

TD1 — Statistique descriptive

Exercice 1.1. La liste des 30 communes françaises les plus peuplées en 2011 est donnée ci-dessous (figure 1). Pour plus de lisibilité, on pourra convertir (et arrondir) les effectifs des communes en milliers d'habitants.

- 1) Tracer un histogramme des effectifs des 30 premières communes françaises.
- 2) Quel est l'effectif moyen parmi ces 30 communes ? Que vaut la médiane ? Commenter la différence.
- 3) Que vaut l'écart-type des effectifs des 30 communes ? Est-il élevé ?
- 4) Déterminer l'écart interquartile des effectifs des 30 communes, puis tracer une boîte à moustaches.

Paris	2 249 975	Grenoble	157 424
Marseille	850 636	Dijon	151 672
Lyon	491 268	Angers	148 803
Toulouse	447 340	Saint-Denis	145 347
Nice	344 064	Villeurbanne	145 034
Nantes	287 845	Nîmes	144 940
Strasbourg	272 222	Le Mans	143 240
Montpellier	264 538	Clermont-Ferrand	140 957
Bordeaux	239 399	Aix-en-Provence	140 684
Lille	227 533	Brest	140 547
Rennes	208 033	Limoges	137 758
Reims	180 752	Tours	134 633
Le Havre	174 156	Amiens	133 327
Saint-Étienne	170 049	Metz	119 962
Toulon	163 974	Perpignan	118 238

FIGURE 1 – Effectifs des 30 premières communes françaises.

Exercice 1.2. Manipulation du symbole Σ . On considère la suite des 7 valeurs $x_1, x_2, \dots, x_7$ suivantes :

x_1	x_2	x_3	x_4	x_5	x_6	x_7
2	1	5	3	1,5	4	0,5

- 1) Calculer $\sum_{i=1}^7 x_i$ et $\sum_{i=2}^6 x_i$.
- 2) Calculer $\sum_{i=1}^7 x_i^2$.
- 3) Calculer la moyenne et l'écart-type du jeu de données précédent.

Exercice 1.3. Les tailles en pouces des 105 étudiants d'un cours de biostatistique aux États-Unis sont réparties selon l'histogramme de la figure 2 (NB : 1 pouce = 2,54 cm).

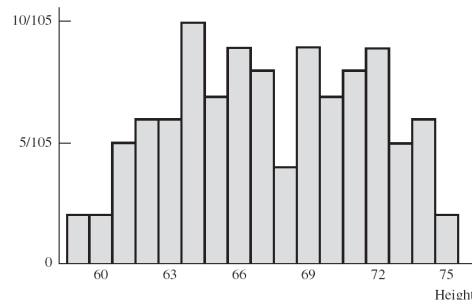


FIGURE 2 – Les tailles des 105 étudiants.

- 1) Décrire la forme de l'histogramme.
- 2) Cet histogramme présente-t-il une caractéristique particulière ?
- 3) Voyez-vous une raison particulière expliquant les deux pics de l'histogramme ? Une autre variable que la taille serait-elle utile pour expliquer ces deux pics ?

Exercice 1.4. *Formule de la variance.* On va expliquer pourquoi la variance d'un caractère statistique X peut s'exprimer des deux façons équivalentes suivantes :

$$v(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{X}^2.$$

- 1) Montrez que $(x_i - \bar{X})^2 = x_i^2 - 2\bar{X}x_i + \bar{X}^2$ pour tout $i \in \{1, \dots, N\}$.
- 2) En déduire que

$$\frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2 = \left(\frac{1}{N} \sum_{i=1}^N x_i^2 \right) - \bar{X}^2.$$

Exercice 1.5. Cet exercice est extrait de l'ouvrage *Mathematical statistics with Applications* de Wackerly, Mendenhall et Scheaffer (2008, chapitre 1). À vous de traduire...

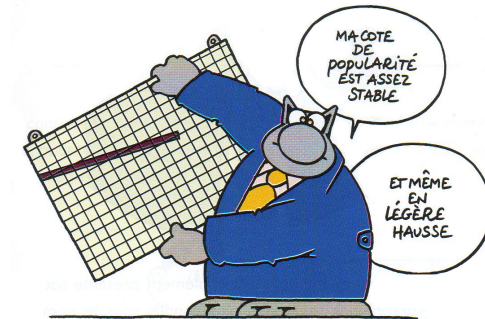
Are some cities more windy than others? Does Chicago deserve to be nicknamed "The Windy City"? Given below are the average wind speeds (in miles per hour) for 45 selected U.S. cities:

8.9	12.4	8.6	11.3	9.2	8.8	35.1	6.2	7.0
7.1	11.8	10.7	7.6	9.1	9.2	8.2	9.0	8.7
9.1	10.9	10.3	9.6	7.8	11.5	9.3	7.9	8.8
8.8	12.7	8.4	7.8	5.7	10.5	10.5	9.6	8.9
10.2	10.3	7.7	10.6	8.3	8.8	9.5	8.8	9.4

Source: *The World Almanac and Book of Facts*, 2004.

- a Construct a relative frequency histogram for these data. (Choose the class boundaries without including the value 35.1 in the range of values.)
- b The value 35.1 was recorded at Mt. Washington, New Hampshire. Does the geography of that city explain the magnitude of its average wind speed?
- c The average wind speed for Chicago is 10.3 miles per hour. What percentage of the cities have average wind speeds in excess of Chicago's?
- d Do you think that Chicago is unusually windy?

Exercice 1.6. Quel message peut-on tirer de ce dessin humoristique ?



Source : *La mathématique du chat* de Philippe Geluck, Daniel Justens.

TD2 — Modélisation probabiliste

Exercice 2.1. Carine joue au casino sur une machine à sous avec des pièces de 2 euros. À chaque tour, sa machine délivre un gain (en euros) avec les probabilités suivantes :

gain (en €)	0	1	2	10
probabilité	0,8	0,1	0,07	0,03

- 1) Si on appelle X le gain de Carine après un essai, à quoi correspond le tableau précédent ?
- 2) Soit Y le gain net de Carine après un essai. Déterminer la loi de la variable aléatoire Y .
- 3) Combien Carine peut-elle espérer gagner en moyenne après un essai ? Commenter.

Exercice 2.2. *Différentes façons de choisir.* Dans la grille suivante, on doit choisir une case blanche au hasard. On propose trois procédures pour ce choix :

- (A) On choisit une case blanche uniformément au hasard parmi les cinq cases blanches.
- (B) On choisit uniformément au hasard une ligne puis, dans la ligne, une case blanche.
- (C) On choisit uniformément au hasard une colonne puis, dans la colonne, une case blanche.

a		
b		c
	d	e

- 1) On modélise la procédure (A) par une variable aléatoire X . Donner la loi de X .
- 2) On modélise la procédure (B) par une variable aléatoire Y . Donner la loi de Y .
- 3) On modélise la procédure (C) par une variable aléatoire Z . Donner la loi de Z .

On a simulé n choix d'une case blanche selon l'une des trois procédures ci-dessus, pour $n = 100$, puis $n = 1000$ et enfin $n = 10000$. On a obtenu les fréquences ci-dessous :

Case	a	b	c	d	e
Fréquence ($n = 100$)	0,19	0,16	0,22	0,31	0,12
Fréquence ($n = 1000$)	0,150	0,159	0,174	0,368	0,149
Fréquence ($n = 10000$)	0,167	0,168	0,164	0,336	0,165

- 4) Au vu de ce tableau, peut-on deviner la procédure utilisée ?

Remarque : L'observation des résultats simulés ne permet pas de remonter au modèle à coup sûr. Mais l'un des objectifs de la statistique inférentielle est d'évaluer le risque de se tromper en « remontant d'une distribution de fréquences à une loi de probabilité ».

Exercice 2.3. On lance un dé à quatre faces (tétraèdre régulier) et un dé à six faces (cube). Les faces des dés sont numérotées de la manière suivante : le dé tétraédrique a trois faces portant le numéro 3 et une face portant le numéro 1, tandis que le dé cubique a trois faces portant le numéro 0, deux faces portant le numéro 2 et une face portant le numéro 4. On appelle X la variable aléatoire égale au numéro de la face cachée du dé tétraédrique, et Y celui de la face cachée du dé cubique.

- 1) Déterminer les lois des variables aléatoires X et Y .
- 2) Supposons que les lancers des deux dés sont indépendants. Donner la loi de $S = X + Y$.
- 3) Calculer l'espérance et l'écart-type de la variable aléatoire S .

Exercice 2.4. *Exercice sur la notion d'indépendance.* Bill, Joe et Sam tirent sur une cible au même instant. Pour chacun d'eux, les probabilités d'atteindre la cible sont respectivement $1/2$, $2/3$ et $1/4$. On suppose leurs tirs indépendants. Quelle est la probabilité pour qu'ils atteignent tous les trois la cible ? Qu'aucun des trois n'atteigne la cible ? Que la cible soit atteinte par au moins l'un d'eux ?

Exercice 2.5. *Deux anniversaires le même jour ?* Pendant un match de foot, deux équipes de 11 joueurs sont sur le terrain.

- 1) Quelle la probabilité que les 22 joueurs sur le terrain soient tous nés un jour de l'année différent ? (Pour simplifier le problème, on oubliera les années bissextiles, et on supposera que tous les jours de l'année sont équiprobables¹.)

1. Si certaines dates sont plus représentées que d'autres, on peut en fait montrer (avec un raisonnement plus sophistiqué) que le phénomène est encore plus marqué, au sens où il est encore plus probable que deux joueurs soient nés le même jour de l'année.

2) En déduire la probabilité qu'au moins deux joueurs sur le terrain soient nés le même jour de l'année. Commentaire ?

Exercice 2.6. Soit X une variable aléatoire de loi donnée par :

a	1	2	3
$\mathbb{P}(X = a)$	1/3	1/3	1/3

- 1) Soit X_1 et X_2 deux variables aléatoires indépendantes de même loi que X (on peut penser à deux tirages indépendants dans la même population). Déterminer la loi de la variable aléatoire $M_2 = (X_1 + X_2)/2$, puis la représenter graphiquement.
- 2) Même question avec trois variables aléatoires X_1, X_2 et X_3 indépendantes et de même loi que X . Quelle est la loi de la variable aléatoire $M_3 = (X_1 + X_2 + X_3)/3$? Représentation graphique ?
- 3) Calculer et comparer les espérances et les écarts-types de X_1, M_2 et M_3 . Commentaires ?

Exercice 2.7. *Deux exemples d'échantillons.* Au cours d'une étude de santé, une équipe de chercheurs d'un hôpital souhaite connaître le rythme cardiaque moyen des 20-50 ans de leur commune. L'un d'entre eux interroge un premier petit groupe de personnes et relève, en nombre de pulsations par minute, les valeurs suivantes :

75 61 72 85 61 70 58 73 83 70

Parallèlement, un collègue interroge un deuxième groupe de personnes et relève les valeurs suivantes :

57 80 86 67 75 57 61 70 68 77

- 1) Décrire l'expérience aléatoire sous-jacente.
- 2) Calculer les moyennes M_1^{obs} et M_2^{obs} des échantillons 1 et 2. (La mention "obs" permet de préciser qu'il s'agit de quantités calculées à partir des données effectivement observées.)
- 3) Calculer les écarts-types $S_1^{\text{obs}}, S_2^{\text{obs}}$ et les écarts-types corrigés $(S_1^{\text{corr}})^{\text{obs}}, (S_2^{\text{corr}})^{\text{obs}}$ des échantillons 1 et 2. Commenter la différence entre ces deux formes d'écarts-types.

Exercice 2.8. *Endocytose (épisode 1).* Lors d'un TP de biologie cellulaire en 2013, des étudiants de L2 Biochimie ont étudié le phénomène d'endocytose chez l'amibe sociale. Le ligand utilisé était le HRP, une enzyme que l'amibe ne connaît pas et n'arrive pas à digérer, donc facile à mesurer. Nous allons analyser les résultats des 18 binômes ayant travaillé avec une concentration initiale de HRP de $30 \mu\text{g mL}^{-1}$. Voici leurs résultats de mesure pour la densité optique par million de cellules, à l'instant $t = 20 \text{ min}$:

-0,0003	0,0132	0,0083	0,0154	0,0251	0,0371	0,0032	0,0059	-0,018
0,0052	0,0885	0,0196	0,0005	0,001	0,0238	0,0138	0,0306	0,0139

- 1) Où est l'aléatoire dans cette expérience de biologie ?
- 2) Calculer la moyenne d'échantillon et l'écart-type d'échantillon corrigé.

Exercice 2.9. Quel message peut-on tirer de ce dessin humoristique ?



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Source : <http://xkcd.com/795/>

TD3 — Tests statistiques pour une moyenne

Exercice 3.1. Une étude a montré que l'apport moyen journalier en vitamine D nécessaire au bon développement des bébés est de $\mu_0 = 10 \mu\text{g}$ par jour. On se demande si la quantité de vitamine D apportée par l'alimentation sous forme de lait en poudre produit par une certaine marque est conforme avec cette norme. On note μ la quantité moyenne journalière de vitamine D apportée par cette alimentation. On considère un échantillon formé de 50 bébés âgés de trois mois et nourris au biberon. On mesure, lors d'une journée bien identifiée, l'apport en vitamine D du lait en poudre pour chacun de ces bébés. On calcule alors l'apport moyen ($9,1 \mu\text{g}$) et l'écart-type d'échantillon ($1,9 \mu\text{g}$).

- 1) Où est l'aléatoire dans ce contexte ?
- 2) Utiliser un test statistique avec le risque de première espèce $\alpha = 0,05$ pour tester l'hypothèse nulle $H_0: \mu = \mu_0$ contre l'hypothèse alternative $H_1: \mu \neq \mu_0$. L'alimentation avec cette marque de lait en poudre fournit-elle la quantité voulue de vitamine D aux bébés ?
- 3) Même question avec le risque de première espèce $\alpha = 0,01$.
- 4) Quelle grandeur pourrait-on fournir en guise de résultat afin de ne pas reconduire le test pour chaque nouvelle valeur de α ?

Exercice 3.2. *Boîtes de conserve (épisode 1).* Un fabricant de boîtes de conserve affirme que le poids moyen des boîtes de sa production est de 495 g. Pour vérifier ses dires, il prélève 43 boîtes dont les poids (en grammes) se répartissent de la façon suivante :

poids (en g)	490	492	495	497	502	505
effectif	6	11	10	8	5	3

- 1) Les données prélevées viennent-elles contredire l'affirmation du fabricant ?

Une société de contrôle vient faire un prélèvement un peu plus important et relève les poids de 149 boîtes :

poids (en g)	490	492	495	497	502	505
effectif	12	31	35	38	20	13

- 2) Les données prélevées viennent-elles contredire l'affirmation du fabricant ? Qu'est-ce qui a changé par rapport à la question précédente ?

Exercice 3.3. Des étudiants de votre promotion viennent vous présenter une étude qu'ils ont conduite sur 25 rats d'une même espèce. Sachant que le poids moyen des rats *Sprague-Dawley* est d'environ 350 g et que les poids des 25 rats étudiés étaient (en grammes) :

386 389 338 356 378 375 339 376 389 376 366 336 385
349 382 342 335 388 375 331 354 363 371 378 361,

pensez-vous qu'ils s'agissait de rats *Sprague-Dawley* ? Quelle hypothèse de modélisation avez-vous faite ?

Exercice 3.4. *Pile ou face consécutifs.* Vous demandez à un ami de lancer 100 fois une pièce et de noter les résultats obtenus (pile ou face). Vous suggérez à votre ami qu'il peut soit effectuer les 100 lancers consciencieusement, soit tricher en imaginant une suite de 100 lancers.

Un calcul mathématique ou une simple simulation numérique permet de montrer la chose suivante : si les 100 lancers sont indépendants et si la pièce est bien équilibrée, alors la loi de la variable aléatoire

$$Z = \text{"nombre maximal de pile ou face consécutifs parmi les 100 lancers"}$$

est donnée approximativement par

a	1	2	3	4	5	6	7	8	9	10
$\mathbb{P}(Z = a)$	0	0	0,000 31	0,029 89	0,169 11	0,266 38	0,224 14	0,143 88	0,080 93	0,042 87
a	11	12	13	14	15	16	17	18	...	
$\mathbb{P}(Z = a)$	0,020 95	0,011 17	0,005 46	0,002 48	0,001 33	0,000 56	0,000 26	0,000 12	...	

Comment allez-vous procéder pour déterminer si votre ami a vraiment lancé les pièces ou s'il a triché ? La réponse doit bien sûr être mathématiquement fondée.

Exercice 3.5. Afin de tester une solution toxique, on fait des injections à un groupe de 80 souris. Il est d'usage de penser que l'injection est mortelle dans 80 % des cas. Le fait que 22 souris ne soient pas mortes est-il compatible au seuil de 5 % avec cette hypothèse ?

Exercice 3.6. Un journaliste souhaiterait savoir si la parité est respectée parmi les agents secrets français. Ces données n'étant pas publiques, il entreprend une petite étude statistique en interrogeant plusieurs contacts indépendants bien placés dans le renseignement. Ses résultats sont les suivants : sur les 73 agents secrets dont il a vu la fiche descriptive, 28 sont des femmes.

1) En prenant un risque de première espèce de 5 %, le journaliste peut-il tirer une conclusion sur la parité parmi les agents secrets français ? Que signifie ce risque de première espèce ?

2) Qu'en est-il avec un risque de première espèce de 2 % ?

Exercice 3.7. Quel message peut-on tirer de ce dessin humoristique ?



Source : <http://xkcd.com/1448/>

TD4 — Intervalles de confiance pour une moyenne

Exercice 4.1. *Boîtes de conserve (épisode 2).* On reprend l'exercice 3.2. Un fabricant de boîtes de conserve veut contrôler le poids des boîtes qu'il a produites. On note μ le poids moyen de toutes les boîtes de sa production, et σ l'écart-type correspondant. Un échantillon de 10 boîtes donne les poids suivants (en grammes) :

490 492 497 502 505 490 495 492 490 497

1) Proposer des estimations pour μ et σ .

Le fabricant décide de tester davantage de boîtes (pour limiter les incertitudes d'estimation). Il contrôle donc 33 boîtes supplémentaires, ce qui donne pour les 43 boîtes au total :

poids (en g)	490	492	495	497	502	505
effectifs	6	11	10	8	5	3

2) Proposer de nouvelles estimations pour μ et σ . À votre avis, ces estimations sont-elles meilleures que les précédentes ? Pourquoi ?

3) Donner un intervalle de confiance pour μ au niveau de confiance 95 %.

4) Même question avec un niveau de confiance de 99 %. Qu'est-ce qui change ? Pourquoi ?

Exercice 4.2. *Endocytose (épisode 2).* On reprend l'exercice 2.8 correspondant au TP sur le phénomène d'endocytose chez l'amibe sociale. Rappelons les résultats de mesure (densité optique par million de cellules à l'instant $t = 20$ min) des 18 binômes ayant travaillé avec une concentration initiale de HRP de $30 \mu\text{g mL}^{-1}$:

-0,0003 0,0132 0,0083 0,0154 0,0251 0,0371 0,0032 0,0059 -0,018
0,0052 0,0885 0,0196 0,0005 0,001 0,0238 0,0138 0,0306 0,0139

Donner un intervalle de confiance à 95 % pour la densité optique par million de cellules à l'instant $t = 20$ min qu'on observerait en moyenne sur toute la population des amibes sociales, dans des conditions expérimentales parfaites, et sans bruit de mesure.

Exercice 4.3. Sur 12 000 individus d'une espèce de lapins, on a dénombré 13 albinos.

1) Estimer la proportion p d'albinos dans l'espèce. On donnera un intervalle de confiance pour p au niveau de confiance 95 %.

2) Même question avec un niveau de confiance de 99 %. Quelle la différence avec la réponse à la question précédente ?

Exercice 4.4. *Êtes-vous un citoyen éclairé ?* Imaginons une situation politique en 2022 : Nicolas Hulot se présente aux élections présidentielles, et nous sommes le jour du second tour, en fin d'après-midi. Peu de bulletins ont été dépouillés, mais l'institut IFOP a réalisé un sondage à la sortie des urnes afin d'estimer la proportion des électeurs qui ont voté pour Nicolas Hulot.

1) Après avoir interrogé 1225 personnes, l'institut de sondage a compté 637 électeurs ayant voté pour Nicolas Hulot. Donner un intervalle de confiance à 95 % pour la proportion des électeurs français qui ont voté pour lui.

2) Vous êtes journaliste sur une chaîne de télévision, et vous devez parler des élections. Quel pronostic annoncerez-vous à l'antenne ?

Exercice 4.5. Quel message peut-on tirer de ce dessin humoristique ?



Source : *La mathématique du chat* de Philippe Geluck, Daniel Justens.

TD5 — D'autres exemples de tests statistiques

Exercice 5.1. Une étude médicale vise à connaître les risques de vertiges liés à la prise d'un certain médicament. Au total, 1163 patients soignés avec ce médicament ont été suivis et on a relevé les résultats suivants :

vertiges	pas de vertiges
108	1055

Ces résultats mettent-ils en doute l'affirmation du laboratoire pharmaceutique selon laquelle la proportion de gens subissant des vertiges après prise du médicament est inférieure à 8 % ?

Exercice 5.2. Dans l'article *Sexual activity and the lifespan of male fruitflies*, Nature, 1981, Partridge et Farquhar rapportent des résultats d'expérience visant à étudier le coût de reproduction en termes de longévité pour les drosophiles mâles. Parmi les expériences menées, deux groupes de 45 mâles ont été comparés : dans le premier groupe, chacun des 45 mâles a été isolé avec une femelle vierge réceptive ; dans le second groupe, chacun des 45 mâles a été isolé avec une femelle non réceptive (rôle de contrôle). La durée de vie des mâles a été mesurée dans chaque groupe (en jours) :

groupe de 45 mâles avec	durée de vie moyenne	écart-type corrigé
1 femelle réceptive	56,76	14,9284
1 femelle non réceptive	64,80	15,6525

Ces données permettent-elles d'affirmer qu'une activité sexuelle plus importante est un facteur de réduction de longévité chez la drosophile mâle ? Par ailleurs, sur quelle hypothèse votre test repose-t-il ?

Exercice 5.3. Un chercheur laisse tomber (volontairement) un stylo dans un ascenseur et note si l'autre occupant l'aide à le ramasser. Il relève les résultats suivants, en fonction du sexe de l'occupant :

	a aidé	n'a pas aidé	total
hommes	370	950	1320
femmes	300	1003	1303

Peut-on en déduire que, dans un ascenseur, les hommes et les femmes sont autant prêts à aider un autre occupant venant de laisser tomber son stylo ?

Exercice 5.4. Un médecin cherche à savoir si la pratique régulière du jogging entraîne une réduction de la fréquence cardiaque. Huit volontaires, qui ne pratiquaient pas le jogging auparavant, acceptent de suivre un programme d'entraînement de jogging pendant un mois. Leurs fréquences cardiaques moyennes ont été mesurées avant et après le programme d'entraînement (en nombre de battements par minute) :

fréquence cardiaque avant	74	86	98	102	78	84	79	70
fréquence cardiaque après	70	85	90	110	71	80	69	74

- 1) Ces données mettent-elles en évidence une réduction du rythme cardiaque après un mois d'entraînement ?
- 2) Quelle hypothèse de modélisation avez-vous faite à la question précédente ? Comment pourrait-on vérifier que cette hypothèse est valide ?
- 3) Parmi les huit volontaires, combien ont diminué leur rythme cardiaque ? Est-ce en contradiction avec votre réponse à la première question ?

Exercice 5.5. On souhaite savoir si les chimpanzés préfèrent le rouge, le bleu, le vert ou le jaune. Un chercheur réalise donc l'expérience suivante sur 83 chimpanzés : il leur montre 4 cartons (un par couleur, les 4 cartons sont de forme identique et sont présentés dans une disposition aléatoire) ; il attend ensuite que le chimpanzé choisisse l'un des cartons et note la couleur choisie. Les résultats obtenus sont les suivants :

rouge	bleu	vert	jaune
30	22	15	16

Cette expérience est-elle une preuve expérimentale du fait que les chimpanzés ont une préférence de couleur dans le choix des cartons ?

Exercice 5.6. Imaginons le scénario suivant : un sondage citoyen effectué auprès de 937 couples en France vise à savoir si chacune des deux personnes du couple serait prête à partager la voiture de son conjoint pour aller au travail le matin. On obtient les résultats suivants :

	pour	contre	total
femmes	106	831	937
hommes	87	850	937

1) Les résultats obtenus sur ces deux échantillons d'hommes et de femmes mettent-ils en évidence un comportement différent des hommes et des femmes face au covoiturage ?

On dispose en fait des données plus précises suivantes : les 937 couples interrogés se répartissent selon

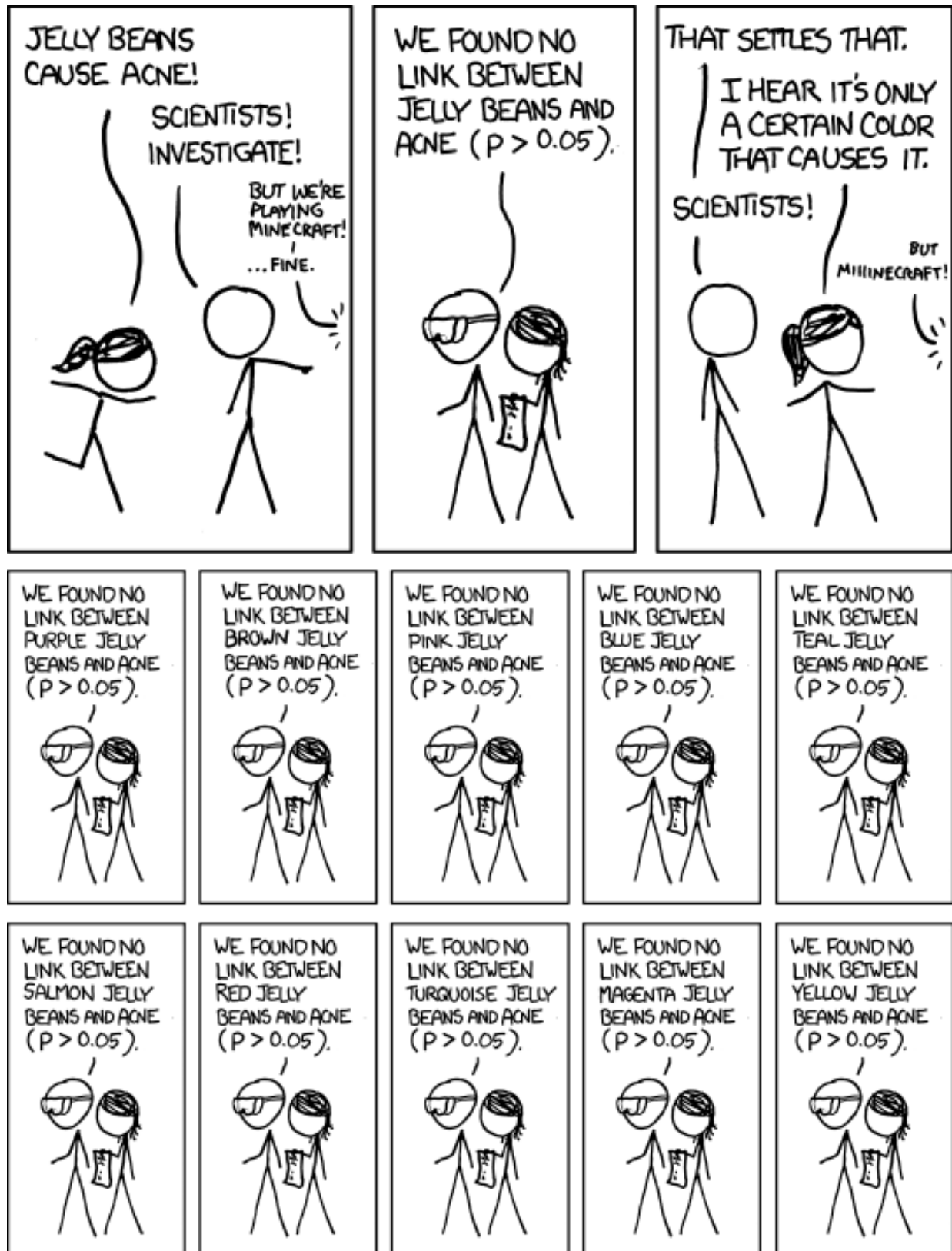
	hommes pour	hommes contre
femmes pour	73	33
femmes contre	14	817

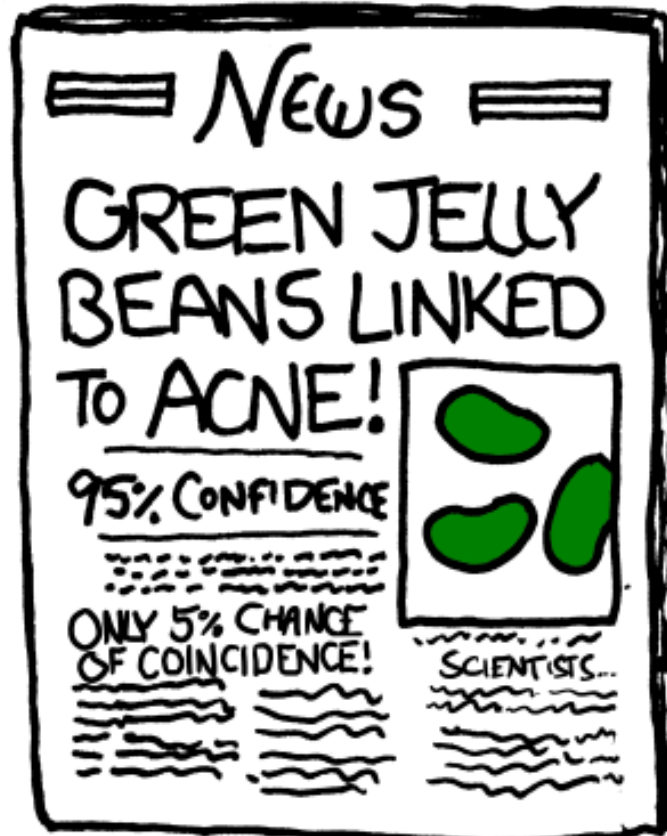
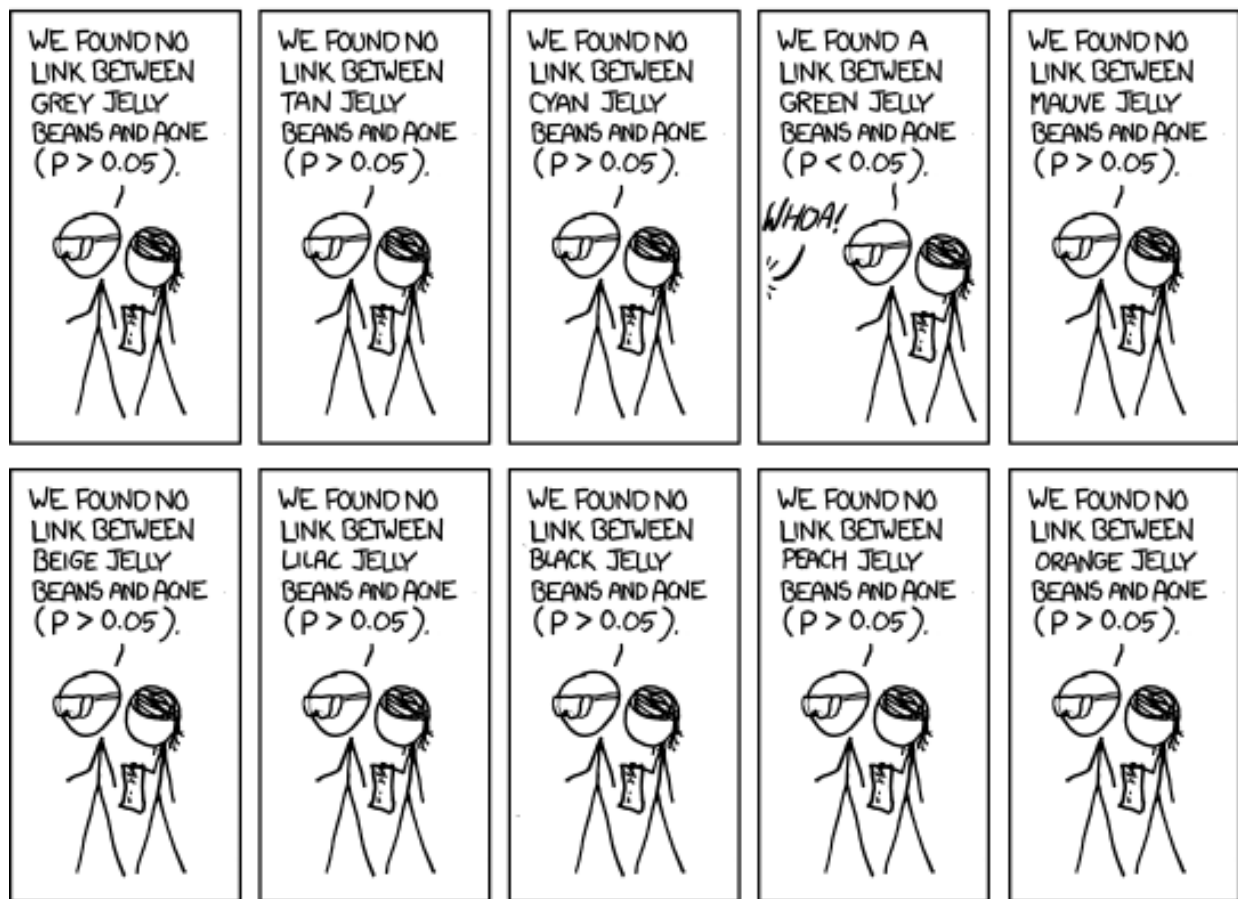
2) Que remarquez-vous dans ces données ?

3) Ces données permettent-elles de mettre en évidence un comportement différent des hommes et des femmes face au covoiturage au sein d'un couple ?

4) Que vaut la p-valeur du test ?

Exercice 5.7. Quel message peut-on tirer de cette bande dessinée humoristique ?





Source : <http://xkcd.com/882/>