the alternate hypothesis has two non-zero variances. Thus the test statistic follows a 50%-50% mixture of
Chi-Square Distributions with 1 and 2 degrees of freedom.
One uses Restricted Maximum Likelihood when testing random effects / covariance parameters.
The test statistic is: (2) {(-569.398) - (-572.413)} = 6.03.
Using the Chi-Square Table, $1.0\% < \text{Prob}[\chi^2_1 > 6.03] < 2.5\%$.

$2.5\% < \text{Prob}[\chi^2_2 > 6.003] < 5.0\%$. Thus the p-value is between 1.25% and 3.75%.

For 1 degree of freedom the 2.5% and 1.0% critical values are 5.02 and 6.33.
So the p-value for 6.03 is very approximately 1.4%.
For 2 degrees of freedom the 5% and 2.5% critical values are 5.99 and 7.38.
So the p-value for 6.03 is very approximately 4.9%.
Thus the average of the two p-values is about: (50%)(1.4%) + (50%)(4.9%) = 3.2%.
Therefore, do not reject H_0 at 0.025, and reject H_0 at 0.050.
Comment: Similar to Q.6.10 in "Mahler's Guide to Linear Mixed Models".
Using a computer the p-value is: (1.406% + 4.905%)/2 = 3.16%.
Likelihood Ratio Tests are applied to fixed effects and to random effects / covariance parameters. One uses Maximum Likelihood when testing fixed effects. Instead one uses Restricted Maximum Likelihood when testing random effects / covariance parameters.

While these solutions are believed to be correct, anyone can make a mistake.
If you believe you've found something that may be wrong, send any corrections or comments to:
Howard Mahler, Email: hmahler@mac.com