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

BUT Science des données : compétence C2

Session 2026 • Code sujet : **26-BUT-SD-EP1** • Durée : **2 heures** • Coefficient : **3,0**
indicatif

Sommaire

Cliquez sur une partie pour y accéder directement

1. **Décrire et contrôler les données de délais de prise en charge** (4 questions) — **Page 3**
2. **Estimer et tester une proportion** (4 questions) — **Page 5**
3. **Étudier une relation et produire une prévision par régression linéaire** (4 questions) — **Page 7**
4. **Restituer les résultats et encadrer leur usage** (3 questions) — **Page 10**
5. **Ce qu'il faut retenir & Questions fréquentes (FAQ)** (Synthèse & Q&A Examen) — **Page 12**

Sujet officiel (PDF) : https://butsd.fr/assets/annales/but-sd/C2 - Analyser statistiquement les données/2026/BUT Science des Données - C2 - Analyser statistiquement les données - Session 2026 - 1-Sujet.pdf

1. Décrire et contrôler les données de délais de prise en charge

Alpina Mobilités fournit les délais de prise en charge de dix trajets clôturés pendant une même semaine, ainsi que les métadonnées et contrôles recommandés. Le candidat décrit l'échantillon avec des indicateurs statistiques et apprécie la qualité et la représentativité des données.

Partie 1 : Décrire et contrôler les données

Question 1.1. Préciser la population étudiée, l'unité statistique, la variable observée et la nature de cette variable.

La population étudiée est l'ensemble des trajets clôturés par le service pendant la semaine considérée. L'unité statistique est un trajet. La variable observée est la durée entre la validation de la réservation et la première prise en charge proposée. C'est une variable quantitative continue, mesurée en minutes.

> ✓ **RÉPONSE À RETENIR**
>
> Population : trajets clôturés pendant la semaine étudiée. Unité : un trajet. Variable : délai de prise en charge, quantitative continue, en minutes.

Question 1.2. Calculer la moyenne, la médiane et l'étendue des délais. Présenter les étapes essentielles du calcul et exprimer chaque résultat avec son unité.

La somme des dix délais est 200 minutes, donc la moyenne est 200 ÷ 10 = 20 min. Les deux valeurs centrales de la série ordonnée sont 20 et 20 : la médiane vaut (20 + 20) ÷ 2 = 20 min. Enfin, l'étendue est le maximum moins le minimum, soit 29 - 12 = 17 min.

> ✓ **RÉPONSE À RETENIR**
>
> Moyenne : 20 min ; médiane : 20 min ; étendue : 17 min.

Question 1.3. Calculer le premier quartile, le troisième quartile et l'écart interquartile en appliquant la convention donnée, puis compléter le DR 1.

Selon la convention demandée, on prend la médiane de chaque moitié de la série ordonnée. Les valeurs centrales des cinq premières observations donnent Q1 = 15 min ; celles des cinq dernières donnent Q3 = 24 min. L'écart interquartile est Q3 - Q1.

Indicateur	Résultat à inscrire au DR 1
Premier quartile Q1	15 min
Troisième quartile Q3	24 min
Écart interquartile	9 min

DR 1 complété : quartiles

Question	Indicateur	Résultat
1.3	Premier quartile Q1 (min)	15 min
1.3	Troisième quartile Q3 (min)	24 min
1.3	Écart interquartile (min)	9 min
2.2	Borne inférieure de l'IC à 95 %	
2.2	Borne supérieure de l'IC à 95 %	

✓ RÉPONSE À RETENIR

Q1 = 15 min ; Q3 = 24 min ; écart interquartile = 9 min.

Question 1.4. À partir des indicateurs obtenus, décrire le niveau central et la dispersion des délais. Indiquer une vérification de qualité pertinente avant de généraliser cette description à l'ensemble des trajets du service.

Dans cet échantillon, le délai moyen et médian est de 20 min. Les quartiles délimitent l'intervalle central de 15 à 24 min, et les délais s'étendent de 12 à 29 min.

Avant de généraliser, vérifier l'unicité des identifiants, la validité des durées et de leur unité, l'exclusion des trajets annulés et la représentativité des trajets observés par rapport au périmètre visé.

✓ RÉPONSE À RETENIR

Échantillon : centre à 20 min, quartiles de 15 à 24 min et étendue de 17 min. Toute généralisation dépend de la qualité et de la représentativité de l'échantillon.

2. Estimer et tester une proportion

La campagne numérique est étudiée à partir d'un échantillon aléatoire de 200 visites éligibles, classées selon qu'elles ont donné lieu ou non à une réservation confirmée dans les 24 heures. Le candidat calcule une fréquence et un intervalle de confiance, réalise un test unilatéral et examine les limites de l'inférence.

Partie 2 : Estimer et tester une proportion

Question 2.1. Calculer la fréquence observée des conversions. Vérifier que les conditions fournies dans le DT 2 permettent l'utilisation de l'approximation normale indiquée pour l'intervalle de confiance.

La fréquence observée vaut 126/200 = 0,630, soit 63 %. Les effectifs vérifiant les conditions du DT 2 sont 200 × 0,630 = 126 et 200 × (1 − 0,630) = 74. Tous deux dépassent 5 : l'approximation normale est recevable.

> ✓ **RÉPONSE À RETENIR**
>
> Fréquence observée : 0,630, soit 63 %. Les deux effectifs dépassent 5, donc l'approximation normale est recevable.

Question 2.2. Calculer l'intervalle de confiance à 95 % de la proportion de conversions à l'aide de la formule du DT 2. Compléter le DR 1 et interpréter l'intervalle dans le contexte de la campagne.

Avec $\hat{p} = 126/200 = 0,630$, la marge d'erreur vaut $1,96\sqrt{0,630(1-0,630)/200} \approx 0,067$. Ainsi, l'intervalle à 95 % est $[0,630 - 0,067\ ;\ 0,630 + 0,067]$, soit les bornes 0,563 et 0,697 à inscrire au DR 1. Dans le périmètre étudié, ces données donnent une estimation de la proportion de visites converties comprise entre 56,3 % et 69,7 %.

DR 1 complété, bornes de l'IC à 95 %

Question	Indicateur	Résultat
1.3	Premier quartile Q1 (min)	.....
1.3	Troisième quartile Q3 (min)	.....
1.3	Écart interquartile (min)	.....
2.2	Borne inférieure de l'IC à 95 %	0,563
2.2	Borne supérieure de l'IC à 95 %	0,697

> ✓ **RÉPONSE À RETENIR**
>
> IC à 95 % : [0,563 ; 0,697]. La proportion de conversions est estimée entre 56,3 % et 69,7 % dans le périmètre étudié.

Question 2.3. **Au seuil de 5 %, réaliser le test unilatéral de l'affirmation selon laquelle le taux de conversion de la population est supérieur à 55 %. Formuler les hypothèses, calculer la statistique de test, appliquer la règle de décision et rédiger une conclusion contextualisée. Compléter les éléments demandés dans le DR 2.**

On teste $H_0 : p = 0,55$ contre $H_1 : p > 0,55$. Sous H_0, $Z = (0,630 - 0,55)/\sqrt{0,55(1-0,55)/200} \approx 2,274$. Comme 2,274 est supérieur à la valeur critique 1,645, on rejette H_0 au seuil de 5 %. Les données soutiennent donc l'affirmation d'un taux supérieur à 55 % dans la population visée par le plan d'échantillonnage, sans l'établir avec certitude.

DR 2 complété, test de proportion

Question	Élément à déterminer	Résultat
2.3	Hypothèse nulle H0	p = 0,55
2.3	Hypothèse alternative H1	p > 0,55
2.3	Statistique de test Z	Z = 2,274
2.3	Décision au seuil de 5 %	Rejeter H_0
3.1	Pente b	

✓ RÉPONSE À RETENIR

Rejeter H_0 au seuil de 5 % : les données soutiennent l'affirmation que le taux de conversion de la population dépasse 55 %.

Question 2.4. **Indiquer une limite de l'inférence si l'échantillon, malgré sa taille, ne couvre pas certaines catégories de visites ou certaines périodes de la campagne. Distinguer cette limite de la seule variabilité d'échantillonnage.**

Si des catégories de visites ou des périodes sont absentes, l'échantillon peut mal représenter la population cible : l'estimation peut alors être biaisée, même avec un grand effectif. C'est un défaut de couverture, distinct de la variabilité d'échantillonnage, qui correspond aux fluctuations aléatoires entre échantillons représentatifs. L'intervalle de confiance tient compte de cette variabilité, mais ne corrige pas un biais de représentation.

✓ RÉPONSE À RETENIR

Un défaut de couverture peut biaiser l'estimation; augmenter la taille de l'échantillon ne corrige pas ce biais.

3. Étudier une relation et produire une prévision par régression linéaire

Cinq journées d'activité comparables sont décrites par le nombre moyen de demandes entrantes par tranche de quinze minutes et le délai moyen de première réponse. Le candidat ajuste une droite de régression, interprète la part de variance expliquée et encadre une prévision et l'interprétation de la relation.

Partie 3 : Étudier une relation et produire une prévision

Question 3.1. Calculer les moyennes de x et de y, puis déterminer les coefficients a et b de la droite de régression de y en x, écrite sous la forme ŷ = a + bx. Utiliser les sommes centrées fournies dans le DT 3 et compléter le DR 2.

Les cinq relevés donnent $\bar{x} = (1+2+3+4+5)/5 = 3$ demandes moyennes par tranche de 15 minutes et $\bar{y} = (11+15+16+19+19)/5 = 16$ min. D'après le DT 3, la pente vaut $b = S_{xy}/S_{xx} = 20/10 = 2$. L'ordonnée à l'origine est $a = \bar{y} - b\bar{x} = 16 - 2 \times 3 = 10$. La pente indique une hausse ajustée de 2 minutes par demande moyenne supplémentaire.

Coefficient	Valeur
Pente b	2 min par demande moyenne
Ordonnée à l'origine a	10 min

La droite de régression est donc $\hat{y} = 10 + 2x$.

DR 2 complété : coefficients de régression

Question	Élément à déterminer	Résultat
2.3	Hypothèse nulle H0	.....
2.3	Hypothèse alternative H1	.....
2.3	Statistique de test Z	.....
2.3	Décision au seuil de 5 %	.....
3.1	Pente b	..... 2 .....
3.1	Ordonnée à l'origine a	..... 10 .....
3.2	Coefficient de détermination R²	.....

✓ RÉPONSE À RETENIR

$\bar{x} = 3$, $\bar{y} = 16$ min, $b = 2$, $a = 10$; $\hat{y} = 10 + 2x$.

Question 3.2. Calculer le coefficient de détermination R². Interpréter sa valeur comme part de la variance observée de y expliquée par l'ajustement linéaire dans cet ensemble de relevés, puis compléter le DR 2.

Pour les cinq journées, le coefficient de détermination est $R^2 = S_{xy}^2/(S_{xx}S_{yy}) = 20^2/(10 \times 44) = 0{,}909$. L'ajustement linéaire rend ainsi compte d'environ 90,909 % de la variance observée des délais dans ces relevés. Cette proportion décrit uniquement la qualité d'ajustement sur les cinq journées étudiées. Elle ne démontre ni une relation causale ni la validité du modèle pour d'autres journées ou hors du domaine observé.

DR 2 complété : coefficient de détermination

Question	Élément à déterminer	Résultat
2.3	Hypothèse nulle H0	.....
2.3	Hypothèse alternative H1	.....
2.3	Statistique de test Z	.....
2.3	Décision au seuil de 5 %	.....
3.1	Pente b	.....
3.1	Ordonnée à l'origine a	.....
3.2	Coefficient de détermination R²	..... 0,909 .....

✓ RÉPONSE À RETENIR

$R^2 = 0{,}909$: environ 90,909 % de la variance observée des délais est expliquée par l'ajustement linéaire.

Question 3.3. À l'aide du modèle obtenu, estimer le délai moyen lorsque le nombre moyen de demandes vaut 6 par tranche de quinze minutes. Donner le résultat en minutes et préciser si cette estimation est une interpolation ou une extrapolation.

Le modèle obtenu est $\hat{y} = 10 + 2x$. Pour $x = 6$ demandes moyennes par tranche de quinze minutes, il donne $\hat{y} = 10 + 2 \times 6 = 22$ min. Les valeurs de x observées vont de 1 à 5 selon le DT 3. Comme 6 est hors de cet intervalle, la valeur annoncée est une extrapolation et non une interpolation. Le modèle ne garantit donc pas que la relation linéaire se prolonge à ce niveau de demandes.

✓ RÉPONSE À RETENIR

Pour $x = 6$, le délai estimé est de 22 min : c'est une extrapolation.

Question 3.4. Formuler une conclusion prudente sur le lien observé entre le volume de demandes et le délai. Citer une information ou une analyse supplémentaire nécessaire avant d'interpréter cette relation comme un effet causal.

Sur les cinq journées étudiées, le lien observé est positif : les délais moyens tendent à augmenter lorsque le volume moyen de demandes augmente. Cet ensemble restreint de relevés met en évidence une association, mais ne permet pas d'établir que la hausse des demandes cause celle des délais.

- Il faudrait examiner des facteurs susceptibles d'influencer les deux variables, comme les effectifs disponibles, les horaires ou la composition des demandes.
- Une analyse portant sur davantage de journées et contrôlant ces facteurs, ou l'évaluation d'une intervention sur la capacité de réponse, aiderait à étudier un effet causal.

> **✓ RÉPONSE À RETENIR**
>
> L'association observée est positive, mais les cinq journées ne suffisent pas à établir un lien causal.

4. Restituer les résultats et encadrer leur usage

La responsable d'Alpina Mobilités demande une synthèse destinée à une réunion opérationnelle, en distinguant les constats sur les données de l'échantillon des conclusions portant sur une population. Le candidat doit aussi expliquer la prudence requise pour une extrapolation et proposer des vérifications pour renforcer la qualité des analyses.

Partie 4 : Restituer les résultats et encadrer leur usage

Question 4.1. Rédiger une synthèse de deux phrases destinée à la responsable. Inclure un résultat descriptif sur les délais et un résultat inférentiel sur le taux de conversion, en utilisant les unités et le niveau de confiance appropriés.

Le délai moyen descriptif est calculé sur les dix trajets observés : (12 + 15 + 15 + 18 + 20 + 20 + 22 + 24 + 25 + 29) / 10 = 20 min. La fréquence de conversion est 126/200 = 0,63. Avec la formule normale du DT 2, l'intervalle à 95 % est [0,563087 ; 0,696913], soit environ [56,309 % ; 69,691 %] pour la proportion dans la population cible, sous les conditions d'échantillonnage de l'étude.

> ✓ RÉPONSE À RETENIR
>
> Sur l'échantillon, le délai moyen est de 20 min. Le taux de conversion estimé dans la population est de 63 %, avec un IC à 95 % de [56,309 % ; 69,691 %].

Question 4.2. Expliquer pourquoi la valeur prédite pour x égal à 6 doit être présentée avec davantage de prudence que les valeurs ajustées pour x compris entre 1 et 5. Donner une conséquence opérationnelle possible de cette précaution.

Les observations couvrent x de 1 à 5, tandis que x = 6 est hors de cet intervalle : la prévision est une extrapolation. La droite ajuste les cinq relevés, mais sa forme linéaire n'est pas vérifiée au-delà du domaine observé ; l'estimation est donc moins étayée. Le modèle donne 10 + 2 × 6 = 22 min. En pratique, il faut éviter de dimensionner les moyens sur cette seule valeur et la confronter à des relevés effectués lors de journées à forte demande.

> ✓ RÉPONSE À RETENIR
>
> À x = 6, la prévision extrapolée est de 22 min ; elle doit être considérée avec prudence et confirmée par de nouvelles observations.

Question 4.3. Proposer deux vérifications à effectuer avant la prochaine analyse pour renforcer la qualité et la pertinence des résultats. Elles doivent porter sur des aspects distincts, par exemple la qualité des données, le plan d'échantillonnage ou les hypothèses du modèle.

Avant la prochaine analyse, réaliser deux vérifications portant sur des aspects distincts :

- **Qualité des données :** contrôler l'unicité des identifiants, les valeurs manquantes ou non positives, les unités et l'exclusion des trajets annulés.
- **Plan d'échantillonnage :** vérifier que les catégories de visites et les périodes de la campagne sont couvertes, afin de limiter un biais de couverture.

Ces contrôles réduisent respectivement les erreurs dans les enregistrements et le risque de généraliser à partir d’un échantillon peu représentatif.

✓ RÉPONSE À RETENIR

Vérifier la cohérence des données et la couverture des catégories de visites et des périodes de la campagne.

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

Bilan de l'épreuve : Cette épreuve du BUT Science des données, session 2026, code 26-BUT-SD-EP1, évalue en deux heures la compétence C2, avec un coefficient indicatif de 3,0. À partir du contexte d'Alpina Mobilités, elle porte sur les délais de prise en charge, la conversion d'une campagne numérique et la relation entre demandes et délais. Les tâches mobilisent statistique descriptive, estimation, test d'hypothèse, régression et interprétation prudente.

Notions-clés, formules & règles à maîtriser

- Pour une série ordonnée paire, la médiane est la moyenne des deux valeurs centrales ; l'étendue est le maximum moins le minimum.
- Avec la convention du sujet, Q1 et Q3 sont les médianes des deux moitiés de la série ordonnée ; l'écart interquartile vaut Q3 - Q1.
- L'intervalle de confiance fourni est $\hat{p} \pm 1{,}96\sqrt{(\hat{p}(1-\hat{p})/n)}$, sous les conditions $n\hat{p} \geq 5$ et $n(1-\hat{p}) \geq 5$.
- En régression de y sur x, $b = Sxy/Sxx$, $a = \bar{y} - b\,\bar{x}$ et $R^2 = Sxy^2/(Sxx\,Syy)$; R^2 ne prouve pas la causalité.

Points de vigilance & attentes du jury

- Préciser population, unité statistique, variable et nature avant de calculer les indicateurs descriptifs.
- Respecter la convention de quartiles donnée et exprimer les indicateurs avec leurs unités.
- Pour un test, annoncer H0 et H1, calculer Z, comparer à la valeur critique et conclure dans le contexte.
- Vérifier le domaine observé avant toute prévision ; x = 6 est hors de l'intervalle observé [1 ; 5].

Questions fréquentes sur cette épreuve (Q&A Examen)

Q : Comment vérifier que l'approximation normale convient pour l'intervalle de confiance ?

Calculer la fréquence observée $\hat{p}$, puis vérifier les deux conditions du sujet : $n\hat{p}$ doit être au moins égal à 5 et $n(1-\hat{p})$ doit aussi être au moins égal à 5. Si elles sont satisfaites, utiliser la formule fournie.

Q : Comment décider si le taux de conversion est significativement supérieur à 55 % ?

Poser H0 : p = 0,55 et H1 : p > 0,55. Calculer Z avec l'erreur-type évaluée sous H0, puis comparer Z à 1,645. Rejeter H0 seulement si Z est strictement supérieur à cette valeur.

Q : Comment organiser le travail dans les deux heures d'épreuve ?

Le sujet comporte quatre parties totalisant 20 points : 5, 5,5, 5,5 et 4 points. Répartir le temps selon ces pondérations, garder du temps pour les interprétations et conclusions, et rendre les documents réponses demandés.

Q : Qu'attend-on dans l'interprétation des résultats et des prévisions ?

Le jury attend des unités, une conclusion replacée dans le contexte et une distinction entre description d'échantillon et inférence. Une valeur prédite hors du domaine observé doit être qualifiée d'extrapolation et présentée avec prudence.

Sujet officiel associé (PDF) : https://butsd.fr/assets/annales/but-sd/C2 - Analyser statistiquement les données/2026/BUT Science des Données - C2 - Analyser statistiquement les données - 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.