Ce document a été mis en ligne par l'organisme FormaV®.

Toute reproduction, représentation ou diffusion, même partielle, sans autorisation préalable, est strictement interdite.

Pour en savoir plus sur nos formations disponibles, veuillez visiter :
formav.co/explorer

BUT SD

BTS Science des données, C6 : Analyse exploratoire

Session 2026 • Code sujet : **26-BUT-SD-EP1** • Durée : **2 h 30** • Coefficient : **3,0 (indicatif)**

Sommaire

Cliquez sur une partie pour y accéder directement

1. **Cadrage de l'analyse et contrôle des données** (3 questions) — **Page 3**
2. **Description statistique et choix des représentations** (3 questions) — **Page 5**
3. **Analyse exploratoire, diagnostic et communication** (3 questions) — **Page 7**
4. **Préparation du traitement et prévention des risques** (2 questions) — **Page 9**
5. **Ce qu'il faut retenir & Questions fréquentes (FAQ)** (Synthèse & Q&A Examen) — **Page 11**

Sujet officiel (PDF) : https://butsd.fr/assets/annales/but-sd/C6 - Mettre en œuvre une analyse exploratoire/2026/BUT Science des Données - C6 - Mettre en œuvre une analyse exploratoire - Session 2026 - 1-Sujet.pdf

1. Cadrage de l’analyse et contrôle des données

L’Observatoire des formations de l’Institut Data Loire étudie les résultats de dix étudiants à une évaluation de statistiques, à partir de données pseudonymisées sur le temps de révision, les scores, les absences et les modalités. Il faut définir le cadre de l’étude, vérifier la qualité des données et préparer une communication prudente.

Définition du besoin

Question 1.1. Formulez l’objectif statistique de l’étude, identifiez la population effectivement observée et l’unité statistique, puis expliquez pourquoi les résultats ne permettent pas, à eux seuls, de généraliser à l’ensemble des étudiants de l’établissement.

L’objectif est de décrire les scores et les autres variables du groupe, puis d’explorer l’association entre le temps de révision déclaré et le score obtenu. La population effectivement observée est constituée des dix étudiants du jeu fourni; l’unité statistique est un étudiant. Ces résultats décrivent ce groupe précis. Le faible effectif et l’absence d’indication d’un échantillonnage représentatif ne permettent pas d’étendre les constats aux autres étudiants de l’établissement.

> ✓ **RÉPONSE À RETENIR**
>
> Étude descriptive des dix étudiants observés, avec une observation par étudiant; aucune généralisation à l’établissement n’est justifiée.

Contrôles de structure et de domaine

Question 1.2. À partir des règles du DT 2, indiquez les contrôles à appliquer au fichier et leurs résultats sur les dix lignes fournies : nombre d’observations, valeurs manquantes, identifiants dupliqués et valeurs hors domaine pour le score et les absences.

Contrôle	Résultat sur le fichier
Structure et valeurs manquantes	10 lignes, une par étudiant; aucune valeur manquante.
Identifiants	10 identifiants non vides et distincts; aucun doublon.
Score	Valeurs numériques comprises entre 42 et 78, toutes dans [0; 100].
Absences	Valeurs entières de 0 à 2, toutes positives ou nulles.
Autres domaines	Temps renseignés et positifs; modalités toutes Présentiel ou Distance.

Les contrôles portent sur le nombre de lignes, la complétude, l’unicité des identifiants, les types et les domaines attendus. Les dix observations respectent ces contrôles : aucun score ne sort de l’intervalle autorisé et aucune absence n’est négative ou non entière. Le temps et les modalités sont également conformes aux règles du dictionnaire.

✓ RÉPONSE À RETENIR

10 observations, aucune valeur manquante ni aucun identifiant dupliqué; aucun score ou nombre d’absences hors domaine.

Précautions de diffusion

Question 1.3. Indiquez deux précautions à prendre avant de transmettre les résultats du tableau de bord à un destinataire. Prenez en compte la pseudonymisation et le risque d’interprétation abusive d’une association statistique.

Avant la transmission du tableau de bord, appliquer deux précautions distinctes :

- **Confidentialité :** limiter l’accès aux destinataires autorisés, sécuriser le fichier et ne diffuser que des résultats agrégés utiles. Les identifiants pseudonymisés restent des données à protéger.
- **Interprétation :** présenter le lien observé entre temps de révision et score comme une association descriptive, sans affirmer un effet causal. Rappeler le faible effectif et vérifier que les unités, intitulés et graphiques correspondent aux calculs.

✓ RÉPONSE À RETENIR

Protéger et minimiser les données diffusées; présenter l’association observée sans conclusion causale.

2. Description statistique et choix des représentations

Les dix observations du DT 1 fournissent les scores et les temps de révision, également représentés dans un nuage de points. Il faut calculer les indicateurs du score selon la convention de quartiles imposée, examiner les clôtures de Tukey et décrire la relation entre les deux variables.

Résumé statistique du score

Question 2.1. Complétez le tableau de résumé statistique pour la variable score : effectif, minimum, médiane, Q1, Q3, maximum, moyenne et écart interquartile. Présentez les étapes essentielles du calcul de la moyenne, de la médiane, de Q1, de Q3 et de l'écart interquartile en appliquant la convention indiquée dans le sujet.

Scores triés : 42, 45, 48, 50, 55, 61, 63, 68, 72, 78. Les deux valeurs centrales sont 55 et 61, donc la médiane vaut $(55 + 61)/2 = 58$. Les moitiés ordonnées donnent $Q_1 = 48$ et $Q_3 = 68$. La somme des dix scores est 582, d'où une moyenne de $582/10 = 58,20$. Enfin, l'écart interquartile est $Q_3 - Q_1 = 68 - 48 = 20$.

Indicateur	Valeur
Effectif	10
Minimum	42
Médiane	58
Q1	48
Q3	68
Maximum	78
Moyenne	58,20
Écart interquartile	20

Tableau de résumé statistique complété

Résumé statistique à compléter

Indicateur	Valeur
Effectif	10
Minimum	42
Médiane	58
Q1	48
Q3	68
Maximum	78
Moyenne	58,20
Écart interquartile	20

✓ RÉPONSE À RETENIR

Effectif : 10 ; minimum : 42 ; médiane : 58 ; Q1 : 48 ; Q3 : 68 ; maximum : 78 ; moyenne : 58,20 ; écart interquartile : 20.

Clôtures de Tukey

Question 2.2. Déterminez les clôtures de Tukey pour les scores et indiquez, en justifiant, si une observation est signalée par cette règle. Précisez qu'une observation non signalée ne doit pas être considérée automatiquement comme exempte de toute erreur.

Avec $Q_1 = 48$, $Q_3 = 68$ et $IQR = 20$, les clôtures sont $48 - 1,5 \times 20 = 18$ et $68 + 1,5 \times 20 = 98$. La règle signale les scores strictement inférieurs à 18 ou supérieurs à 98. Or les scores observés vont de 42 à 78 : aucun ne franchit ces limites. L'absence de signalement par Tukey ne prouve pas qu'une donnée est exacte ; toute valeur douteuse doit être vérifiée auprès de sa source.

✓ RÉPONSE À RETENIR

Clôtures de Tukey : [18 ; 98]. Aucun score n'est signalé.

Lecture du nuage de points

Question 2.3. Décrivez la forme générale du nuage de points et la situation de l'observation J par rapport aux autres observations. Proposez une vérification à effectuer avant de qualifier J d'erreur ou de la supprimer, puis précisez si le nuage établit un lien de causalité.

Le nuage présente une tendance globalement croissante : pour les observations A à I, les scores augmentent avec le temps de révision. J se distingue avec 10 heures et 45 points : son score est inférieur à ceux de B à I, mais supérieur à celui de A. Avant de la qualifier d'erreur ou de la supprimer, vérifier auprès de la source la saisie des deux valeurs et leurs unités. Le nuage décrit une association, sans établir de lien de causalité.

✓ RÉPONSE À RETENIR

Association globalement croissante ; J (10 h, 45 points) s'écarte de la tendance et doit être vérifiée. Aucune causalité n'est établie.

3. Analyse exploratoire, diagnostic et communication

Le responsable de formation attend une note de synthèse fondée uniquement sur les dix observations fournies. Il faut choisir des représentations adaptées, compléter la grille de diagnostic en distinguant constats et limites, puis rédiger une conclusion professionnelle qui traite prudemment l'observation J.

Choix des représentations

Question 3.1. Choisissez une représentation adaptée pour décrire la distribution des scores et une représentation adaptée pour étudier la relation entre le temps de révision et le score. Justifiez brièvement chaque choix en fonction du type des variables.

Le score est une variable quantitative : une boîte à moustaches résume sa position, sa dispersion et ses valeurs extrêmes ; un histogramme montre la forme de sa distribution. Le temps de révision et le score sont deux variables quantitatives : un nuage de points permet de décrire leur association et de repérer l'écart de J. Ces graphiques ne permettent pas de conclure à une causalité.

> ✓ **RÉPONSE À RETENIR**
>
> Distribution des scores : boîte à moustaches ou histogramme. Relation entre deux variables quantitatives : nuage de points.

Grille de diagnostic DR 1

Question 3.2. Complétez la grille DR 1 avec, pour chacun des axes « Distribution des scores », « Temps de révision et score » et « Qualité des données », un constat chiffré ou directement vérifiable et une limite ou vérification pertinente. Pour l'axe « Temps de révision et score », tenez compte de l'observation J.

Les scores vont de 42 à 78 points et leur médiane est de 58. Les identifiants A à J sont distincts ; les dix lignes sont renseignées et les scores, absences, temps et modalités respectent les domaines attendus. Le nuage montre une hausse générale des scores avec le temps de révision pour A à I, tandis que J s'écarte de cette tendance.

Axe	Constat appuyé par une donnée	Limite ou vérification nécessaire
Distribution des scores	Médiane : 58 ; minimum : 42 ; maximum : 78.	Effectif de 10 : vérifier les valeurs à la source.
Temps de révision et score	Hausse générale ; J : 10 h et 45 points.	Vérifier les données de J ; une association ne prouve pas une causalité.
Qualité des données	10 identifiants distincts ; aucune valeur manquante ou hors domaine.	Avant diffusion, contrôler unités, modalités et destinataires.

Les constats décrivent seulement les dix observations fournies ; l'observation J doit être vérifiée avant toute correction ou exclusion.

DR 1 complété

Axe de diagnostic	Constat appuyé par une donnée	Limite ou vérification nécessaire
Distribution des scores	Médiane 58 ; min. 42 ; max. 78	n = 10 ; vérifier les scores à la source
Temps de révision et score	Hausse générale ; J : 10 h, 45 points	Vérifier J ; pas de causalité
Qualité des données	10 identifiants distincts ; aucune valeur manqu	Contrôler unités, modalités et destinataires

Un constat et une limite ou vérification sont attendus dans chaque ligne.

✓ RÉPONSE À RETENIR

Scores : médiane 58, minimum 42, maximum 78. J : 10 h et 45 points, à vérifier. Les constats concernent uniquement les dix observations.

Conclusion professionnelle

Question 3.3. Rédigez une conclusion professionnelle de trois à cinq phrases destinée au responsable. Elle doit comporter un résultat descriptif, une réserve sur l'observation J et une limite d'interprétation liée à l'effectif ou au caractère observationnel des données.

Dans le groupe observé, les scores vont de 42 à 78 points et la médiane est de 58. Les scores augmentent généralement avec le temps de révision déclaré pour les observations A à I. J s'écarte de cette tendance ; il convient de vérifier son temps et son score à la source avant toute correction ou exclusion. L'effectif de dix étudiants et le caractère observationnel des données limitent l'interprétation : aucune causalité ni généralisation à l'établissement ne peut être établie.

✓ RÉPONSE À RETENIR

La médiane des scores est de 58/100. J doit être vérifié ; ces dix observations ne permettent ni causalité ni généralisation à l'établissement.

4. Préparation du traitement et prévention des risques

Avant de produire un tableau de bord, il faut définir un traitement reproductible depuis la réception du fichier jusqu'à la validation de la restitution. Le candidat doit également identifier des risques liés à la confidentialité ou à l'interprétation des résultats et proposer des mesures de prévention adaptées.

Protocole opératoire DR 2

Question 4.1. Complétez le protocole opératoire DR 2 en quatre étapes chronologiques, de la réception du fichier à la validation de la restitution. Pour chaque étape, indiquez au moins un contrôle observable permettant de la valider et de rendre l'analyse reproductible par une autre personne.

1. **Réception et importation :** sécuriser le fichier source, puis importer les colonnes prévues. Contrôle : les cinq variables attendues sont présentes et lisibles.
2. **Contrôle des données :** vérifier types, unités, valeurs manquantes, identifiants et domaines. Contrôle : consigner le bilan des dix lignes et toute anomalie.
3. **Analyse :** calculer les indicateurs selon les conventions du sujet et produire les représentations adaptées. Contrôle : conserver les paramètres et vérifier la cohérence des résultats avec les données.
4. **Validation et restitution :** contrôler calculs, titres, unités, arrondis, effectifs, destinataires et protection des données. Contrôle : restitution validée et diffusion limitée au besoin.

DR 2 complété : protocole opératoire

Protocole de traitement et de restitution

Étape	Opération à effectuer	Contrôle ou résultat attendu
1	Importer et sécuriser le fichier	5 variables présentes
2	Contrôler types et domaines	Bilan des 10 lignes consigné
3	Calculer indicateurs et graphiques	Résultats cohérents et reproductibles
4	Valider et préparer la diffusion	Calculs et confidentialité vérifiés

Quatre étapes attendues, de la réception du fichier à la validation finale.

✓ RÉPONSE À RETENIR

Importer le fichier, contrôler les données, réaliser l'analyse, puis valider les résultats et sécuriser leur restitution.

Risques et prévention

Question 4.2. Identifiez deux risques professionnels ou organisationnels liés au traitement ou à la diffusion de ces données, puis proposez une mesure de prévention adaptée à chacun. Une mesure doit concerner la confidentialité ou la minimisation des données, et l'autre la qualité ou l'interprétation des résultats.

Deux risques principaux appellent des mesures adaptées :

- **Atteinte à la confidentialité :** les identifiants pseudonymisés restent des données à protéger. Sécuriser le fichier, limiter les accès et ne diffuser que des résultats agrégés utiles.
- **Erreur ou interprétation abusive :** une association observée ne démontre pas une causalité, et une restitution incohérente peut induire le responsable en erreur. Vérifier calculs, unités et graphiques, faire relire la restitution et signaler les limites liées aux dix observations.

✓ RÉPONSE À RETENIR

Limiter et sécuriser les données diffusées ; contrôler les résultats et préciser les limites d’interprétation.

Ce qu'il faut retenir de ce sujet (Session 2026)

Bilan de l'épreuve : Épreuve du BTS Science des données, compétence C6, session 2026, code 26-BUT-SD-EP1, d'une durée de 2 h 30 et de coefficient indicatif 3,0. À partir de dix observations fictives sur la révision et les scores d'étudiants, elle évalue le contrôle des données, les statistiques descriptives, l'analyse graphique, l'interprétation prudente, la restitution reproductible et la protection des données.

Notions-clés, formules & règles à maîtriser

- Pour dix valeurs ordonnées, la médiane est la moyenne des deux valeurs centrales.
- Dans ce sujet, Q1 et Q3 sont les médianes des deux moitiés de même effectif de la série ordonnée.
- L'écart interquartile vaut IQR = Q3 - Q1 ; les clôtures de Tukey sont Q1 - 1,5 IQR et Q3 + 1,5 IQR.
- Un nuage de points décrit l'association entre deux variables quantitatives, sans démontrer de lien de causalité.

Points de vigilance & attentes du jury

- Contrôler le nombre de lignes, les valeurs manquantes, l'unicité des identifiants et les domaines des variables.
- Appliquer strictement la convention de quartiles donnée et montrer les étapes essentielles des calculs.
- Une valeur signalée par Tukey est à examiner, mais n'est pas automatiquement erronée ; vérifier sa source avant correction ou exclusion.
- Limiter les conclusions aux dix observations et protéger les identifiants pseudonymisés lors de la diffusion.

Questions fréquentes sur cette épreuve (Q&A Examen)

Q : Comment calculer les quartiles demandés dans ce sujet ?

Triez les dix scores, puis séparez-les en deux groupes de cinq valeurs. Q1 est la médiane des cinq scores inférieurs et Q3 celle des cinq scores supérieurs, conformément à la convention précisée dans le sujet.

Q : Que signifie une observation signalée par les clôtures de Tukey ?

Elle se situe à l'extérieur des bornes Q1 - 1,5 IQR et Q3 + 1,5 IQR et doit être examinée. Ce signalement ne prouve pas une erreur : vérifiez la donnée auprès de sa source avant toute correction ou exclusion.

Q : Comment organiser le temps pour cette épreuve de 2 h 30 ?

Répartissez le temps entre les quatre parties selon leurs barèmes de 5, 6, 5 et 4 points. Gardez aussi du temps pour vérifier les calculs, compléter les documents réponses et relire la restitution.

Q : Que faut-il respecter dans les conclusions et la notation ?

Appuyez les constats sur les données et les calculs, mentionnez les limites ou vérifications demandées et n'inférez pas de causalité. Vérifiez également les unités, les effectifs, les arrondis et la cohérence des graphiques avec les calculs.

Sujet officiel associé (PDF) : https://butsd.fr/assets/annales/but-sd/C6 - Mettre en œuvre une analyse exploratoire/2026/BUT Science des Données - C6 - Mettre en œuvre une analyse exploratoire - Session 2026 - 1-Sujet.pdf

Ressources BUT Science des Données : Toutes les annales & corrigés • Fiches de révision complètes • Simulateur de notes d'examen

Copyright © 2027 FormaV®. Tous droits réservés.

FormaV® est une marque déposée auprès de l'INPI sous le numéro 5227897.

Ce document a été élaboré par FormaV® avec le plus grand soin afin d'accompagner chaque apprenant vers la réussite de ses examens. Son contenu (textes, graphiques, méthodologies, tableaux, exercices, concepts, mises en forme) constitue une œuvre protégée par le droit d'auteur.

Toute copie, partage, reproduction, diffusion ou mise à disposition, même partielle, gratuite ou payante, est strictement interdite sans accord préalable et écrit de FormaV®, conformément aux articles L.111-1 et suivants du Code de la propriété intellectuelle. Dans une logique anti-plagiat, FormaV® se réserve le droit de vérifier toute utilisation illicite, y compris sur les plateformes en ligne ou sites tiers.

En utilisant ce document, vous vous engagez à respecter ces règles et à préserver l'intégrité du travail fourni. La consultation de ce document est strictement personnelle.

Merci de respecter le travail accompli afin de permettre la création continue de ressources pédagogiques fiables et accessibles.