UE4 : Biostatistiques

Chapitre 6 :

Problème récapitulatif

Professeur Philippe CINQUIN

Année universitaire 2010/2011

Plan

- A) Introduction
- B) Statistiques descriptives
- C) Probabilités
- D) Estimation
- E) Intervalles de confiance
- F) Problème récapitulatif
- G) Résumé des objectifs

Problème récapitulatif

NB vous n'avez pas besoin de calculette pour ce problème (vous n'aurez pas droit aux calculettes au concours). Les calculs sont ou bien suffisamment simples pour pouvoir se faire de tête (par exemple, sommes ou soustractions de nombres entiers, divisions par 1000, divisions donnant des résultats entiers, …), ou bien plus difficiles, mais dans ce cas vous trouverez dans l'énoncé des tableaux résumant les résultats de certaines opérations, parmi lesquelles celles dont vous aurez besoin

- 1) On étudie les mesures de la concentration sanguine en glucose (glycémie) sur un échantillon de n = 1000 prélèvements réalisés sur un an à l'entrée des patients, choisis par tirage au sort sur des patients différents, sur la base du numéro de patient, dans un hôpital qui reçoit environ 10 000 patients/an.
  - 1.1) Quelle est selon vous la population qu'un tel échantillonnage permet d'étudier ?
  - 1.2) Que pensez-vous de la méthode d'échantillonnage retenue ?
  - 1.3) quel type de variable est la variable « glycémie » ?
  - 1.4) quels autres types de variable connaissez-vous ?

Problème récapitulatif

- 2) Les données sont connues sous la forme du tableau 1.
  - 2.1) Comment s'appelle un tel tableau ? Que permet-il d'étudier ? Pourquoi est-il licite de regrouper les données par classes ?
  - 2.2) Calculez la fréquence de la classe [5 , 6[

Classes (glycémie en mmoles/L)	*Effectifs*
2-2.99	0
3 - 3.99	11
4 - 4.99	80
5 - 5.99	166
6 - 6.99	159
7 - 7.99	116
8 - 8.99	95
9 - 9.99	112
10 10.99	133
11 - 11.99	72
12 - 12.99	38
13 - 13.99	17
14 - 14.99	1
ou plus...	0

Problème récapitulatif

- On complète ce tableau de la manière suivante
  - 2.3) Calculez la moyenne de la glycémie à partir de ce tableau
  - 2.4) quel est le moyen graphique le mieux adapté pour représenter ces données ? Pourquoi ?
  - 2.5 Mettez en œuvre ce moyen graphique et commentez le résultat
  - 2.6 Quels paramètres pouvez-vous estimer à partir de ce graphique ? Donnez leurs valeurs
  - 2.7) Proposez un moyen graphique d'estimer la médiane, et le 3ème quartile, et mettez le en œuvre.

Classes (glycémie en mmoles/L)	*Effectifs*	*Somme des valeurs observées dans la classe*
2-2.99	0	0
3 - 3.99	11	38,5
4 - 4.99	80	360
5 - 5.99	166	913
6 - 6.99	159	1033,5
7 - 7.99	116	870
8 - 8.99	95	807,5
9 - 9.99	112	1064
10 10.99	133	1396,5
11 - 11.99	72	828
12 - 12.99	38	475
13 - 13.99	17	229,5
14 - 14.99	1	14,5
ou plus...	0	0
Total	1000	8030

Problème récapitulatif

- 3) Intrigué par le caractère bi-modal de l'échantillon, le statisticien a demandé des données cliniques complémentaires, et s'est aperçu que le CHU concerné disposait d'un service de diabétologie important, dont l'activité représente environ la moitié de celle de l'hôpital. Il reprend l'analyse sur les seuls patients issus de ce service de diabétologie.

  Les données sont fournies sous la forme suivante (où x_i représente la valeur du $i^{ème}$ individu de l'échantillon) :
  taille de l'échantillon : n = 501;
  Σx_i = 4 509mmoles/l;
  Σx_i^2 = 42 000 $(mmoles/l)^2$
- Pour les calculs, on pourra noter que $4\,509^2/501 = 40\,581$ et utiliser le tableau des racines de nombres ci-joint :

nombre	racine
1,419	1,19
2,838	1,68
4,257	2,06
5,676	2,38

Problème récapitulatif

- 3.1) quelle peut être la population de référence ?
- 3.2) estimer la moyenne et l'écart-type de la variable aléatoire « glycémie » dans la population d'où est extrait l'échantillon observé
- 3.3) quelle est la probabilité qu'un patient de la population ait une glycémie inférieure à 12 mmoles/l ?

Problème récapitulatif

- On donne Racine(500) = 22.3 et

nombre	nombre/22.3	3/nombre
1,19	0,053	2,52
1,68	0,075	1,79
2,06	0,092	1,46
2,38	0,107	1,26

- 3.4) quels sont les intervalles de confiance à 95%, à 99%, à 999‰, de l'estimation de la moyenne de la glycémie dans la population d'où est extrait l'échantillon ?
- 3.5) admettons que la variance de la glycémie dans la population soit exactement celle que l'on a estimée avec cet échantillon : combien de prélèvements faudrait-il inclure dans l'échantillon pour que l'intervalle de confiance à 95% de la moyenne ait une étendue réduite de moitié par rapport à celle que vous avez calculée dans la question précédente ?

Problème récapitulatif

- 4) On sait qu'une personne sur 100 présente une pathologie associée, indépendante du diabète, mais qui aggrave ce dernier. Quelle est la probabilité que, dans un échantillon de dix diabétiques, on trouve au moins une personne présentant cette pathologie associée ?
  On donne :
  Ln (99) = 4.595
  Ln (100) = 4.605
  exp(0.1) = 1.11
  exp(0) = 1
  exp(-0.1) = 0.90
  exp(-0.2) = 0.82

Corrigé

- 1) On étudie les mesures de la concentration sanguine en glucose (glycémie) sur un échantillon de n = 1000 prélèvements réalisés sur un an à l'entrée des patients, choisis par tirage au sort sur des patients différents, sur la base du numéro de patient, dans un hôpital qui reçoit environ 10 000 patients/an.
  - 1.1) Quelle est selon vous la population qu'un tel échantillonnage permet d'étudier ? Pourquoi ?
    - population utilisant l'hôpital étudié (NB acceptée aussi toute autre formulation équivalente ou montrant qu'on a compris que l'échantillon ne pouvait être représentatif de la glycémie de la population de la planète…)
  - 1.2) Que pensez-vous de la méthode d'échantillonnage retenue ?
    - Satisfaisante (randomisée) : le tirage au sort garantit le caractère randomisé de l'échantillonage, qui permet donc d'étudier la glycémie dans la population dont est extrait l'échantillon
  - 1.3) quel type de variable est la variable « glycémie » ?
    - Quantitative, continue
  - 1.4) quels autres types de variable connaissez-vous ?
    - Qualitative, ordinale, quantitative et discrète, …

Corrigé

- 2) Les données sont connues sous la forme du tableau 1.
  - 2.1) Comment s'appelle un tel tableau ? Que permet-il d'étudier ? Pourquoi est-il licite de regrouper les données par classes ?
    - Tableau des effectifs. Il permet d'étudier la distribution dans l'échantillon considéré de la variable aléatoire « glycémie ». Cette variable étant continue, il est impératif de regrouper les données par classes (NB un tel regroupement peut aussi être utile pour une variable discrète présentant un grand nombre de valeurs d'effectif non nul).
  - 2.2) Calculez la fréquence de la classe [5 , 6[
    - 166/1000 = 0.166

Classes (glycémie en mmoles/L)	*Effectifs*
2-2.99	0
3 - 3.99	11
4 - 4.99	80
5 - 5.99	166
6 - 6.99	159
7 - 7.99	116
8 - 8.99	95
9 - 9.99	112
10 10.99	133
11 - 11.99	72
12 - 12.99	38
13 - 13.99	17
14 - 14.99	1
ou plus...	0

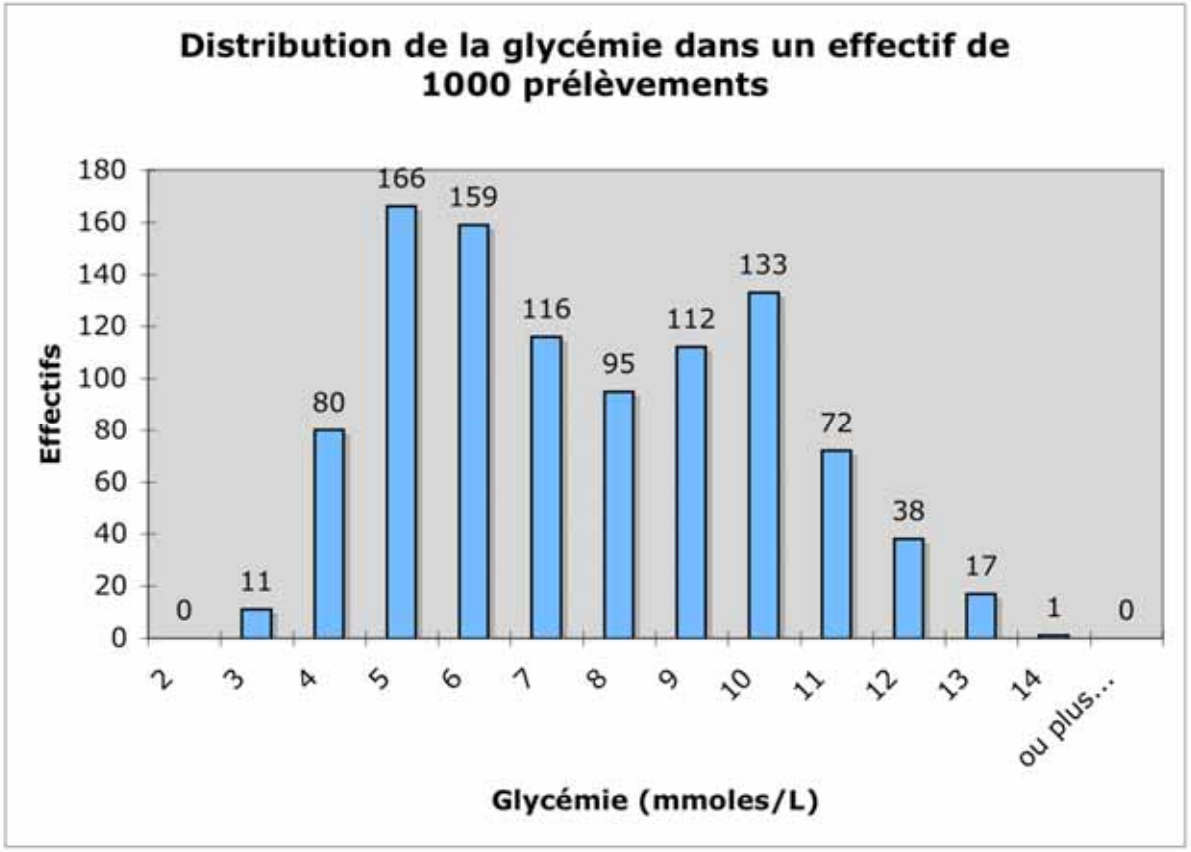
Corrigé

- On complète ce tableau de la manière suivante
  - 2.3) Calculez la moyenne de la glycémie à partir de ce tableau
    - Moyenne = 8030/1000 = 8.03 mmoles/L
  - 2.4) quel est le moyen graphique le mieux adapté pour représenter ces données ? Pourquoi ?
    - Histogramme (variable quantitative)

Classes (glycémie en mmoles/L)	*Effectifs*	*Somme des valeurs observées dans la classe*
2-2.99	0	0
3 - 3.99	11	38,5
4 - 4.99	80	360
5 - 5.99	166	913
6 - 6.99	159	1033,5
7 - 7.99	116	870
8 - 8.99	95	807,5
9 - 9.99	112	1064
10 10.99	133	1396,5
11 - 11.99	72	828
12 - 12.99	38	475
13 - 13.99	17	229,5
14 - 14.99	1	14,5
ou plus...	0	0
Total	1000	8030

Corrigé

- 2.5 Mettez en œuvre ce moyen graphique et commentez le résultat
  - Histogramme bimodal, asymétrique
- 2.6 Quels paramètres pouvez-vous estimer à partir de ce graphique ? Donnez leurs valeurs
  - Modes (classes [5,6[et [10,11[). On peut aussi donner les valeurs centrales des classes modales, soit 5.5. et 10.5
  - Minimum : 3 et Maximum : 15



Corrigé

- 2.7) Proposez un moyen graphique d'estimer la médiane, et le 3ème quartile, et mettez le en œuvre.
  - Calculer le tableau des fréquences cumulées. On peut y lire directement la classe médiane (la première à égaler ou dépasser 50%)
  - et la classe « 3ème Quartile » (la première à égaler ou dépasser 75%)

Classes (glycémie en mmoles/L)	Effectifs	Somme des valeurs observées dans la classe
2-2.99	0	0
3 - 3.99	11	38,5
4 - 4.99	80	360
5 - 5.99	166	913
6 - 6.99	159	1033,5
7 - 7.99	116	870
8 - 8.99	95	807,5
9 - 9.99	112	1064
10 10.99	133	1396,5
11 - 11.99	72	828
12 - 12.99	38	475
13 - 13.99	17	229,5
14 - 14.99	1	14,5
ou plus...	0	0
Total	1000	8030

Exemple de calcul des fréquences cumulées : valeur de la classe [5,6[

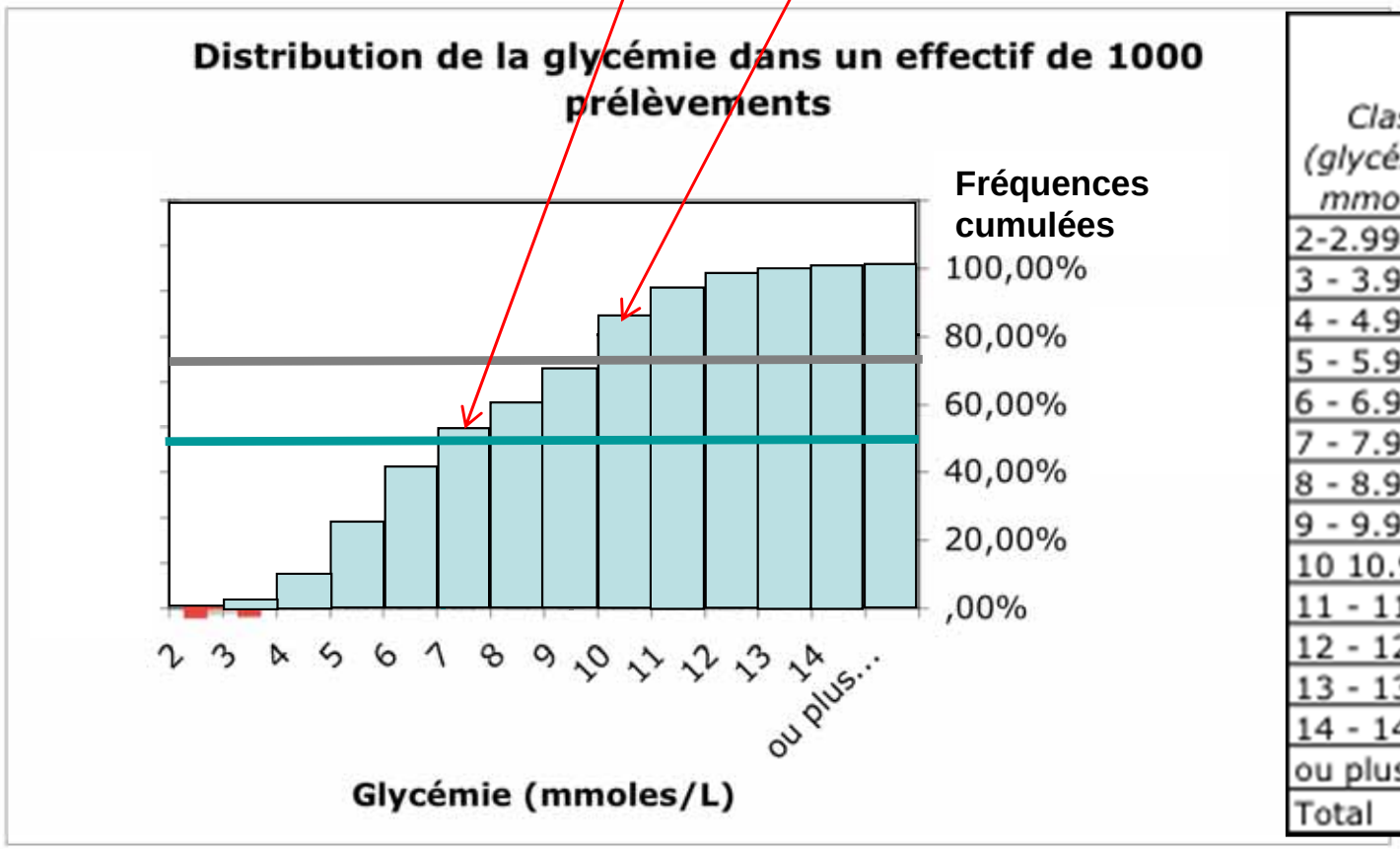
0.26 = (0 + 11 + 80 +166)/1000 =

257/1000 ≈ 26%

Classes (glycémie en mmoles/L)	% cumulé
2 - 2.99	0
3 - 3.99	1
4 - 4.99	9
5 - 5.99	26
6 - 6.99	42
7 - 7.99	53
8 - 8.99	63
9 - 9.99	74
10 - 10.99	87
11 - 11.99	94
12 - 12.99	98
13 - 13.99	100
14 - 14.99	100
ou plus...	100

Corrigé

- 2.7) Proposez un moyen graphique d'estimer la médiane, et le 3ème quartile, et mettez le en œuvre (*suite*).
  - La classe Médiane est [7,8[parce que c'est la première classe dont la représentation graphique est coupée par la ligne horizontale 50%
  - La classe 3ème Quartile est [10,11[parce que c'est la première classe dont la représentation graphique est coupée par la ligne horizontale 75%



Classes (glycémie en mmoles/L)	*Effectifs*	*Somme des valeurs observées dans la classe*
2-2.99	0	0
3 - 3.99	11	38,5
4 - 4.99	80	360
5 - 5.99	166	913
6 - 6.99	159	1033,5
7 - 7.99	116	870
8 - 8.99	95	807,5
9 - 9.99	112	1064
10 10.99	133	1396,5
11 - 11.99	72	828
12 - 12.99	38	475
13 - 13.99	17	229,5
14 - 14.99	1	14,5
ou plus...	0	0
Total	1000	8030

Corrigé

- 3) taille de l'échantillon : n = 501;
  Σx_i = 4 509mmoles/l;
  Σx_i^2 = 42 000 (mmoles/l)^2
- Pour les calculs, on pourra noter que
  4 509^2/501 = 40 581 et utiliser le tableau des racines de nombres ci-joint :

nombre	racine
1,419	1,19
2,838	1,68
4,257	2,06
5,676	2,38

- 3.1) quelle peut être la population de référence ?
  - Population des diabétiques utilisant l'hôpital étudié
- 3.2) estimer la moyenne et l'écart-type de la variable aléatoire « glycémie » dans la population d'où est extrait l'échantillon observé
  - La moyenne inconnue μ est estimée par
    m = Σx_i/n = 4 509 / 501 mmoles/L = 9 mmoles/L
  - L'écart type inconnu σ est estimée p

$$\sigma = \sqrt{\frac{\sum x_i^2 - \left(\sum x_i\right)^2 / n}{n-1}} = \sqrt{\frac{42000 - 4509^2/501}{500}}$$

$$\sigma = \sqrt{\frac{42000 - 40581}{500}} = \sqrt{\frac{1419}{500}} = \sqrt{\frac{2838}{1000}} = \sqrt{2.838} = 1.68$$

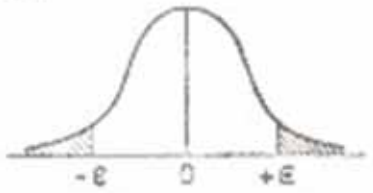
Corrigé

- 3.3) quelle est la probabilité qu'un patient de la population ait une glycémie inférieure à 12 mmoles/l ?
- ε = (12-9)/1.68 = 3/1.68 = 1.79
- α = 0.075 (milieu entre 0.07 et 0.08)
- P(G < 12) = P[(G-9)/1.68 < 1.79] = 1 - α /2 = 1 - 0.0375 = 0.96

nombre	nombre/22.3	3/nombre
1,19	0,053	2,52
1,68	0,075	1,79
2,06	0,092	1,46
2,38	0,107	1,26

Table de l'écart-réduit (loi normale) (*).

La table donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée ε, c'est-à-dire la probabilité extérieure à l'intervalle $(-\varepsilon, +\varepsilon)$.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058

Corrigé

- On donne Racine(500) = 22.3 et

nombre	nombre/22.3
1,19	0,053
1,68	0,075
2,06	0,092
2,38	0,107

- 3.4) quels sont les intervalles de confiance à 95%, à 99%, à 999‰, de l'estimation de la moyenne de la glycémie dans la population d'où est extrait l'échantillon ?

$$\sigma_e = \frac{\sigma}{\sqrt{n}} = \frac{1.68}{\sqrt{501}} \approx \frac{1.68}{\sqrt{500}} = \frac{1.68}{22.3} = 0.075$$

  - Il faut aller lire dans la table de l'écart réduit les valeurs de ε pour les trois valeurs de $(1-\alpha)$ choisies

Corrigé

- Pour $\alpha = 0.05$, on sait que $\varepsilon = 2$ (1.96 si on veut être précis, mais retenez la valeur 2)
- Pour $\alpha = 0.01$, on lit $\varepsilon = 2.576 = 2.6$ (une décimale suffit ici)
- Pour $\alpha = 0.001$, on lit $\varepsilon = 3.29 = 3.3$ (une décimale suffit ici)

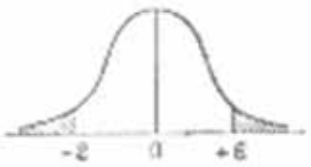
- En appliquant la formule $[m-\varepsilon\sigma_e; m+\varepsilon\sigma_e]$ nous obtenons donc les 3 intervalles :
  [8.85 ; 9.15] pour 95%
  [8.81 ; 9.19] pour 99%
  [8.75 ; 9.25] pour 999‰

  NB il faut ici effectuer les multiplications de 0.075 par 2, par 2.6 et par 3.3; qui doivent pouvoir être réalisées manuellement

TABLE I

Table de l'écart-réduit (loi normale) (*).

La table donne la probabilité α pour que l'écart-réduit égale ou dépasse, en valeur absolue, une valeur donnée ε, c'est-à-dire la probabilité extérieure à l'intervalle $(-\varepsilon, -\varepsilon)$.



ε	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311

α	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
ε	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

Corrigé

- 3.5) admettons que la variance de la glycémie dans la population soit exactement celle que l'on a estimée avec cet échantillon : combien de prélèvements faudrait-il inclure dans l'échantillon pour que l'intervalle de confiance à 95% de la moyenne ait une étendue réduite de moitié par rapport à celle que vous avez calculée dans la question précédente ?
  - Il en faut 4 fois plus, soit 4 x 501 = 2004
  - En effet, l'étendue de l'intervalle de confiance est proportionnelle à σ_e. Or, n intervient dans le calcul de σ_e par sa racine, et est au dénominateur.

Corrigé

- 4) On sait qu'une personne sur 100 présente une pathologie associée, indépendante du diabète, mais qui aggrave ce dernier. Quelle est la probabilité que, dans un échantillon de dix diabétiques, on trouve au moins une personne présentant cette pathologie associée ?
- Il s'agit d'une loi binomiale, avec n = 10 et p = 0.01 (donc q = 0.99). Soit B cette loi, on cherche $P(B \geq 1)$
- Or $P(B \geq 1) = 1 - P(B=0)$
- $P(B=0) = q^{10} = 0.99^{10}$
- Or, $\text{Ln}(0.99^{10}) = 10 \times [\text{Ln}(99) - \text{Ln}(100)] = 10 \times (4.595 - 4.605)$ $= 10 \times (-0.01) = -0.1$
- $P(B=0) = 0.99^{10} = \exp(\text{Ln}0.99^{10}) = \exp(-0.1) = 0.90$
- Donc $P(B \geq 1) = 1-0.9 = 0.1$

Mentions légales

Ce document a été réalisé par la Cellule TICE de la Faculté de Médecine et de Pharmacie de Grenoble
(Université Joseph Fourier – Grenoble 1)