

EPSOEXAMS

EPSO PREPARATION

AST3

MOCK EXAM 01

DATA PROCESSING, ANALYTICS AND STATISTICS



30 QUESTIONS | 40 MINUTES

MODERATE TO VERY DIFFICULT

DETAILED EXPLANATIONS FOR EVERY OPTION

Pattakou Eva - Pattakos Pantelis

epsoexams.eu / Independent practice material

Before you start

Data Processing, Analytics and Statistics

Your task

Answer 30 questions in 40 minutes. Select exactly one answer (A, B, C or D) for each question. Award one point per correct answer and zero for an incorrect or unanswered item. Maximum score: 30.

How to use the mock

Set a timer, use the answer grid below and keep the solutions closed until the timer ends. For numerical items, show your intermediate calculations on separate paper. If you are stuck, record your best answer, flag the item and return later.

Scope and difficulty

This is an independently authored practice test informed by the typical duties in the competition notice. It combines concepts, applied decisions, numerical reasoning where relevant and EU legal scenarios. Difficulty labels are editorial judgements, not calibrated EPSO ratings. This mock does not reproduce an official exam or guarantee a selection outcome.

YOUR ANSWERS

01 A B C D

02 A B C D

03 A B C D

04 A B C D

05 A B C D

06 A B C D

07 A B C D

08 A B C D

09 A B C D

10 A B C D

11 A B C D

12 A B C D

13 A B C D

14 A B C D

15 A B C D

16 A B C D

17 A B C D

18 A B C D

19 A B C D

20 A B C D

21 A B C D

22 A B C D

23 A B C D

24 A B C D

25 A B C D

26 A B C D

27 A B C D

28 A B C D

29 A B C D

30 A B C D

Questions

Data Processing, Analytics and Statistics

1. Two mutually exclusive groups contain 120 and 80 observations. Their means are 15 and 25 respectively. All observations have equal weight. What is the mean for the combined dataset?

- A. 21
 - B. 19
 - C. 40
 - D. 20
-

2. A SQL table contains four rows whose amount values are 10, NULL, 0 and 20. What does `SELECT COUNT(*), COUNT(amount), AVG(amount)` return?

- A. 4, 3, 10
 - B. 3, 3, 10
 - C. 4, 4, 7.5
 - D. 4, 2, 15
-

3. Before splitting a labelled dataset, an analyst estimates scaling means and variances using every row. No labels enter the calculation. Which correction is most appropriate?

- A. Fit preprocessing on training data and apply it unchanged to validation and test data.
 - B. Remove the scaled variables and evaluate using the original split.
 - C. Estimate separate scaling parameters on the test set before evaluation.
 - D. Keep the preprocessing because the target labels were not used.
-

4. Region A reports 180 incidents among 90,000 residents; region B reports 150 incidents among 50,000 residents. Which statement correctly compares incident rates?

- A. B has 300 per 100,000 residents and a rate 50% above A.
 - B. A has 200 per 100,000 residents and a rate 20% above B.
 - C. A and B have equal rates because both populations exceed the counts.
 - D. B has 150 per 100,000 residents and a rate 25% below A.
-

Questions

Data Processing, Analytics and Statistics

5. For an approximately normal estimator, an estimate is 72 and its standard error is 3. Use 1.96 as the critical value. What is the approximate 95% confidence interval?

- A. 70.04 to 73.96
 - B. 69.00 to 75.00
 - C. 66.12 to 77.88
 - D. 60.24 to 83.76
-

6. An API returns case identifiers such as 000731. The specification treats them as fixed-length identifiers, not quantities. Which ingestion design best preserves meaning?

- A. Store them as integers and discard their original representation.
 - B. Remove leading zeros before checking whether identifiers are unique.
 - C. Convert them to floating-point values to allow future decimals.
 - D. Store them as strings and validate the required length.
-

7. A production notebook must be rerun by another team after a dependency update. Which evidence is most useful for reproducing the original result?

- A. The latest input data and whichever dependencies now install successfully.
 - B. The notebook filename, last edit date and final output total.
 - C. The code revision, input-data version, dependency versions and relevant random seeds.
 - D. The chart image, analyst's name and current package list.
-

8. A pandas merge joins on case_id. The left table has two rows with case_id = 7; the right table has three rows with case_id = 7. No other rows exist. How many rows does an inner merge produce?

- A. 3
 - B. 6
 - C. 2
 - D. 5
-

Questions

Data Processing, Analytics and Statistics

9. A population has 800 urban and 200 rural households. A sample has 80 urban households with mean consumption 12 and 40 rural households with mean 30. What is the population-composition-weighted estimate of mean consumption?

- A. 18.0
 - B. 21.0
 - C. 15.6
 - D. 26.4
-

10. A price index using the old base is 125 in 2024 and 150 in 2025. It is rebased so that 2024 equals 100, without changing the underlying series. What is the 2025 value and the growth rate from 2024?

- A. 120 and 20%
 - B. 125 and 25%
 - C. 120 and 25%
 - D. 150 and 50%
-

11. A screening test has 90% sensitivity and 95% specificity in a population where prevalence is 2%. For 10,000 people, use expected counts. Approximately what share of positive results are true positives?

- A. 64.3%
 - B. 95.0%
 - C. 26.9%
 - D. 90.0%
-

12. All applications must remain in a result, with only approved payments attached where available. Which SQL pattern correctly implements that requirement?

- A. `LEFT JOIN payments p ON a.id = p.application_id WHERE p.status = 'approved'`
 - B. `LEFT JOIN payments p ON a.id = p.application_id AND p.status = 'approved'`
 - C. `LEFT JOIN payments p ON a.id = p.application_id WHERE p.status IS NOT NULL`
 - D. `INNER JOIN payments p ON a.id = p.application_id AND p.status = 'approved'`
-

Questions

Data Processing, Analytics and Statistics

13. An exclusion subquery returns identifiers 2 and NULL. A non-null candidate identifier is 5. Under standard SQL three-valued logic, why might WHERE id NOT IN (subquery) fail to keep that candidate?

- A. The NULL comparison makes the predicate unknown, which WHERE does not retain.
 - B. The candidate is retained because 5 differs from the only non-null value.
 - C. NOT IN treats every NULL as the integer zero.
 - D. NOT IN removes only values appearing at least twice in the subquery.
-

14. A table has scores 90, 90 and 80. Evaluate DENSE_RANK() OVER (ORDER BY score DESC). Which output is correct in that score order?

- A. 1, 1, 3
 - B. 1, 2, 3
 - C. 2, 2, 3
 - D. 1, 1, 2
-

15. Programme P records 81 successes out of 90 easy cases and 2 out of 10 difficult cases. Programme Q records 19 out of 20 easy cases and 24 out of 80 difficult cases. Which interpretation is justified?

- A. Equal totals of 100 cases make case-mix adjustment unnecessary.
 - B. Q performs better in each stratum, while P has the higher aggregate rate because case mix differs.
 - C. P performs better in each stratum and in the aggregate.
 - D. Q performs better in each stratum and therefore must have the higher aggregate rate.
-

16. A validation set contains 100 actual positive cases. At threshold 0.5, TP = 60 and FP = 15. At threshold 0.3, TP = 80 and FP = 40. Which change is correct?

- A. Recall and precision both rise by 20 percentage points.
 - B. Recall rises from 80% to 100%, while precision remains 80%.
 - C. Recall remains 60%, while precision rises from 66.7% to 80%.
 - D. Recall rises from 60% to 80%, while precision falls from 80% to about 66.7%.
-

Questions

Data Processing, Analytics and Statistics

17. Missing income is replaced with zero before calculating a mean. Missingness is much more common among high-income respondents. Which claim is most defensible?

- A. Zero replacement guarantees an unbiased mean because every row remains.
 - B. Deleting all missing rows guarantees removal of the bias.
 - C. The mean may be biased downward; document and assess missingness before choosing treatment.
 - D. The missingness pattern cannot affect the mean if sample size is large.
-

18. A fitted logistic model has an odds ratio of 2 for a binary feature. At feature value 0, the estimated event probability is 0.20. Holding other covariates fixed, what is the probability at feature value 1?

- A. 1/3
 - B. 0.40
 - C. 0.25
 - D. 0.60
-

19. A table publishes counts for two categories and their total. A sensitive category is suppressed, but the other category count and total remain visible. What is the principal disclosure problem?

- A. The remaining values become statistically independent of the hidden count.
 - B. The total must be retained because it never reveals component counts.
 - C. Suppression guarantees anonymity whenever a cell is blank.
 - D. The suppressed count can be reconstructed by subtraction.
-

20. Five-fold cross-validation evaluates a pipeline containing imputation and feature selection. Which workflow best avoids optimistic evaluation?

- A. Fit feature selection on all labels, but imputation within each fold.
 - B. Fit both steps once on the full dataset before forming folds.
 - C. Fit preprocessing on each validation fold before predicting that fold.
 - D. Fit imputation and feature selection separately inside each training fold.
-

Questions

Data Processing, Analytics and Statistics

21. A large voluntary survey excludes people without internet access. Increasing the sample from 20,000 to 100,000 respondents using the same channel will primarily do what?

- A. Correct non-response by treating every missing person as a zero outcome.
 - B. Eliminate coverage bias once the sample is sufficiently large.
 - C. Reduce some sampling variability without automatically correcting coverage bias.
 - D. Make the sample representative because five times as many people respond.
-

22. A public EU dataset can be downloaded, but it includes photographs licensed from third parties. Which conclusion should guide a reuse request?

- A. Public access automatically transfers copyright to every downloader.
 - B. Check the relevant reuse framework and third-party rights before authorising the photographs' reuse.
 - C. EU-origin data are exempt from all third-party intellectual-property restrictions.
 - D. Removing the photographer's name makes commercial reuse unrestricted.
-

23. An anti-fraud model predicts risk when an application arrives. One proposed feature is whether an investigator later confirms fraud. What is the strongest objection?

- A. The feature is unusable only because binary features cannot enter risk models.
 - B. The feature is unavailable at prediction time and leaks later outcome information.
 - C. The feature is valid if it raises cross-validation accuracy.
 - D. The feature should be retained and replaced by zero for every live case.
-

24. A statistical release revises last year's denominator while leaving this year's numerator unchanged. A dashboard combines the revised numerator table with an old denominator cache. What is the best first response?

- A. Round the new percentage to fewer decimals to hide minor differences.
 - B. Align the input vintages and rerun validation before interpreting the reported change.
 - C. Assume revisions cannot affect ratios because the numerator is unchanged.
 - D. Replace the denominator with this year's numerator to ensure consistency.
-

Questions

Data Processing, Analytics and Statistics

25. Independent observations have unchanged variance and the same sampling design. A mean estimated from 400 observations has standard error 6. What standard error is expected from 900 observations?

- A. 9
 - B. 2.67
 - C. 6
 - D. 4
-

26. The same 60 people are measured before and after an intervention. A researcher treats the measurements as two unrelated samples. Which issue should be addressed first?

- A. Use the within-person pairing when assessing changes and their uncertainty.
 - B. Delete one measurement from every person to remove dependence.
 - C. Double the sample size to 120 independent people without altering the design.
 - D. Assume a difference in averages proves the intervention caused the change.
-

27. In R, `x <- c(2, NA, 8)`. Consider `mean(x)` and `mean(x, na.rm = TRUE)`. Which pair of results is correct?

- A. 5 and 5
 - B. NA and 10/3
 - C. 0 and 5
 - D. NA and 5
-

28. For ordinary administrative processing of staff personal data by an EU institution, which EU data-protection framework is specifically designed for that institutional processing?

- A. Directive 2019/1024 as the staff-data processing code
 - B. Regulation (EU) 2016/679 alone, with no institutional regime
 - C. Regulation (EU) 2018/1725
 - D. Regulation (EC) 1049/2001 as a complete substitute for data-protection law
-

Questions

Data Processing, Analytics and Statistics

29. An approved AI assistant summarises external documents. One PDF includes an instruction to ignore the user's task and reveal internal files. Which design addresses the relevant risk most directly?

- A. Treat document text as untrusted data and enforce tool-access limits independently of model output.
 - B. Give the assistant broader file access so it can decide what is confidential.
 - C. Increase temperature so the assistant becomes less obedient to the PDF.
 - D. Move the PDF instruction above the user's message so it is processed first.
-

30. A provider changes a field from whole euros to euro cents without changing its name. The schema still validates as an integer, and reported expenditure rises roughly 100-fold. Which control most directly detects the defect?

- A. Accept the values because the data type remains unchanged.
 - B. Validate semantic units and magnitude ranges alongside the structural schema.
 - C. Divide every future release by 100 without consulting the provider.
 - D. Cast every integer to text before calculating expenditure.
-

Answer key

Data Processing, Analytics and Statistics

Stop here until you have finished the timed test.

01	B	11	C	21	C
02	A	12	B	22	B
03	A	13	A	23	B
04	A	14	D	24	B
05	C	15	B	25	D
06	D	16	D	26	A
07	C	17	C	27	D
08	B	18	A	28	C
09	C	19	D	29	A
10	A	20	D	30	B

Score: ____ / 30

Review every explanation, including answers you guessed correctly. Record whether each error came from a knowledge gap, misreading, calculation or time pressure. Revise the relevant topic before attempting another timed mock.

Solutions and explanations

Data Processing, Analytics and Statistics

1. Correct answer: B

Difficulty: Moderate

- A. Incorrect. 21 reverses the group-size weights.
 - B. Correct. $(120 \times 15 + 80 \times 25) / 200 = 3,800 / 200 = 19$. The group sizes are the weights.
 - C. Incorrect. 40 adds the group means rather than combining their totals.
 - D. Incorrect. 20 gives equal weight to the two groups despite their different sizes.
-

2. Correct answer: A

Difficulty: Moderate

- A. Correct. COUNT(*) includes every row. COUNT(amount) and AVG(amount) ignore NULL; zero remains a valid value. The mean is $30 / 3 = 10$.
 - B. Incorrect. COUNT(*) also counts the row containing NULL.
 - C. Incorrect. NULL is not counted as a non-null value or included as zero in AVG.
 - D. Incorrect. Zero is not a missing value and must be counted.
-

3. Correct answer: A

Difficulty: Moderate

- A. Correct. The evaluation data influenced the fitted transformation. Fitting preprocessing on training data prevents this information leakage while preserving a common transformation.
 - B. Incorrect. Deleting variables is unnecessary and changes the intended modelling problem.
 - C. Incorrect. Test-specific fitting still learns from the held-out evaluation data.
 - D. Incorrect. Leakage can occur through feature statistics without using labels.
-

Solutions and explanations

Data Processing, Analytics and Statistics

4. Correct answer: A

Difficulty: Moderate

A. Correct. $A: 180 / 90,000 \times 100,000 = 200$. $B: 150 / 50,000 \times 100,000 = 300$. Relative excess: $(300 - 200) / 200 = 50\%$.

B. Incorrect. A has the lower rate; absolute incident counts do not determine the comparison.

C. Incorrect. Different count-to-population ratios do not become equal because populations are large.

D. Incorrect. B's population denominator produces 300, not 150.

5. Correct answer: C

Difficulty: Moderate

A. Incorrect. The critical value must be multiplied by the standard error.

B. Incorrect. A one-standard-error margin is not the specified 95% margin.

C. Correct. The margin is $1.96 \times 3 = 5.88$. Subtracting and adding it to 72 gives 66.12 and 77.88.

D. Incorrect. This doubles the correct margin of error.

6. Correct answer: D

Difficulty: Moderate

A. Incorrect. Integer conversion loses the specified representation.

B. Incorrect. Normalising without authority may merge identifiers that should remain distinct.

C. Incorrect. Floating-point storage introduces irrelevant numeric behaviour.

D. Correct. A string preserves leading zeros and supports the specified length check. Identifiers represent identity, not arithmetic magnitude.

Solutions and explanations

Data Processing, Analytics and Statistics

7. Correct answer: C

Difficulty: Moderate

- A. Incorrect. Latest versions may produce a different computation from the original run.
 - B. Incorrect. Names and timestamps do not reconstruct the computation.
 - C. Correct. Reproduction requires the actual code, data and environment used, plus seeds where stochastic behaviour matters. A final image or total is not a sufficient execution specification.
 - D. Incorrect. A chart and current environment do not identify the original inputs or dependencies.
-

8. Correct answer: B

Difficulty: Moderate

- A. Incorrect. It does not automatically retain only the right-table count.
 - B. Correct. Every left match pairs with every right match: $2 \times 3 = 6$. Repeated keys cause a many-to-many expansion.
 - C. Incorrect. An inner merge does not automatically retain only the left-table count.
 - D. Incorrect. Matching rows form combinations; the counts are not added.
-

9. Correct answer: C

Difficulty: Hard

- A. Incorrect. 18 weights by the sample's 80:40 composition instead of the population's 800:200.
 - B. Incorrect. 21 gives each stratum equal weight.
 - C. Correct. Population shares are 0.8 and 0.2: $0.8 \times 12 + 0.2 \times 30 = 15.6$. The sample deliberately overrepresents rural households.
 - D. Incorrect. 26.4 reverses the urban and rural population weights.
-

Solutions and explanations

Data Processing, Analytics and Statistics

10. Correct answer: A

Difficulty: Hard

- A. Correct. The rebased value is $150 / 125 \times 100 = 120$. Growth is $(150 - 125) / 125 = 20\%$, unchanged by rebasing.
- B. Incorrect. The 2024 old-base value is not the 2025 rebased value.
- C. Incorrect. 25 is the old index-point difference, not the percentage change.
- D. Incorrect. 150 remains on the old base and 50% compares it with the wrong denominator.
-

11. Correct answer: C

Difficulty: Very difficult

- A. Incorrect. This value is not supported by the prevalence-adjusted expected counts.
- B. Incorrect. Specificity conditions on actual negatives.
- C. Correct. There are 200 actual positives: 180 test positive. Of 9,800 actual negatives, 490 test positive. Positive predictive value is $180 / 670 \approx 26.9\%$.
- D. Incorrect. Sensitivity conditions on actual positives, not on positive test results.
-

12. Correct answer: B

Difficulty: Hard

- A. Incorrect. The WHERE filter rejects NULL-extended rows and loses unmatched applications.
- B. Correct. Filtering matches in the ON clause preserves applications without an approved match as NULL-extended rows. It does not filter those applications out of the result.
- C. Incorrect. This removes unmatched applications and allows statuses other than approved.
- D. Incorrect. An inner join cannot preserve applications with no qualifying payment.
-

Solutions and explanations

Data Processing, Analytics and Statistics

13. Correct answer: A

Difficulty: Very difficult

- A. Correct. For 5, NOT IN requires the conjunction of $5 \neq 2$ and $5 \neq \text{NULL}$. The second comparison is unknown; WHERE retains only true predicates. A suitably correlated NOT EXISTS avoids this trap.
- B. Incorrect. The NULL introduces an unknown comparison even when non-null values differ.
- C. Incorrect. NULL is not implicitly the integer zero.
- D. Incorrect. Multiplicity is not the rule defining NOT IN.
-

14. Correct answer: D

Difficulty: Hard

- A. Incorrect. 1, 1, 3 describes RANK, which leaves a gap after a tie.
- B. Incorrect. Distinct sequential values describe row numbering, not dense ranking.
- C. Incorrect. Dense ranking starts at 1 for the first group.
- D. Correct. DENSE_RANK assigns the same rank to ties and leaves no gap between distinct score groups. The two 90s share rank 1; 80 receives rank 2.
-

15. Correct answer: B

Difficulty: Very difficult

- A. Incorrect. Equal overall sample sizes do not imply equal case composition.
- B. Correct. Easy: P 90%, Q 95%; difficult: P 20%, Q 30%. Aggregate: P 83%, Q 43%. The reversal reflects different stratum weights, illustrating Simpson's paradox.
- C. Incorrect. P's within-stratum rates are lower in both strata.
- D. Incorrect. Different mixture weights can reverse the aggregate comparison.
-

Solutions and explanations

Data Processing, Analytics and Statistics

16. Correct answer: D

Difficulty: Hard

- A. Incorrect. The numerical precision calculation shows a decline, not an increase.
 - B. Incorrect. The stated TP counts give recall of 60% and 80%.
 - C. Incorrect. The threshold change increases recall and reduces precision in these data.
 - D. Correct. Recall is $TP / 100$. Precision is $TP / (TP + FP)$: $60 / 75 = 80\%$, then $80 / 120 \approx 66.7\%$. Lowering the threshold finds more positives but also more false positives.
-

17. Correct answer: C

Difficulty: Hard

- A. Incorrect. Row retention does not make invented zero incomes unbiased.
 - B. Incorrect. Complete-case deletion can preserve selection bias.
 - C. Correct. Replacing missing high incomes with zero can depress the estimate. Neither retaining rows nor increasing sample size corrects systematic missingness; treatment requires assumptions and sensitivity checks.
 - D. Incorrect. Large samples do not eliminate systematic non-response bias.
-

18. Correct answer: A

Difficulty: Very difficult

- A. Correct. Baseline odds are $0.20 / 0.80 = 0.25$. Doubling gives 0.50; probability is $0.50 / 1.50 = 1/3$. An odds ratio multiplies odds, not probability.
 - B. Incorrect. Doubling probability confuses probability with odds.
 - C. Incorrect. 0.25 is the baseline odds, not the new probability.
 - D. Incorrect. 0.60 does not follow from the stated odds transformation.
-

Solutions and explanations

Data Processing, Analytics and Statistics

19. Correct answer: D

Difficulty: Very difficult

- A. Incorrect. The accounting identity links all three values directly.
 - B. Incorrect. A total combined with a complementary count can reveal the component.
 - C. Incorrect. A blank cell does not prevent reconstruction from other published values.
 - D. Correct. If $\text{total} = \text{hidden category} + \text{visible category}$, subtraction reveals the hidden count. Disclosure control must consider relationships between released values, including complementary suppression where appropriate.
-

20. Correct answer: D

Difficulty: Very difficult

- A. Incorrect. Feature selection using held-out labels leaks target information.
 - B. Incorrect. Full-data fitting allows validation data to influence the pipeline.
 - C. Incorrect. Fitting on validation rows makes evaluation data part of training.
 - D. Correct. Every fitted preprocessing or selection step must learn only from the training portion of each fold. Validation rows should be transformed using those fitted training steps.
-

21. Correct answer: C

Difficulty: Hard

- A. Incorrect. Zero substitution does not supply the unobserved people's actual outcomes.
 - B. Incorrect. Size does not restore the missing population coverage.
 - C. Correct. Sample size can improve precision within the observed sampling mechanism, but excluded groups remain excluded. Coverage and self-selection require changes to design or justified adjustment.
 - D. Incorrect. More respondents from the same biased channel need not represent the population.
-

Solutions and explanations

Data Processing, Analytics and Statistics

22. Correct answer: B

Difficulty: Hard

- A. Incorrect. Downloading does not transfer ownership of copyright.
 - B. Correct. Access and reuse are different questions. Third-party rights need their own assessment; making a dataset public does not transfer those rights to users.
 - C. Incorrect. An EU source does not extinguish third-party rights.
 - D. Incorrect. Removing attribution does not create permission to reuse.
-

23. Correct answer: B

Difficulty: Very difficult

- A. Incorrect. Binary variables can be valid predictors when appropriately available.
 - B. Correct. A feature reflecting a later investigation cannot support the stated decision-time prediction. Its apparent predictive power results from future information and may also encode the target.
 - C. Incorrect. Validation can look excellent despite a prediction-time availability error.
 - D. Incorrect. A constant live replacement creates severe train-serve mismatch rather than solving leakage.
-

24. Correct answer: B

Difficulty: Hard

- A. Incorrect. Rounding does not correct incompatible inputs.
 - B. Correct. A ratio depends on both components. Mixing incompatible release vintages can distort comparisons; align versions and document the revision before interpreting results.
 - C. Incorrect. Changing the denominator can change the ratio materially.
 - D. Incorrect. Using the numerator as denominator destroys the meaning of the measure.
-

Solutions and explanations

Data Processing, Analytics and Statistics

25. Correct answer: D

Difficulty: Hard

- A. Incorrect. Larger samples reduce, rather than increase, this standard error.
 - B. Incorrect. 2.67 uses the sample-size ratio rather than its square root.
 - C. Incorrect. The standard error changes even when the population variance stays fixed.
 - D. Correct. For the mean, standard error scales as $1 / \sqrt{n}$. The new value is $6 \times \sqrt{400 / 900} = 6 \times 2/3 = 4$.
-

26. Correct answer: A

Difficulty: Hard

- A. Correct. The observations are paired. Within-person differences preserve the design and account for the covariance between measurements; causal interpretation still needs a justified design.
 - B. Incorrect. Removing one time point prevents the intended change analysis.
 - C. Incorrect. There are 60 people, not 120 independent participants.
 - D. Incorrect. A before-after difference alone does not establish causality.
-

27. Correct answer: D

Difficulty: Moderate

- A. Incorrect. The first expression does not remove NA by default.
 - B. Incorrect. The second expression uses two non-missing observations, not three.
 - C. Incorrect. An NA is not implicitly converted into zero.
 - D. Correct. The default mean returns NA when NA is present. Removing missing values leaves 2 and 8, whose mean is $10 / 2 = 5$.
-

Solutions and explanations

Data Processing, Analytics and Statistics

28. Correct answer: C

Difficulty: Hard

- A. Incorrect. Open-data reuse legislation is not the staff-data processing regime.
 - B. Incorrect. The question concerns the specific EU institutional regime.
 - C. Correct. Regulation 2018/1725 specifically governs personal-data processing by Union institutions, bodies, offices and agencies within its scope. It must be distinguished from the GDPR framework.
 - D. Incorrect. Public-access rules are not a substitute for the institutional data-protection framework.
-

29. Correct answer: A

Difficulty: Very difficult

- A. Correct. This is an indirect prompt-injection attempt. Separation of instructions and external data helps, while independently enforced permissions prevent text from granting unauthorised access.
 - B. Incorrect. Broader permissions increase the impact of an injection.
 - C. Incorrect. Temperature is not an access-control mechanism.
 - D. Incorrect. Giving the malicious text instruction priority worsens the vulnerability.
-

30. Correct answer: B

Difficulty: Very difficult

- A. Incorrect. The same type can encode a different unit and meaning.
 - B. Correct. Type validation alone does not capture units. Metadata and plausibility checks should detect the semantic change; correction must be confirmed and documented with the provider.
 - C. Incorrect. A blind permanent correction may mis-handle subsequent releases.
 - D. Incorrect. Casting does not establish or correct the unit.
-

Sources and study follow-up

Data Processing, Analytics and Statistics

Legal and technical reference points

The scenarios and numerical data are original educational examples. The references below support the legal rules and technical concepts tested; they are not an official EPSO reading list. Check the applicable consolidated legislation and its scope. Internal Commission rules do not automatically bind every EU institution.

Competition notice EPSO/AST/158/26

<https://eur-lex.europa.eu/eli/C/2026/4963/oj/eng>

EU institutions data protection - Regulation 2018/1725

<https://eur-lex.europa.eu/eli/reg/2018/1725/oj/eng>

PostgreSQL: aggregates

<https://www.postgresql.org/docs/current/functions-aggregate.html>

PostgreSQL: subquery expressions

<https://www.postgresql.org/docs/current/functions-subquery.html>

PostgreSQL: window functions

<https://www.postgresql.org/docs/current/functions-window.html>

scikit-learn: common pitfalls and leakage

https://scikit-learn.org/stable/common_pitfalls.html

R: mean function

<https://stat.ethz.ch/R-manual/R-devel/library/base/html/mean.html>

Reuse of Commission documents - Decision 2011/833/EU

<https://eur-lex.europa.eu/eli/dec/2011/833/oj/eng>