SMI Practice Papers (Story-Based)

Sets 9 to 14 — with detailed answers directly below each question

How to use

Attempt each set under timed conditions. Do not jump to the answer until you have written your method choice, formulas, and numeric steps. The wording is intentionally general so you must infer the correct technique.

Contents

SMI PRACTICE PAPER — SET 9 (TOTAL 35)

(Story-first, method not named; you must infer what to use)

General Instructions:
- Attempt ALL questions.
- Statistical tables / critical values are provided inside the question.
- Show steps and interpret in plain business terms.

Question 1 (9 marks)

A subscription app is planning server capacity for a premium rollout. Leadership wants a reliable estimate of the fraction of users who will convert to premium.

Part (a) The team wants a 98% confidence estimate with margin of error at most 0.04.
If they have no prior information about conversion, what minimum sample size should they co
Given: z0.99 = 2.326 [5]

Part (b) After the survey, 212 out of 680 users said they would upgrade.
Construct a 98% confidence interval for the true conversion proportion and interpret it for
Given: z0.99 = 2.326 [4]

Answer 1(a)

Decoding cue: 'fraction/share/proportion' + 'margin of error' + 'confidence'.
Use worst-case p = 0.5 when no prior estimate is available.

```
n = (z^2 * p(1-p)) / E^2
  = (2.326^2 * 0.25) / 0.04^2
  = (5.411 * 0.25) / 0.0016
  = 1.35275 / 0.0016
  = 845.47  =>  n = 846 (round up)
```

Answer 1(b)

```
p_hat = 212/680 = 0.3118
SE = sqrt( p_hat(1-p_hat) / n )
   = sqrt(0.3118*0.6882/680)
   = sqrt(0.2146/680)
   = sqrt(0.0003156) = 0.01777
Margin = z0.99 * SE = 2.326 * 0.01777 = 0.0413
98% CI = 0.3118 ± 0.0413 = (0.2705, 0.3531)
```

Interpretation: plan for true conversion roughly 27.1% to 35.3%.

Question 2 (9 marks)

A hospital studies recovery time (days) under two decisions: Treatment Type (Standard vs New) and Therapy Intensity (Low vs High).

Mean recovery times (lower is better):

	Low	High	Row Mean
Standard	16	10	13.0
New	9	15	12.0
Column Mean	12.5	12.5	12.5

Two-factor results summary:

Source	SS	df	MS	F	p
Treatment Type	2.0	1	2.0	0.50	0.49
Intensity	0.0	1	0.0	0.00	1.00
Treatment×Intensity	72.0	1	72.0	18.00	0.001
Error	48.0	12	4.0		
Total	122.0	15			

Part (a) From the mean table, describe the pattern. At α=0.05, interpret which effects are
Part (b) Explain what a significant Treatment×Intensity effect means in practical terms.
Also explain why column means are identical (12.5 for Low and High) but the interaction is

Answer 2(a)

From means:

- Standard: High reduces recovery time (16 -> 10) i.e., High helps.
- New: High increases recovery time (9 -> 15) i.e., High hurts.

This reversal indicates strong interaction.
At α=0.05: Treatment p=0.49 (NS), Intensity p=1.00 (NS), Interaction p=0.001 (Significant).

Answer 2(b)

Meaning: there is no single 'best intensity' overall; best intensity depends on treatment.
Paradox: column means average across treatments; opposite effects cancel out:

- Under Standard, High is 6 days better than Low.
- Under New, High is 6 days worse than Low.

So the overall average difference is 0, yet conditional differences are large => interactio

----

Question 3 (9 marks)

A retail team models Monthly Revenue (lakhs) using Marketing Spend (lakhs) and Footfall (thousands). Using 20 months of data, they fit: Revenue = $\beta0 + \beta1$(Marketing) + $\beta2$(Footfall) + ε.

Regression output:

Predictor	Estimate	Std.Error	t-stat	p-value
Intercept	7.6	2.4	3.17	0.005
Marketing	2.1	0.5	4.20	0.0004
Footfall	1.4	0.8	1.75	0.096

Additional: R^2=0.74, Adj R^2=0.71, Overall model F-test p-value=0.0001, α=0.05

Part (a) Decide whether the overall model is useful. Then evaluate each predictor.
A manager says: 'Drop Footfall to simplify; it's not needed.' Do you agree? Justify. [4]
Part (b) Revenue has an upward trend and repeating quarterly seasonal spikes.
Explain why a method that models level+trend+seasonality is better than one that only updat
Also give one advantage and one disadvantage of a 3-month moving average vs exponential smo

Answer 3(a)

Overall F-test p=0.0001 < 0.05 => the model is useful (predictors jointly matter).
Predictors:

- Marketing p=0.0004 < 0.05 => significant.
- Footfall p=0.096 > 0.05 => not significant at 5% (weak evidence).

Do not drop Footfall automatically:

- p=0.096 can be 'borderline' depending on context (forecasting vs explanation).
- Footfall may be correlated with Marketing (multicollinearity), inflating SE and p-value.

```
- Best practice: compare reduced vs full model (Adj R², AIC/BIC, out-of-sample error).

Answer 3(b)
---------------
Trend + seasonality implies forecasts should include components for both, otherwise forecas
trend and miss seasonal spikes. Methods like Holt-Winters explicitly model level+trend+seas
3-month moving average:
+ Advantage: very simple; smooths random noise.
- Disadvantage: lags turning points; handles seasonality poorly; discards older data abrupt
Exponential smoothing:
+ Advantage: weights recent data more smoothly; adapts better.
- Disadvantage: basic form models only level unless extended (Holt/Holt-Winters).

-----------------------------------------
Question 4 (8 marks)
-----------------------------------------
A company runs a product trial: 40 users test a feature; 26 continue using it.

Part (a) Using Maximum Likelihood Estimation, derive the estimate of p = P(user continues).
Show: (i) likelihood, (ii) log-likelihood, (iii) derivative and solution. Interpret p■. [4

Part (b) Adoption is modeled using Age with:
log(odds) = 2.9 − 0.07×Age
Compute adoption probability for Age=35.
Then explain why this approach is more appropriate than ordinary linear regression for a 0/
(give 2 concrete issues with linear regression here). [4]

Answer 4(a)
---------------
Decoding cue: 26 successes out of 40 => Bernoulli/Binomial parameter p.
Let x=26, n=40.
L(p) = p^x (1-p)^(n-x) = p^26 (1-p)^14
■(p) = 26 ln p + 14 ln(1-p)
d■/dp = 26/p − 14/(1-p) = 0
=> 26(1-p) = 14p  => 26 = 40p  => p_hat = 26/40 = 0.65
Interpretation: estimated continuation probability is 65%.

Answer 4(b)
---------------
Age=35: z = 2.9 − 0.07*35 = 0.45
p = 1/(1+e^{−z}) = 1/(1+e^{−0.45}) ≈ 0.610
Why logistic beats linear regression for 0/1:
1) Logistic keeps predicted p in [0,1]; linear regression can predict <0 or >1.
2) 0/1 errors are non-normal and variance is p(1-p) (heteroscedastic); OLS inference breaks

================================================================================
```

SMI PRACTICE PAPER — SET 10 (TOTAL 35)

(Story-first, same pattern, different numbers)

General Instructions:

- Attempt ALL questions.
- Use values provided; do not use external tables.

Question 1 (9 marks)

A fintech app wants to estimate the proportion of users who will enable AutoPay. They want 99% confidence and margin of error ≤ 0.03, with no prior estimate.

Given: z0.995 = 2.576

Part (a) Find the required sample size. [5]

Part (b) In the collected sample, 384 out of 1200 enabled AutoPay.
Construct the 99% CI and interpret. [4]

Answer 1(a)

No prior estimate => use worst-case p=0.5.

```
n = (z^2 * 0.25) / E^2
  = (2.576^2 * 0.25) / 0.03^2
  = (6.635 * 0.25) / 0.0009
  = 1.65875 / 0.0009
  = 1843.06  => n = 1844
```

Answer 1(b)

```
p_hat = 384/1200 = 0.32
SE = sqrt(0.32*0.68/1200) = sqrt(0.2176/1200) = sqrt(0.0001813) = 0.01346
Margin = 2.576*0.01346 = 0.0347
99% CI = 0.32 ± 0.0347 = (0.2853, 0.3547)
```

Interpretation: true enable rate is likely 28.5%–35.5%.

Question 2 (9 marks)

A lab studies two factors affecting completion time (minutes): Tool (Old vs New) and Training (Basic vs Advanced).

Mean completion times (lower is better):

	Basic	Advanced	Row Mean
Old	40	28	34
New	26	32	29
Column Mean	33	30	31.5

ANOVA summary:

Source	SS	df	MS	F	p
Tool	5	1	5	1.11	0.31
Training	0	1	0	0.00	1.00
Tool×Training	60	1	60	13.33	0.003
Error	54	12	4.5		
Total	119	15			

Part (a) Describe pattern + which effects are significant at α=0.05. [4]

Part (b) Explain the equal Training column means but strong interaction.

```
         Give a practical recommendation.                          [5]

Answer 2(a)
---------------
Pattern:
- Old tool: Advanced is much faster than Basic (40 -> 28).
- New tool: Advanced is slower than Basic (26 -> 32).
=> reversal => interaction.
Significance (α=0.05): Tool NS (p=0.31), Training NS (p=1.00), Interaction significant (p=0

Answer 2(b)
---------------
Column means are averages across tools; opposite training effects cancel when averaged.
Recommendation: choose training conditional on tool:
- If Old tool => Advanced.
- If New tool => Basic.

-----------------------------------------
Question 3 (9 marks)
-----------------------------------------
A firm predicts Profit (lakhs) from AdSpend (lakhs) and PriceDiscount (%). Using n=18
months, they fit a multiple regression model.

Output:
Predictor   Estimate  Std.Error  t-stat  p-value
Intercept     5.9       1.9       3.11    0.006
AdSpend       1.8       0.5       3.60    0.003
Discount     -0.9       0.4      -2.25    0.038
Additional: Overall F-test p=0.0005, R²=0.70, Adj R²=0.66

Part (a) Decide overall model significance and interpret signs of coefficients. [4]
Part (b) Management wants a simpler model with only AdSpend.
         Would you accept, given Discount is significant? Justify.   [5]

Answer 3(a)
---------------
Overall F-test p=0.0005 < 0.05 => model significant.
Coefficient interpretation (holding other predictor fixed):
- AdSpend +1.8 => higher ad spend increases profit.
- Discount -0.9 => higher discount decreases profit.

Answer 3(b)
---------------
Since Discount is significant (p=0.038), dropping it can:
- reduce explanatory power,
- bias AdSpend effect if Discount correlates with AdSpend,
- worsen prediction if Discount carries signal.
Decision should be based on model comparison (Adj R²/AIC/BIC and out-of-sample error).
If goal is policy/explanation => keep Discount. If goal is pure simplicity => validate redu

-----------------------------------------
Question 4 (8 marks)
-----------------------------------------
A feature is shown to 30 users; 21 continue using it.
Part (a) Using MLE, estimate p and show likelihood -> log-likelihood -> derivative. [4]
Part (b) Adoption is modeled by age: log(odds) = 3.4 − 0.09×Age.
         Compute adoption probability at Age=40 and give 2 reasons logistic is suitable for
Answer 4(a)
---------------
n=30, x=21
L(p) = p^21 (1-p)^9
■(p) = 21 ln p + 9 ln(1-p)
d■/dp = 21/p − 9/(1-p) = 0
=> 21(1-p)=9p => 21=30p => p_hat=0.70

Answer 4(b)
---------------
Age=40: z = 3.4 − 0.09*40 = -0.2
p = 1/(1+e^{−z}) = 1/(1+e^{0.2}) = 1/(1+1.221) ≈ 0.450
Reasons:
1) Ensures predicted probabilities in [0,1].
2) Correct error structure for Bernoulli outcomes; avoids OLS heteroscedasticity/non-normal

================================================================================
```

SMI PRACTICE PAPER — SET 11 (TOTAL 35)
(Story-first. You must decide the right technique from the wording.)

Instructions:
- Attempt all 4 questions.
- Use ONLY the values provided (no table lookup).
- Write conclusions in plain language.

Question 1 (9 marks)

A fintech product lead wants a reliable estimate of the "share of users" who will enable UPI A

(a) They want 95% confidence with margin of error at most 0.05 and have no prior estimate.
What minimum sample size should be collected?
Given: z0.975 = 1.96 [5]

(b) In a survey of 520 users, 162 enabled AutoPay. Build a 95% confidence interval for the true enable rate and interpret it for planning. [4]

Question 2 (9 marks)

A clinic compares recovery time (days) under two decisions:
- Protocol: Standard vs New
- Support: Basic vs Intensive

Mean recovery time (lower is better):

	Basic	Intensive	Row Mean
Standard	18	12	15.0
New	11	17	14.0
Column Mean	14.5	14.5	14.5

Two-factor analysis summary:

Source	SS	df	MS	F	p
Protocol	2.0	1	2.0	0.50	0.49
Support	0.0	1	0.0	0.00	1.00
Protocol×Support	64.0	1	64.0	16.00	0.002
Error	48.0	12	4.0		
Total	114.0	15			

(a) From the mean table, describe what is happening.
At alpha = 0.05, which effects matter statistically? [4]

(b) Explain in practical terms why the "Support" averages look identical
(Basic mean = Intensive mean = 14.5) but the interaction is significant. [5]

Question 3 (9 marks)

A retail team builds a model to explain Monthly Revenue (lakhs) using:
- Marketing Spend (lakhs)
- Avg Basket Size (items)

They fit: Revenue = beta0 + beta1(Spend) + beta2(Basket) + error using n=22 months.

Output:

```
Predictor        Estimate   Std.Error   t-stat   p-value
Intercept         6.4         2.1        3.05     0.006
Spend             1.9         0.4        4.75     0.0001
Basket            0.7         0.6        1.17     0.255

Overall model F-test p-value = 0.0002, R^2 = 0.71, Adj R^2 = 0.68
Assume alpha = 0.05.

(a) A manager says: "The model is clearly useful, so every predictor is useful."
    Correct or incorrect? Justify using the output above.             [4]

(b) The revenue series shows an upward trend and strong repeating monthly seasonality.
    Which forecasting approach is appropriate and why?
    Also: one advantage and one disadvantage of moving average vs exponential smoothing. [5]

-----------------------------------------
Question 4 (8 marks)
-----------------------------------------
A delivery platform tracks number of customer complaints per day for 12 days:
2, 1, 0, 3, 4, 1, 2, 3, 1, 5, 3, 2

(a) Treating the daily complaint count as coming from a single-parameter count model,
    find the maximum-likelihood estimate of the rate parameter and interpret it.  [4]

(b) Adoption is modeled as:
    log(odds of adoption) = 1.8 − 0.06 × Age
    Compute adoption probability for Age = 30.
    Give TWO reasons this modeling approach is preferred over ordinary linear regression
    for a 0/1 outcome.                                                 [4]
===============================================================================

===============================================================================
SET 11 — ANSWERS (Show steps; exam-style)
===============================================================================

-----------------------------------------
Answer 1(a)
-----------------------------------------
Decoding cue: "share", "confidence", "margin of error", "no prior estimate" -> proportion samp
Worst-case p(1-p)=0.25.

n = (z^2 * 0.25) / E^2
  = (1.96^2 * 0.25) / 0.05^2
  = (3.8416 * 0.25) / 0.0025
  = 0.9604 / 0.0025
  = 384.16  -> round up
n = 385

-----------------------------------------
Answer 1(b)
-----------------------------------------
p_hat = x/n = 162/520 = 0.3115

SE = sqrt( p_hat(1-p_hat) / n )
   = sqrt(0.3115*0.6885/520)
   = 0.02031

Margin = z * SE = 1.96 * 0.02031 = 0.03981

95% CI = 0.3115 ± 0.0398 = (0.2717, 0.3513)
```

```
Interpretation: True enable rate is likely ~27.2% to 35.1%. Plan capacity using this range.

-----------------------------------------
Answer 2(a)
-----------------------------------------
From means:
Standard: Basic 18 -> Intensive 12 (Intensive helps a lot)
New:      Basic 11 -> Intensive 17 (Intensive hurts!)
The direction flips across protocol -> interaction pattern.

At alpha=0.05 using p-values:
Protocol p=0.49 (not significant)
Support  p=1.00 (not significant)
Protocol×Support p=0.002 (significant)

Only the interaction effect is statistically significant.

-----------------------------------------
Answer 2(b)
-----------------------------------------
Why column means are identical (14.5 and 14.5) but interaction is significant:
When you average across protocols, opposite effects cancel out.

Effect of Intensive within Standard: 12 - 18 = -6 (reduces time by 6)
Effect of Intensive within New:      17 - 11 = +6 (increases time by 6)

Averaging these gives 0 overall, so main effect of Support looks zero,
even though conditional effects are large and opposite.
Practical conclusion: choose Support level conditional on Protocol; do NOT interpret Support a

-----------------------------------------
Answer 3(a)
-----------------------------------------
Overall model is useful because F-test p=0.0002 < 0.05 (predictors jointly help).

But NOT every predictor is useful:
- Spend p=0.0001 < 0.05 -> significant predictor
- Basket p=0.255 > 0.05 -> not significant (given other variables in model)

So manager statement is incorrect.

-----------------------------------------
Answer 3(b)
-----------------------------------------
Trend + strong monthly seasonality -> use a method that models level + trend + seasonality
(e.g., Holt-Winters / seasonal exponential smoothing).

Moving average vs exponential smoothing:
+ Moving average advantage: very simple, smooths noise.
- Moving average disadvantage: lags trend/seasonality; drops older data abruptly.

+ Exponential smoothing advantage: adaptive; weights recent data more; can be extended to tren
- Exponential smoothing disadvantage: choice of alpha (and seasonal parameters) matters; may b

-----------------------------------------
Answer 4(a)
-----------------------------------------
Decoding cue: "complaints per day" = count data; single-parameter Poisson is standard.
For Poisson, MLE of lambda is sample mean.
Sum = 2+1+0+3+4+1+2+3+1+5+3+2 = 27
```

```
n = 12
lambda_hat = 27/12 = 2.25

Interpretation: estimated average complaints per day is 2.25.

-----------------------------------------
Answer 4(b)
-----------------------------------------
Age=30:
z = 1.8 - 0.06*30 = 1.8 - 1.8 = 0
p = 1/(1+e^{-z}) = 1/(1+1) = 0.50

Two reasons logistic is preferred for 0/1 outcomes:
1) Predicted probabilities stay in [0,1] (linear regression can give <0 or >1).
2) Error structure is not normal and variance is p(1-p) (heteroscedastic); OLS assumptions fai
===============================================================================
===============================================================================
SMI PRACTICE PAPER — SET 12 (TOTAL 35)
(Story-first. Technique not named; infer from scenario.)
===============================================================================
Instructions:
- Attempt all 4 questions.
- Use ONLY provided critical values.
- Write a clear decision + interpretation.

-----------------------------------------
Question 1 (9 marks)
-----------------------------------------
A product team compares satisfaction ratings for the SAME 9 users before and after a UI change
Define d = After − Before. Summary of differences:
d_bar = 2.8,   s_d = 1.4,   n = 9

Given:
t0.95,8  = 1.860  (one-tailed)
t0.975,8 = 2.306  (two-tailed, 95% CI)

(a) Is there evidence at alpha = 0.05 that satisfaction increased?              [5]
(b) Build a 95% CI for the average increase and interpret.                     [4]

-----------------------------------------
Question 2 (9 marks)
-----------------------------------------
A company tests three warehouse layouts (A, B, C) for average order-pick time (minutes).
They observe:

A: 22, 20, 23, 21, 24
B: 25, 27, 26, 28, 24
C: 19, 18, 20, 17, 21

Given: F0.95,(2,12) = 3.885

(a) Decide if the layouts differ in mean pick time at alpha = 0.05. Show steps. [6]
(b) What should be done NEXT if you find a difference? Name one post-hoc and one check. [3]

-----------------------------------------
Question 3 (9 marks)
-----------------------------------------
A sales analyst studies how revenue changes with the number of store visits (X).
They computed the following summary from n=10 months:

x_bar = 5,   y_bar = 18
Sxx = 28,    Sxy = 56
SSE = 20

Given: t0.975,8 = 2.306

(a) Fit the line: Y_hat = b0 + b1 X.                                          [4]
(b) Test whether the relationship is real at alpha = 0.05.                    [3]
(c) Give a 95% CI for the slope and interpret.                                [2]

-----------------------------------------
Question 4 (8 marks)
-----------------------------------------
A questionnaire has 6 items (V1..V6). Suitability checks:
KMO = 0.73, Bartlett p < 0.001
Eigenvalues: 2.7, 1.1, 0.8, 0.6, 0.5, 0.3
```

```
Rotated loadings (2 factors kept):
V1: (0.83, 0.10)
V2: (0.79, 0.15)
V3: (0.74, 0.20)
V4: (0.12, 0.81)
V5: (0.18, 0.77)
V6: (0.30, 0.35)

(a) How many factors would you keep and why?                             [3]
(b) Compute communalities for V1 and V6. Which is poorly represented?     [5]
===============================================================================


===============================================================================
SET 12 — ANSWERS
===============================================================================

-----------------------------------------
Answer 1(a)
-----------------------------------------
Decoding cue: SAME users measured twice -> test mean of paired differences.
H0: mu_d = 0
H1: mu_d > 0

SE = s_d / sqrt(n) = 1.4 / sqrt(9) = 1.4/3 = 0.4667
t  = d_bar / SE = 2.8 / 0.4667 = 6.000

Critical (one-tailed): t0.95,8 = 1.860
Decision: 6.000 > 1.860 -> Reject H0
Conclusion: strong evidence average satisfaction increased.

-----------------------------------------
Answer 1(b)
-----------------------------------------
95% CI: d_bar ± t0.975,8 * SE
SE = 0.4667
Margin = 2.306 * 0.4667 = 1.076

CI = 2.8 ± 1.076 = (1.724, 3.876)

Interpretation: average increase is likely between ~1.72 and ~3.88 points.

-----------------------------------------
Answer 2(a)
-----------------------------------------
Decoding cue: 3 groups, numeric response -> compare group means -> one-way ANOVA.

Compute group means:
A mean = 22
B mean = 26
C mean = 19
Grand mean = (22+26+19)/3 = 22.3333 (since equal n, same as overall)

Between SS:
SSB = sum( n*(mean_i - grand)^2 ) with n=5
    = 5[(22-22.3333)^2 + (26-22.3333)^2 + (19-22.3333)^2]
    = 123.3333

Within SS:
SSW = sum within-group squared deviations
A: (22-22)^2+(20-22)^2+(23-22)^2+(21-22)^2+(24-22)^2 = 10
```

```
B: deviations from 26 -> 0+1+0+4+4 = 10
C: deviations from 19 -> 0+1+1+4+4 = 10
SSW = 30

dfB = k-1 = 2
dfW = N-k = 15-3 = 12

MSB = 123.3333/2  = 61.6667
MSW = 30/12       = 2.5000

F = 61.6667/2.5 = 24.6667

Compare with F0.95,(2,12)=3.885
Decision: 24.667 > 3.885 -> Reject H0
Conclusion: mean pick times differ across layouts.

-----------------------------------------
Answer 2(b)
-----------------------------------------
Next steps:
- Post-hoc: Tukey HSD (or Bonferroni pairwise comparisons) to identify which pairs differ.
- Assumption checks: equal variances (Levene test) and normal residuals (QQ plot).

-----------------------------------------
Answer 3(a)
-----------------------------------------
b1 = Sxy/Sxx = 56/28 = 2.0
b0 = y_bar - b1*x_bar = 18 - 2*5 = 8
So: Y_hat = 8 + 2X

-----------------------------------------
Answer 3(b)
-----------------------------------------
Test H0: beta1 = 0 vs H1: beta1 != 0
df = n-2 = 8
MSE = SSE/df = 20/8 = 2.5

SE(b1) = sqrt(MSE/Sxx) = sqrt(2.5/28) = 0.2988
t = b1/SE(b1) = 2/0.2988 = 6.69

Critical two-sided: t0.975,8 = 2.306
Decision: |6.69| > 2.306 -> Reject H0
Conclusion: visits have a significant relationship with revenue.

-----------------------------------------
Answer 3(c)
-----------------------------------------
95% CI for slope:
b1 ± t*SE(b1) = 2 ± 2.306(0.2988)
Margin = 0.689
CI = (1.311, 2.689)

Interpretation: each +1 visit increases mean revenue by about 1.31 to 2.69 units (lakhs).

-----------------------------------------
Answer 4(a)
-----------------------------------------
KMO=0.73 (>0.6) and Bartlett significant -> factor model is appropriate.
Kaiser rule: keep eigenvalues > 1 -> 2 factors (2.7 and 1.1).

-----------------------------------------
```

```
Answer 4(b)
-----------------------------------------
Communality h^2 = sum of squared loadings (across retained factors).

V1: (0.83, 0.10)
h1^2 = 0.83^2 + 0.10^2 = 0.6889 + 0.0100 = 0.6989

V6: (0.30, 0.35)
h6^2 = 0.30^2 + 0.35^2 = 0.0900 + 0.1225 = 0.2125

V6 is poorly represented (low communality -> most variance is uniqueness/error).
===============================================================================
```

==

SMI PRACTICE PAPER — SET 13 (TOTAL 35)

(Story-first. You must infer the tool.)

==

Instructions:
- Attempt all questions.
- Use provided critical values only.

Question 1 (9 marks)

A manufacturing unit samples 16 batches to estimate mean defect count per 1000 items.
They observe: x_bar = 52, s = 8, n = 16
Given: t0.975,15 = 2.131

(a) Construct a 95% confidence interval for the true mean defect count. [5]
(b) Explain what changes to the interval width if n is doubled (qualitative). [4]

Question 2 (9 marks)

A website team checks whether device type and response category are related.

	Buy	Maybe	No	Total
Mobile	50	30	20	100
Desktop	30	20	50	100
Total	80	50	70	200

Given: chi^2_0.95,2 = 5.991

(a) Write H0 and H1. [2]
(b) Compute expected counts for all 6 cells. [3]
(c) Compute the chi-square statistic and conclude at alpha=0.05. [4]

Question 3 (9 marks)

A churn model uses a single score X:
log(odds of churn) = -0.8 + 0.4X

(a) Compute churn probability when X = 5. [4]
(b) Interpret exp(0.4) as an odds ratio (plain language). [3]
(c) Give TWO practical reasons this model form is chosen for a yes/no event. [2]

Question 4 (8 marks)

A pilot test: 50 users receive a feature; 18 continue using it after 1 week.

(a) Using maximum likelihood, estimate p = P(continue). Show the key steps. [4]
(b) Interpret the estimate as a business metric. [4]

==

==

SET 13 — ANSWERS

==

Answer 1(a)

```
Decoding cue: estimate mean, sigma unknown, small n -> t-interval.

SE = s/sqrt(n) = 8/sqrt(16) = 8/4 = 2
Margin = t0.975,15 * SE = 2.131 * 2 = 4.262

95% CI = 52 ± 4.262 = (47.738, 56.262)

-----------------------------------------
Answer 1(b)
-----------------------------------------
CI width is proportional to 1/sqrt(n).
Doubling n reduces SE by factor sqrt(2), so the interval becomes narrower
(roughly 1/sqrt(2) times the previous half-width).

-----------------------------------------
Answer 2(a)
-----------------------------------------
H0: Device type and response category are independent.
H1: They are associated (not independent).

-----------------------------------------
Answer 2(b)
-----------------------------------------
Expected E_ij = (row total * column total) / N

Row totals: 100 each, N=200
Column totals: Buy 80, Maybe 50, No 70

Mobile:
E(Buy)   = 100*80/200 = 40
E(Maybe) = 100*50/200 = 25
E(No)    = 100*70/200 = 35

Desktop (same because row total is same):
E(Buy)=40, E(Maybe)=25, E(No)=35

-----------------------------------------
Answer 2(c)
-----------------------------------------
chi^2 = sum (O-E)^2/E

Mobile:
(50-40)^2/40 = 100/40 = 2.5
(30-25)^2/25 = 25/25  = 1.0
(20-35)^2/35 = 225/35 = 6.4286

Desktop:
(30-40)^2/40 = 100/40 = 2.5
(20-25)^2/25 = 25/25  = 1.0
(50-35)^2/35 = 225/35 = 6.4286

Total chi^2 = 19.8571
df = (2-1)(3-1) = 2

Compare with chi^2_0.95,2 = 5.991
Decision: 19.857 > 5.991 -> Reject H0
Conclusion: device type and response category are associated.

-----------------------------------------
Answer 3(a)
```

```
-----------------------------------------
X=5:
z = -0.8 + 0.4*5 = -0.8 + 2 = 1.2
p = 1/(1+e^{-1.2}) = 0.7685 (approx)

-----------------------------------------
Answer 3(b)
-----------------------------------------
Odds ratio for +1 increase in X:
OR = exp(0.4) = 1.4918

Interpretation: each 1-point increase in X multiplies churn odds by ~1.49
(about a 49% increase in odds), holding other factors constant.

-----------------------------------------
Answer 3(c)
-----------------------------------------
Two practical reasons:
1) Keeps predicted probabilities within [0,1].
2) Correctly models non-constant variance of binary data (p(1-p)) and supports odds interpreta

-----------------------------------------
Answer 4(a)
-----------------------------------------
Decoding cue: 50 trials, 18 successes -> Bernoulli/Binomial likelihood.

L(p) = p^18 (1-p)^(32)
log L = 18 ln p + 32 ln(1-p)

d/dp: 18/p - 32/(1-p) = 0
=> 18(1-p) = 32p
=> 18 = 50p
p_hat = 18/50 = 0.36

-----------------------------------------
Answer 4(b)
-----------------------------------------
Interpretation: estimated continuation rate is 36%.
Out of 100 similar users, about 36 are expected to remain active after 1 week.
===============================================================================
```

SMI PRACTICE PAPER — SET 14 (TOTAL 35)

(Story-first; mixed numericals like sample.)

Instructions:
- Attempt all questions.
- Use provided critical values only.

Question 1 (9 marks)

Two independent teams use different onboarding flows. Time to first purchase (minutes) is meas

Team A: n1=15, mean=72, sd=10
Team B: n2=12, mean=65, sd=8

Given:
t0.95,25 = 1.708 (one-tailed)
t0.975,25 = 2.060 (two-tailed, 95% CI)

(a) Is there evidence at alpha=0.05 that Flow A has higher mean time than Flow B? [5]
(b) Build a 95% CI for (muA - muB) and interpret. [4]

Question 2 (9 marks)

A food-delivery firm tests two factors impacting delivery time (minutes):
- Routing: Old vs New
- Rider Incentive: Off vs On

Mean delivery times:

	Incentive Off	Incentive On	Row Mean
Old Routing	34	24	29
New Routing	22	30	26
Column Mean	28	27	27.5

ANOVA summary:

Source	SS	df	MS	F	p
Routing	8.0	1	8.0	1.60	0.23
Incentive	1.0	1	1.0	0.20	0.66
Routing×Incentive	45.0	1	45.0	9.00	0.01
Error	60.0	12	5.0		

(a) Which effects are significant at alpha=0.05? [4]
(b) Explain the business meaning of the interaction and what decision it implies. [5]

Question 3 (9 marks)

A demand series (units) for months 1..6:
120, 128, 133, 145, 150, 160

(a) Compute the 3-month moving average forecast for month 7. [3]
(b) Using simple exponential smoothing with alpha=0.3 and F1=120, compute F7. [5]
(c) If both trend and seasonality exist, which method should be preferred and why? [1]

Question 4 (8 marks)

```
-----------------------------------------
A clustering algorithm is used to segment 2D customer points.
Points: A(1,1), B(2,1), C(4,3), D(5,4)
Initial centroids: C1=(1,1), C2=(5,4)

(a) Assign each point to nearest centroid (Euclidean distance).                   [4]
(b) Compute updated centroids after assignment.                                   [4]
===============================================================================


===============================================================================
SET 14 — ANSWERS
===============================================================================

-----------------------------------------
Answer 1(a)
-----------------------------------------
Decoding cue: two independent groups, SDs differ -> use unequal-variance (Welch) t logic.

H0: muA - muB = 0
H1: muA - muB > 0

diff = 72 - 65 = 7

SE = sqrt( s1^2/n1 + s2^2/n2 )
   = sqrt(10^2/15 + 8^2/12)
   = sqrt(100/15 + 64/12)
   = sqrt(6.6667 + 5.3333)
   = sqrt(12.0000)
   = 3.4641

t = diff/SE = 7/3.4641 = 2.0207

Using provided df=25 critical values:
Critical one-tailed t0.95,25 = 1.708
Decision: 2.0207 > 1.708 -> Reject H0
Conclusion: evidence Flow A has higher mean time to first purchase.

-----------------------------------------
Answer 1(b)
-----------------------------------------
95% CI:
diff ± t0.975,25 * SE = 7 ± 2.060(3.4641)

Margin = 7.134
CI = (7 - 7.134, 7 + 7.134) = (-0.134, 14.134)

Interpretation: plausible mean difference ranges from near 0 up to ~14 minutes.
At 95% confidence, difference could be small; treat result with caution for sizing decisions.

-----------------------------------------
Answer 2(a)
-----------------------------------------
At alpha=0.05 using p-values:
Routing p=0.23 not significant
Incentive p=0.66 not significant
Routing×Incentive p=0.01 significant

Only interaction is significant.

-----------------------------------------
```

```
Answer 2(b)
-----------------------------------------
Interaction meaning: the effect of Incentive depends on Routing.

Old routing: Incentive On reduces time (24 vs 34) -> improvement of 10 minutes.
New routing: Incentive On increases time (30 vs 22) -> makes it worse by 8 minutes.

Decision implication:
Do NOT enable incentive blindly. Choose incentive policy conditional on routing:
- If Old routing -> turn Incentive On
- If New routing -> keep Incentive Off

-----------------------------------------
Answer 3(a)
-----------------------------------------
3-month MA for month 7 uses months 4,5,6:
F7 = (145 + 150 + 160)/3 = 455/3 = 151.67

-----------------------------------------
Answer 3(b)
-----------------------------------------
SES formula: F_{t+1} = alpha*A_t + (1-alpha)*F_t
alpha=0.3, F1=120

F2 = 0.3*120 + 0.7*120 = 120
F3 = 0.3*128 + 0.7*120 = 122.4
F4 = 0.3*133 + 0.7*122.4 = 125.58
F5 = 0.3*145 + 0.7*125.58 = 131.406
F6 = 0.3*150 + 0.7*131.406 = 136.984
F7 = 0.3*160 + 0.7*136.984 = 143.889

So F7 ≈ 143.89

-----------------------------------------
Answer 3(c)
-----------------------------------------
Use Holt-Winters (seasonal exponential smoothing) because it models level + trend + seasonalit

-----------------------------------------
Answer 4(a)
-----------------------------------------
Distances:

To C1(1,1):
A: 0
B: sqrt((2-1)^2+(1-1)^2)=1
C: sqrt((4-1)^2+(3-1)^2)=sqrt(13)=3.606
D: sqrt((5-1)^2+(4-1)^2)=5

To C2(5,4):
A: 5
B: sqrt(3^2+3^2)=4.243
C: sqrt(1^2+1^2)=1.414
D: 0

Assignments:
Cluster 1: A, B
Cluster 2: C, D

-----------------------------------------
```

```
Answer 4(b)
-----------------------------------------
Updated centroids are means of points in each cluster.

C1' = mean(A,B) = ((1+2)/2, (1+1)/2) = (1.5, 1.0)
C2' = mean(C,D) = ((4+5)/2, (3+4)/2) = (4.5, 3.5)
===============================================================================
```