Feuille Td 1 : Statistique à une variable

Exercice 1

Pour étudier le nombre d'enfants de moins de 18 ans par famille, on choisit un échantillon de familles et pour chacune d'elles, on note le nombre d'enfants. La répartition des familles de l'échantillon suivant le nombre d'enfants est donnée par le tableau :

nombre k d'enfants	0	1	2	3	4	5	6	7	8
nombre de familles ayant k enfants	91	146	104	63	47	33	10	4	2

1. Construire le diagramme en bâtons de la série statistique.
2. Déterminer et représenter la fonction de répartition.
3. Calculer le nombre moyen d'enfants par famille dans l'échantillon.
4. Donner la médiane et les quartiles de cette série.
5. Dessinez le diagramme en boîte à moustache.

Exercice 2

Les salaires mensuels payés aux ouvriers d'une entreprise se répartissent comme suit :
102 ouvriers gagnent entre 400 et 1000 euros (valeur exlue),
104 ouvriers gagnent entre 1000 et 1250 euros (valeur exlue),
163 ouvriers gagnent entre 1250 et 1500 euros (valeur exlue),
121 ouvriers gagnent entre 1500 et 2000 euros (valeur exlue),
57 ouvriers gagnent entre 2000 et 2500 euros (valeur exlue),
48 ouvriers gagnent entre 2500 et 3500 euros(valeur exlue).

1. Dessinez l'histogramme et le polygone des fréquences cumulées.
2. Calculez le mode, la médiane.
3. Calculez le salaire mensuel moyen.

Exercice 3

On considère les statistiques suivantes sur les taux de réussites au baccalauréat de deux lycées :

	Lycée A	Lycée B	Total
Échecs	63	16	79
Réussites	2037	784	2821
Total	2100	800	2900
Taux d'échec	0,030	0,020	0,027

Quel lycée choisiriez-vous ? Une deuxième étude, plus fine, sépare les individus en deux groupes, ceux qui sont issus d'un milieu défavorisé et les autres :

	Favorisé			Défavorisé		
	Lycée A	Lycée B	Total	Lycée A	Lycée B	Total
Échecs	6	8	14	57	8	65
Réussites	594	592	1186	1443	192	1635
Total	600	600	1200	1500	200	1700
Taux d'échec	0,010	0,013	0,016	0,038	0,040	0,038

Quel lycée choisiriez-vous ? Expliquer le paradoxe (on observera que pour chaque lycée, le taux d'échec du premier tableau est une moyenne pondérée des deux taux du deuxième tableau par une formule que l'on détaillera).

Exercice 4

Les salaires annuels des 30 employés d'une entreprise sont les suivants (en centaines d'euros), présentés par ordre croissant :

100	100	100	110	120	140	150	150	150	160
160	160	180	180	190	190	200	200	200	210
220	230	230	250	260	290	340	410	420	530

1. Donner la médianes et les quartiles de cette série.
2. Calculer la moyenne.
3. Tracer l'histogramme en regroupant les données en classes de longueur 50.

4*. Les temps sont durs ; il faut faire des économies. Peut-on baisser la masse salariale de 25 000 euros en respectant les contraintes suivantes :
– les salaires inférieurs à la médiane ne sont pas modifiés,
– si X gagne plus que Y alors, après la modification, c'est toujours le cas et leur différence de salaire est divisée par au plus 2,
– un salaire ne baisse pas de plus de 10% ?
Si oui, proposer une répartition respectant les contraintes. Sinon dire pourquoi.

Exercice 5

En 2007, le taux brut de mortalité en Inde est inférieur à celui de la France : 8 pour 1000 contre 9 pour 1000. Pourtant à tout âge le taux de mortalité est inférieur en France à ce qu'il est en Inde. Expliquer.

Exercice 6

Soit $(x_1, ...x_n)$ une suite de données numériques. Notons $\bar{x}$ et s les moyennes et écarts type associés.

1. Soit a un réel, que valent les moyennes et écarts type des suites $(x_i - a)$ et (x_i/a) ?
2. Que valent la moyenne et l'écarts type de la suite $(x_i - \bar{x})/s$?

Exercice 7

Soit x un ensemble de données séparé en deux sous-ensembles y et z de taille n_y et n_z, montrer que

$$\bar{x} = p_y\bar{y} + p_z\bar{z}, \quad p_y = \frac{n_y}{n_y + n_z}, \quad p_z = \frac{n_z}{n_y + n_z}$$
$$s_x^2 = \{p_y s_y^2 + p_z s_z^2\} + \{p_y(\bar{y} - \bar{x})^2 + p_z(\bar{z} - \bar{x})^2\}.$$

Pour la seconde identité, on commencera par montrer que

$$\sum (y_i - \bar{x})^2 = \sum (y_i - \bar{y})^2 + n_y(\bar{y} - \bar{x})^2.$$

$\bar{x}$ est donc une moyenne pondérée des moyennes. s_x^2 est la somme de deux termes, le premier étant la moyenne pondérée des variances, appelée variance intra-classe ; montrer que le second, appelé variance inter-classe, peut s'interpréter comme la variance d'une certaine variable aléatoire.

Exercice 8

Les salariés de l'entreprise "VIVE LES GLACES" recoivent les salaires suivants :

Salaires mensuel	Effectif de salariés
[500-1500[	50
[1500-2500[	125
[2500-5500[	25

La population se répartit de manière homogène à l'intérieur de chaque classe.
1. Tracez l'histogramme correspondant à cette distribution.
2. Tracez la courbe de Lorenz.
3. Calculez l'indice de Gini, interprétez.

Exercice 9

Ce qu'on appelle le développement d'un pays est fréquemment mesuré uniquement par le PIB ou le PNB par habitant. D'autres mesures existent qui permettent aussi d'évaluer le niveau d'un pays de façon différente : espérance de vie, taux de mortalité infantile, taux de survie à cinq ans... L'indice de développement humain (IDH) est un indice créé pour tenir compte de différents facteurs importants. On dit que l'IDH est un indice agrégé.

	Mesure	Valeur min.	Valeur max.
Longévité	Espérance de vie à la naissance	20 ans	83,5 ans
Éducation	Durée moyenne de scolarisation	0 an	13,1 ans
Éducation	Durée attendue de scolarisation	0 an	19 ans
Niveau de vie	PIB par hab.	100	108831

Dans chaque domaine on calcule l'indice en posant

$$Indice = \frac{valeur - valeur\ min.}{valeur\ max. - valeur\ min.}.$$

L'IDH est calculé de la façon suivante. On utilise la formule précédente pour les quatre critères pris en compte. Attention, pour l'indice de niveau de vie, on applique la formule non au PNB par habitant mais au logarithme népérien du PIB par habitant. On fait ensuite la moyenne géométrique des deux indices concernant l'éducation. On applique la formule à cette moyenne géométrique (avec l'indice maximale 0,994 et 0 pour l'indice minimal). On obtient ainsi trois indices. L'IDH est obtenu en prenant leur moyenne géométrique.

1. Quel est l'indice de longévité du pays ayant la plus grande espérance de vie à la naissance ? Quel est l'indice d'éducation le plus grand ?
2. Pourquoi prendre le logarithme des revenus ?
3. Pourquoi considérer des moyennes géométriques plutôt qu'arithmétiques ?
4. Au Portugal, l'espérance de vie à la naissance est 77,9 ans, la durée attendue de scolarisation 16 ans, la durée moyenne de scolarisation 7,7 ans, le PIB (PPA) par habitant 21558 dollars. Calculer l'IDH du Portugal.

Exercice 10 *

Soit $(x_1, ...x_n)$ une suite de données numériques. Montrer que la médiane est la valeur pour laquelle la somme des distances des données à cette valeur est minimale. On remarquera que la fonction $y \to \sum_i |x_i - y|$ est continue, affine par morceaux, avec une dérivée entière sur chaque morceau. On pourra traiter séparément les cas "n pair" et "n impair".

Exercice 11*

Tracer la courbe de Lorenz et calculer l'indice de Gini correspondant aux répartitions données dans le tableau suivant.

Déciles	Revenus mensuels perçus	Cumuls des revenus perçus (en 10^7 euros)
1er	20	5
2e	500	135
3e	1000	510
4e	1400	1110
5e	1700	1885
6e	2000	2810
7e	2300	3885
8e	2800	5160
9e	3800	6865
10e	200000	12530

Feuille Td 2 : Régression

Exercice 1

On considère la série statistique à deux variables suivante :

x_i	1	2	3	4	5
y_i	0	0	1	1	2

1. Calculer les moyennes de x et de y.
2. Calculer les variances de x et de y, la covariance de x et y.
3. Donner une équation de la droite de régression de y sur x.
4. Tracer le nuage de points et la droite de régression.
5. Montrer sur ces données que le coefficient de détermination est le carré du coefficient de corrélation linéaire.

Exercice 2

On considère la série statistique à deux variables suivante :

x_i	1	5	6	1	4
y_i	-1	4	6	0	3

1. Calculer les moyennes de x et de y.
2. Calculer les variances de x et de y, la covariance de x et y.
3. Donner une équation de la droite de régression de y sur x.
4. Tracer le nuage de points et la droite de régression.

Exercice 3

On se propose d'étudier l'influence de la température sur la durée d'incubation des oeufs de grenouilles. On choisit 6 échantillons de 200 oeufs chacun. Le nombre x d'éclosion au 22-ème jour est le suivant

température t_i d'incubation en degrés Celsius	6	6,4	6,8	7,2	7,6	8
nombre x_id'éclosions à la température t_i	131	144	157	170	190	189

a) Dessiner le nuage des données et tracer "à l'oeil" une droite D qui a l'air de bien approcher ce nuage.
b) Calculer le coefficient de corrélation observé et écrire l'équation de la droite de régression de x en t. Etudier la qualité de l'ajustement.
c) Calculer le nombre d'éclosions prédit pour un échantillon de 200 oeufs au 22-ème jour pour une température de 7,5 degrés.

Exercice 4

Soient $\{(x_i, y_i)\}_{i=1}^n$ une série statistique à deux variables. À quelle condition les deux droites de régression, de x sur y et de y sur x, coïncident-elles ?

Exercice 5

On a mesuré les variables x et y sur 10 individus et obtenu les résultats suivants :

individu $n^0 i$	1	2	3	4	5	6	7	8	9	10
x_i	13	16	23	29	35	43	49	55	58	63
y_i	16,5	17,9	20,3	22	23,5	25,3	26,5	27,6	28,2	29,1

1) Calculer la droite de régression linéaire de y en x.
2) *On pose $z_i = \log x_i$ (logarithme en base 10). Chercher les valeurs de α, β, γ qui minimisent la somme $\sum_{i=1}^{10}(y_i - \alpha - \beta x_i - \gamma z_i)^2$.
3) *Représenter sur le même graphique : le nuage de points (x_i, y_i), la droite de régression de la question 1), la courbe de la fonction $y = \alpha + \beta x + \gamma \log(x)$ avec les coefficients trouvés dans la question 2).

Exercice 6

Nous souhaitons exprimer le nombre y de glace mangées en 6 mois en fonction du nombre x d'exercices de maths faits en 6 mois. Pour cela, nous avons interrogé 20 étudiants et les résultats ci-dessous sont disponibles : $\frac{1}{20}\sum_{i=1}^{20} x_i = 34.9$ et $\frac{1}{20}\sum_{i=1}^{25} x_i^2 = 1246.3$, $\frac{1}{20}\sum_{i=1}^{20} y_i = 18.34$ et $\frac{1}{20}\sum_{i=1}^{25} y_i^2 = 339.2$ et $\frac{1}{20}\sum_{i=1}^{20} y_i x_i = 646.32$.

a) Calculer l'équation de la droite de régression linéaire de y en x.

b) Calculer le coefficient de corrélation observé et écrire l'équation de la droite de régression de x en t. Etudier la qualité de l'ajustement.

c) Calculer le coefficient de détermination.

Feuille Td 3 :Rappels

Variables aléatoires et modélisations

Exercice 1

On suppose qu'il y a deux météos possibles pour un jour donné : soit il pleut, soit il fait beau. On suppose que la météo de chaque jour est indépendante de celle des autres jours et suit la même loi, il y a une probabilité égale à 1/4 de pleuvoir. Soit X_i la météo du $i$eme jour. Pour simplifier on notera que $X_i = 0$ s'il pleut au i eme jour et $X_i = 1$ s'il fait beau le $i$eme jour.

1) Quelle loi suit X_i ?

2) Quelle est la probabilité qu'il pleuve durant les 5 premiers jours ?

3) Quelle est la probabilité que durant la première semaine il pleuve durant 2 jours et qu'il fasse beau durant 5 jours ?

4) Quelle est la probabilité que durant la première semaine il pleuve durant le premier jour et le dernier jour et qu'il fasse beau durant les autres jours ?

Exercice 2

Soit Y une variable aléatoire uniformément répartie sur l'intervalle $[1, 3]$, c'est-à-dire une variable aléatoire de loi de densité

$$f_Y(y) = \frac{1}{2} \text{ si } y \in [1, 3], \quad 0 \text{ sinon.}$$

Calculer l'espérance et la variance de Y.

Exercice 3

On considère une variable aléatoire continue X dont la densité est $f(x) = 3x^2/8$ si $0 \leq x \leq 2$ et $f(x) = 0$ si $x \notin [0, 2]$.

1) Pourquoi est-ce que f est bien une densité ?

2) Calculer l'espérance et la variance de X.

Exercice 4

On lance douze fois une pièce équilibrée. Quelle est la probabilité d'obtenir six "pile" et six "face" ? Quelle est la probabilité d'obtenir au moins deux foix un "pile" ?

Exercice 5

Accident nucléaire : une certitude statistique. La probabilité d'un accident nucléaire majeur en Europe dans les trente prochaines années serait de plus de 100% (*Libération* le vendredi 3 juin 2011). L'article commence par estimer la probabilité d'un accident majeur par réacteur nucléaire et par année de fonctionnement. Selon l'article, le parc mondial actuel de réacteurs cumule 14000 réacteurs-ans (environ 450 réacteurs pendant 31 ans). Pendant cette période, il y a eu quatre accidents majeurs, ce qui mène à une probabilité d'accident majeur d'environ 0,0003 par an pour chaque réacteur. Les auteurs en « déduisent »donc que la probabilité d'un accident majeur en France (avec ses 58 réacteurs) pendant les trente prochaines années serait de 58 fois 30 fois 0,0003, donc d'environ 50%. Quant à la probabilité d'un accident en Europe (143 réacteurs) dans les trente prochaines années, elle « est »de 143 fois 30 fois 0,003, « donc » d'environ 129%. Commenter.

Exercice 6

On effectue un sondage auprès de 1000 personnes prises au hasard parmi 60 millions. La question posée admet deux réponses : A ou B. Pour simplifier, on suppose que toutes les personnes susceptibles d'être interrogées ont un avis stable sur la question posée et que tous les sondés répondent effectivement. Le nombre de personnes N préférant A est donc supposé bien défini. Le sondage a pour but d'estimer ce nombre. Nous ne nous intéressons ici qu'à la modélisation aléatoire du sondage.

1. Soit k un nombre compris entre 0 et 1000. Exprimer en fonction de k et N la probabilité que les sondeurs obtiennent k réponses A dans leur échantillon. Commencer par décrire l'ensemble des éventualités de l'expérience aléatoire que constitue le sondage.
2. À quelles conditions est-il raisonnable de considérer que les probabilités de la question précédente sont bien approchées par des probabilités associées à une loi binomiale ? Quelle loi binomiale ?

Lois normales

Exercice 7

1. Soit X une v.a. à densité de loi $N(0;1)$. Donner des valeurs approchées de $\mathbb{P}(X > -1)$; $\mathbb{P}(X < -2)$; $\mathbb{P}(1 < X < 2)$ et $\mathbb{P}(|X| < 2)$.
2. Soit Z une v.a. à densité de loi $N(1.75\ ;\ 0.01)$. Donner une valeur approchée de $\mathbb{P}(Z > 1.9)$.

Exercice 8

On suppose que la taille mesurée en mètre des garçons de 20 ans suit une loi normale de moyenne m et d'écart-type σ. On sait que 84% des garçons de 20 ans mesurent moins de 1 m 86 et que 97% mesurent plus de 1 m 58. Déterminer m et σ.

Exercice 9

Pour un échantillon de 300 individus sains, on a étudié la glycémie ; on a constaté que 20% des glycémies sont inférieures à 0.82 g/l et que 30% des glycémies sont supérieures à 0.98 g/l. En supposant que la glycémie suit une loi normale, déterminer la moyenne et l'écart-type de cette loi.

Exercice 10

500 personnes ont postulé pour une place, mais 379 ont été refusées parce qu'elles n'étaient pas assez grandes. La taille d'un individu suivant une loi normale de moyenne 171.5 cm et d'écart-type 5 cm, estimer la taille minimale exigée.

Feuille Td 4 : Estimations

Exercice 1

A la réception de colis, un responsable doute de l'exactitude des masses affichées sur les boîtes. Il prélève, au hasard, 25 boîtes qu'il pèse. Soit x i la masse de ième boîte. Il obtient
$\sum_{i=1}^{25} x_i = 49,5\ kg$ et $\sum_{i=1}^{25} x_i^2 = 98,3\ kg^2$.
On supposera que les masses de la production suivent une loi normale.
Donner une estimation ponctuelle de la moyenne et de la variance de la masse des boîtes de la production.

Exercice 2

Soient $X_1, \dots X_n$ des variables aléatoires i.i.d. La variable aléatoire X_1 est continue et a pour densité,

$$f(y) = 3x^2 \text{ si } y \in [0,1], \quad 0 \text{ sinon.}$$

1) Quelle est l'espérance de X_1 ? Quelle est la variance de X_1 ?
2) Que donne la loi des grands nombres pour les $X_1, \dots X_n$?
3) Que donne le TCL pour les $X_1, \dots X_n$?

Exercice 3

Soient $X_1, \dots X_n$ des variables aléatoires i.i.d. La variable aléatoire X_1 est discrète et prend uniformément ses valeurs sur $0, 1, 2, 3, 4$.
1) Quelle est l'espérance de X_1 ? Quelle est la variance de X_1 ?
2) Que donne la loi des grands nombres pour les $X_1, \dots X_n$?
3) Que donne le TCL pour les $X_1, \dots X_n$?

Exercice 4

On lance 300 fois un dé à 6 faces non truqué. Minorer (par un nombre strictement positif) la probabilité que le nombre d'apparition du 1 soit compris strictement entre 30 et 70. On pourra penser à utiliser l'inégalité de Bienaymé-Tchebychev.

Exercice 5

On lance 4720 fois une pièce équilibrée. Donner un intervalle centré en 0,5 contenant la proportion de "pile" obtenue avec probabilité supérieure à 0,99 (utiliser le théorème limite central).

Exercice 6

Monsieur Raimbourg décide de jouer à la roulette. Il parie systématiquement sur le rouge. Il fait mille parties, misant à chaque fois dix euros. Donner une valeur approchée de la probabilité qu'il perde moins de cent euros (en utilisant le théorème limite central).

Exercice 7

Un homme politique influent et riche voudrait évaluer la proportion p de la population française pensant rouge (plutôt que bleu) sur une question donnée. Il a entendu parler de ce que peut faire un sondage mais ne sait pas comment s'y prendre. Sa commande est la suivante : il veut un intervalle de longueur 2% contenant p avec probabilité supérieure à 0,999. Évaluer la taille de l'échantillon nécessaire pour répondre à sa commande.
1) En utilisant l'inégalité de Bienaymé-Tchebychev.
2) En utilisant le théorème limite central.

Exercice 8

On veut construire un parc d'attraction sur les mathématiques. Afin de savoir si le projet est rentable il faudrait qu'au moins 1 pourcent de la population soit intéréssé par ce projet. On effectue un sondage aléatoire dans la population française dans le but d'estimer la proportion p de personnes qui sont intéressées par ce projet. On interroge 1000 personnes prises au hasard dans la population. On appelle X_i la variable aléatoire définie par $X_i = 1$ si la $i$ème personne interrogée est intéréssée, $X_i = 0$ sinon.

1. Quelle est la loi suivie par chaque X_i ? Quelle est la loi suivie par le nombre de personnes intéressées par un tel projet dans un tel échantillon de 1000 personnes ? On considère que le sondage est fait avec remise.
2. Énoncer le théorème limite central (pour une suite de variables de même loi que les X_i).
3. Sur 1000 personnes interrogées, 45 disent être intéressées par un parc d'attraction sur les mathématiques. En utilisant l'approximation fournie par le théorème limite central (et la table jointe à l'énoncé) donner un intervalle de confiance au niveau 0,95 pour la proportion p. Est-ce que d'après vous il est rentable de construire le parc ?
4. Sur 1000 personnes interrogées, 45 disent être intéressées par un parc d'attraction sur les mathématiques. En utilisant l'inégalité de Bienaymé-Tchebychev donner un intervalle de confiance au niveau 0,95 pour la proportion p. (On pourra penser par que comme $0 \leq p \leq 1$ on peut majorer $p * (1 - p)$ par $1/4$). Est-ce que d'après vous il est rentable de construire le parc ?

Exercice 9

On effectue un sondage aléatoire dans la population française dans le but de déterminer la proportion p d'individus éprouvant de la peur à l'idée d'effectuer un voyage en avion.
On interroge 1000 personnes. Sur ces 1000 personnes, 253 affirment éprouver la peur de l'avion, les 747 restantes n'éprouvant pas d'appréhension particulière. Sur la base de ces données, proposer une fourchette de valeurs plausibles pour la valeur de p (donner un intervalle de confiance à 90%).
1) En utilisant l'inégalité de Bienaymé-Tchebychev.
2) En utilisant le théorème limite central.

Exercice 11

Soient $X_1, \dots X_n$ des variables aléatoires i.i.d. La variable aléatoire X_1 est continue et a pour densité,

$$f(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}} \text{ si } y > 0, \quad 0 \text{ sinon.}$$

1) Quelle est l'espérance de X_1 ?
2) Donner un estimateur $\hat{\theta}$ pour θ.
3) Est ce que $\hat{\theta}$ est un estimateur sans biais de θ ?
4) Quelle est l'erreur quadratique moyenne (MSE) de $\hat{\theta}$?

Exercice 12

Soient $X_1, \dots X_n$ des variables aléatoires i.i.d. On suppose que la variance de X_1 est finie.
1) Donner un estimateur $\hat{\nu}$ pour l'espérance ν de X_1.
2) Donner un estimateur $\hat{\sigma^2}$ pour la variance σ^2 de X_1.
3) En utilisant le TCL et le lemme de Slutsky, donner un intervalle de confiance asymptotique à 95 pourcent de ν. Donner une application numérique en sachant que $n = 60$, $\sum_{i=1}^{60} x_i = 49$ et $\sum_{i=1}^{60} x_i^2 = 138$.

Exercice 13

À la suite d'un premier infarctus du myocarde, 10% des patients font une rechute. La prise d'un anticoagulant pourrait permettre de réduire ce risque. On met en place un essai clinique pour tester l'effet de l'anticoagulant dans la prévention des rechutes. On fait un essai avec 500 patients. On observe un taux de rechute de 8,5%. Peut-on considérer au niveau 0,95 que l'anticoagulant a un effet ? Et si on avait observé le même taux avec une étude portant sur 10000 patients ? Utiliser le théorème limite central.

Exercice 14

Une compagnie d'assurance se propose d'assurer 100000 clients contre le vol. Les sommes en euros (la plupart du temps nulles) $X_1, ..., X_{100000}$ qu'aura à rembourser chaque année la compagnie aux clients sont des v.a. indépendantes d'espérance 75 et d'écart type 750. Quelle somme cette compagnie d'assurance doit-elle faire payer à chaque client par an pour que ses frais évalués à 1,5 millions d'euros soient couverts avec une probabilité supérieure ou égale à 0.999 ? (On utilisera sans justification l'approximation par la loi normale.)

Exercice 15*

Une compagnie aérienne a N places dans ses avions qu'elle vend à un prix P. Ses clients ne se présentent pas à l'embarquement avec probabilité p. Elle décide de vendre plus de places que ce dont elle dispose avec la politique suivante : si un client ne se présente pas il perd le prix de son billet, si un client se présente mais n'a pas de place la compagnie l'indemnise à hauteur de λP ($\lambda > 1$). La compagnie souhaite qu'avec une probabilité supérieure à 95% son chiffre d'affaire soit supérieur à NP. Combien doit-elle vendre de places au maximum (on supposera que N est grand et on utilisera une approximation par la loi normale) ? Est-ce ce critère qu'utiliserait une compagnie pour décider comment elle fixe ses prix ? Application numérique : $N = 5000$, $p = 0,1$, $\lambda = 4$.

Exercice 16*

Monsieur Soal s'intéressait à certains phénomènes paranormaux, par exemple à la télépathie. Il a mené dans ce cadre de multiples expériences dans lesquelles il demandait à différentes personnes de deviner des cartes. L'une de ces expériences a consisté à faire deviner la valeur d'une carte parmi cinq à madame Gloria Stewart. L'expérience a été menée 37100 fois. Madame Stewart a trouvé la bonne valeur 9410 fois. Monsieur Soal avait une formation de mathématicien ; il a donc utilisé les techniques mathématiques pour analyser ce résultat.
Considérons l'expérience comme un moyen pour évaluer la probabilité, notée p, qu'a madame Stewart de deviner la valeur de la carte.
1) Donner des intervalles de confiance pour p aux niveaux 0,95, 0,99, 0,999, 0,99999.
2) Plaçons-nous sous l'hypothèse que madame Stewart réponde au hasard. Donner une majoration de la probabilité que la proportion du nombre de fois où elle devine la valeur de la carte soit supérieure à 0,23.
Vous ne trouverez pas de valeur approchée dans la table. Une possibilité : majorée la probabilité en utilisant (si $a > 0$ et X suit une loi normale centrée réduite) :

$$\mathbb{P}(X > a) \leq \frac{e^{-\frac{a^2}{2}}}{a}$$

3) Les principes de la statistique mathématique ne nous invitent-ils pas ici à considérer que madame Stewart a effectivement un pouvoir qui consisterait à deviner significativement plus souvent la valeur de la carte que le hasard ? Après tout c'est ainsi qu'on procède pour affirmer qu'un médicament est efficace (significativement plus que le placebo), non ?
La réponse à cette question serait affirmative si les données fournies par monsieur Soal étaient fiables. Or ce n'est pas le cas. À plusieurs reprises dans sa carrière, il a procédé à des opérations sur ses données qui les rendent inutilisables. Ces opérations peuvent sembler inoffensives parfois : choisir l'outil d'analyse après avoir recueilli les données plutôt qu'avant (l'un des principes de la statistique est de ne pas le faire, même si l'outil est valable), oublier une parties des données,...