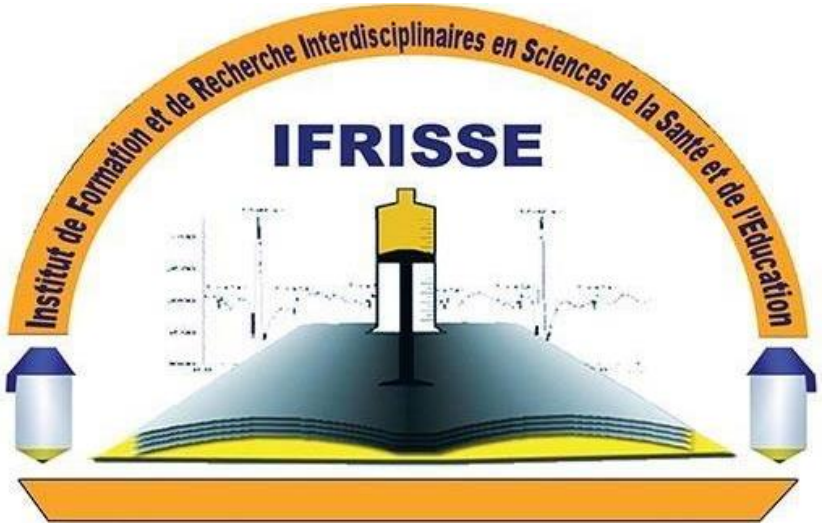


Introduction à la biostatistique

CORRECTION DES EXERCICES DE COURS

Exercice 1.1

- B : Qualitative nominale
- C : Qualitative binaire
- D : Type date (variable quantitative après transformation)
- E : Quantitative continue
- F : Qualitative nominale
- G : Qualitative nominale
- H : Qualitative ordinale

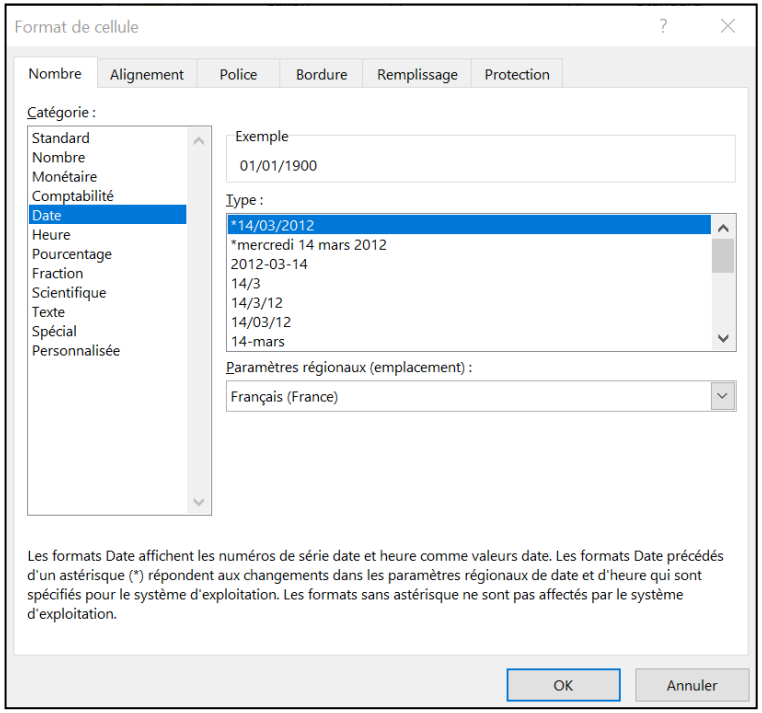
Exercice 1.2

Selon les bornes choisies pour les classes, plusieurs solutions sont possibles :

Etiquette de classe	1	2	3	4	5	6	Total
Centre de classe	44	45	46	47	48	49	
Classe	[43.5 -44.5[	[44.5 -45.5[	[45.5 -46.5[	[46.5 -47.5[	[47.5 -48.5[	[48.5 -49.5[	
Effectifs	4	7	9	8	8	14	50

Etiquette de classe	1	2	3	4	5	6	Total
Centre de classe	44.5	45.5	46.5	47.5	48.5	49.5	
Classe	[44 -45[	[45 -46[	[46 -47[	[47 -48[	[48 -49[	[49 -50[	
Effectifs	4	9	10	10	10	7	50

Exercice 1.3



1	31/07/1969
2	17/08/1969
3	=B2-B1

1	31/07/1969	03:56
2	17/08/1969	12:45
3	17	=C2-C1

1. 1er Janvier 1900
2. Le point de départ des dates Excel est fixé au 01/01/1900 = 1
3. 25415 est le nombre de jours écoulés entre le 01/01/1900 et le 31/07/1969
4. 15 jours
5. 8h49mn

Exercice 2.1

Dilution	n	Fréquence relative	3 classes de fréquences équivalents	2 classes	3 classes sémantiques
1/2	4	4/121=0.033	0.322	7.4	14 % non significatifs
1/4	5	0.041			
1/8	8	0.066		24.8	
1/16	22	0.182			74.4% de suspects
1/32	25	0.207	0.339	33.9	
1/64	16	0.132			
1/128	11	0.091	0.339	14.9	
1/256	7	0.058			
1/512	9	0.074		12.4	
1/1024	6	0.05			11.6% de paludismes évolutifs
1/2048	5	0.041		6.6	
1/4096	3	0.025			
Total	121	1		100	100%

Exercice 3.1 et 3.2

3.1

1. Camembert ou diagramme en barres
2. 2 histogrammes juxtaposés ou une pyramide des âges
3. Histogramme ou polygone de fréquence
4. Diagramme en barres horizontales (un camembert serait moins lisible)

3.2 Il faut se méfier des résultats d'une étude comportant trop de données manquantes.

- Si ceux qui n'ont pas répondu sont des fumeurs qui ont eu « honte » de répondre, alors p=65%
- Si aucun d'eux ne fume p=35%
- S'il y a autant de fumeurs que de non fumeurs parmi les non-répondants, alors p envoisine 50%

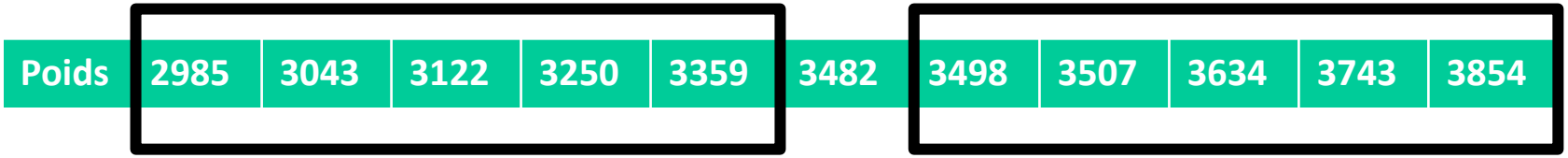
Exercice 3.3

3.3

1. Absence de légende pour les 2 lignes colorées
2. Absence de légende axe des ordonnées gauche
3. Absence de légende axe des ordonnées droite
4. Absence de légende axe des abscisses
5. Echelle non-uniforme sur l'axe des abscisses
6. Titre insuffisamment précis (âge de la population par exemple)
7. Titre insuffisamment précis (lieu et période d'étude)

Exercice 4.1

En classant les données par ordre croissant on obtient :



Médiane = 3482 g

$$Moyenne = \frac{2985 + 3043 + 3122 + 3250 + 3359 + 3482 + 3498 + 3507 + 3634 + 3743 + 3854}{11}$$

Moyenne=3407 g

Exercice 4.2

La série est classée et d'effectif pair.

$$Médiane = Moyenne\ des\ \left(\left[\frac{n}{2}\right]\text{ème} ; \left[\frac{n+2}{2}\right]\text{ème}\right) valeurs$$

Médiane =(3359 + 3482)/2 = 3420 g

Exercice 4.3

Après avoir classé les valeurs de la série par ordre croissant :

Paramètre	Valeur
médiane	9.6
1er quartile	7.6
3ème quartile	14
mode	[9, 9.9[
intervalle interquartile	14-7.6=6.4
minimum	0.6
maximum	20.5
étendue	20.5-0.6=19.9
moyenne	550.3/53=10.38
variance	23.25
écart type	racine(23.25)
coefficient de variation	46.4%

```
> serie=c(9.7,5.8,11.9,16.1,15.7,17.9,2.2,10,15.3,6.6,8.2,4.2,3.6,7,7.9,2.5,8.7,9.3,11.5,9.5,9.6,9.5,16.3,
10.6,10.2,8.9,18.8,14.4,20.5,8.3,17.6,4.5,13.1,14.6,18.6,10.6,8.9,13.7,9.4,14,5.2,7.6,4.9,9.5,6.8,10.8,11.
1,9.7,19.7,4,8,0.6,16.7)
> length(serie)
[1] 53
> median(serie)
[1] 9.6
> quantile(serie, probs = 0.25)
25%
7.6
> quantile(serie, probs = 0.75)
75%
 14
> mode(serie)
[1] "numeric"
> IQR(serie)
[1] 6.4
> min(serie)
[1] 0.6
> max(serie)
[1] 20.5
```

```
> max(serie)-min(serie)
[1] 19.9
> mean(serie)
[1] 10.38302
> var(serie)
[1] 23.24874
> sd(serie)
[1] 4.821695
> sqrt(var(serie))
```

Exercice 6.1

1. Loi normale car biologique variable quantitative continue

2. Loi de Poisson si on connait la moyenne du nombre de mésothéliomes survenant annuellement

3. Loi de Poisson

4. La variables est un nombre k d'évènements survenant dans un échantillon de taille n

Exercice 6.2

x=4; n=10; p(x)=0.15

Loi binomiale

La probabilité d'observer au moins 4 sujets est la sommes des probabilités d'en observer 4, 5, 6,7,8,9 ou 10.

$$\mathrm{P}(\mathrm{X} \geq 4) = P(X = 4) + P(X = 5) + \ \ldots + P(10) = 0.05$$

Ou plus simplement :

$$\mathrm{P}(\mathrm{X} \geq 4) = 1 - \mathrm{P}(\mathrm{X} \leq 4) = 1 - [P(X = 0) + P(X = 1) + \ P(2)] = 0.05$$

Exercice 6.3

Loi binomiale

$$x = 1; \quad n = 300; \quad p(x) = 1/11000$$

$$\mathrm{P(X = 1)} = 2.6\%$$

Compliqué … Vous pouvez passer

Exercice 6.4

Attention : la solution à cet exercice n'est pas triviale. Ne vous offusquez pas si vous ne la comprenez pas parfaitement. Continuez les autres exercices.

Nous sommes dans une situation où nous pouvons calculer le pourcentage de malades dans la population totale de cet établissement : p=25/80=0.31

Lorsqu'on s'interroge sur une chambre double, on s'interroge sur une population de 2 individus. Chaque tirage est indépendant puisque les chambres sont indépendantes

Si les malades sont répartis aléatoirement dans toutes les chambres, alors la loi binomiale est respectée et on peut calculer la probabilité s'avoir 0, 1 ou 2 cas dans une chambre double.

$$P(0) = \frac{2!}{0!\,(2-0)!} \times 0.31^{0} \times 0.69^{2-0} = 0.476$$

$$P(1) = \frac{2!}{1!\,(2-1)!} \times 0.31^{1} \times 0.69^{2-1} = 0.428$$

$$P(1) = \frac{2!}{2!\,(2-2)!} \times 0.31^{2} \times 0.69^{2-2} = 0.225$$

Compliqué … Vous pouvez passer

Exercice 6.4

S'il n'y avait pas de contagion et uniquement une répartition aléatoire, la probabilité d'avoir :

- 2 cas dans une chambre double est 0.096
- 1 cas dans une chambre double est 0.428
- 0 cas dans une chambre double est 0.476

Or, selon les données provenant des 40 chambres doubles, la probabilité observée de :

- 2 cas dans une chambre double est 0.225
- 1 cas dans une chambre double est 0.175
- 0 cas dans une chambre double est 0.6

Entre le postulat aléatoire et la réalité, la probabilité qu'une chambre double ait 2 cas, passe de 9.6% à 22.5%

On suspecte donc que la distribution des cas en chambre double n'est pas aléatoire mais influencée par le phénomène de contagion.

Exercice 6.5 et 6.6

6.5 Loi de Poisson de moyenne 1 $$P(X=k)=\frac{e^{-\mu}\mu^{k}}{k!}$$

1. P(0)=36.8%; P(1)=36.8%; P(2)=8.4%; P(3)=6.1%; P(4)=1.5%
2. 1-P(0)=63.2%
3. P(0)+P(1)=73.6%

6.6 Loi de Poisson de moyenne 1.4 $$P(X=4)=\frac{e^{1.4}\times 1.4^{4}}{4!}$$

1. P(0)=0.247; P(1)=0.345; P(2)=0.242; P(3)=0.113
2. La probabilité d'observer au plus 3 cas est P(0)+P(1)+P(2)+P(3)=0.947
3. La probabilité d'observer au moins 4 cas est la p-value

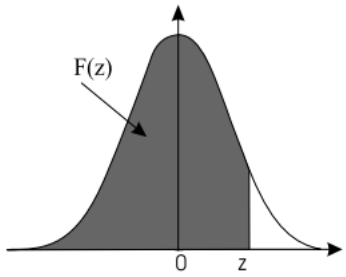
   p-value=1-0.947=0.053=5.3%
4. 5.3% > 5%. Donc il y a plus de 5% de chances d'observer au moins 4 cas au cours de la même année.

→ Au seuil de risque 5%, on ne peut pas conclure à un risque majoré

Exercice 6.7

Utilisation des tables

*Question 1. C*hances qu'un nouveau-né pèse moins de 2700 g



$$\text{Z-score} = \frac{X - \mu}{\sigma}$$

$$Zscore = \frac{X - \mu}{\sigma} = \frac{2700 - 3300}{357} = -1.68$$

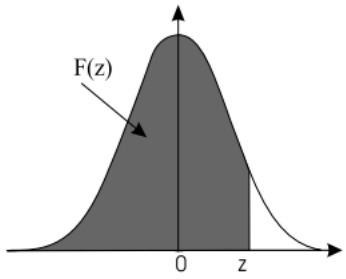
En lisant sur la table de la loi normale centrée réduite, ça correspond à une probabilité de 1-0.9535= **0.0465=4.6%**

Dans la table de la loi normale, valeur à lire à l'intersection entre la ligne **1.6** et la colonne **0.08**

Exercice 6.7

Utilisation des tables

*Question 2. C*hances qu'un nouveau-né pèse plus de 3500 g



$$\text{Z-score} = \frac{X - \mu}{\sigma}$$

$$Zscore = \frac{X - \mu}{\sigma} = \frac{3500 - 3300}{357} = 0.56$$

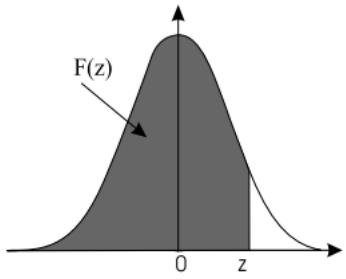
En lisant sur la table de la loi normale centrée réduite, ça correspond à une probabilité de 1-0.7123 = 0.2877=28.77%

Dans la table de la loi normale, valeur à lire à l'intersection entre la ligne **0.5** et la colonne **0.06**

Exercice 6.7

Utilisation des tables

Question 3. Poids en deçà duquel se situent 90% des nnés



Dans la table de la loi normale, le valeur de Z qui correspond à une probabilité de 90% → Donc lire à l'intersection des lignes 1.2 et 0.09

Z=1.29

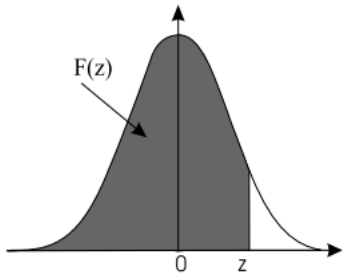
Or, Z est une variable centrée réduite $\text{Z-score} = \frac{X - \mu}{\sigma}$

On en déduit la valeur de X

$$\frac{X-\mu}{\sigma} = Z \qquad \frac{X-3300}{357} = 1.29 \qquad \rightarrow X=3760 \text{ g}$$

Exercice 6.7

Utilisation des tables



Question 4. Probabilité qu'un nné pèse entre 3000 et 3500 g

Pour 3000

$$Zscore = \frac{X - \mu}{\sigma} = \frac{3000 - 3300}{357} = -0.84$$

→ P1=1-0.799=0.201

Pour 3500

$$Zscore = \frac{X - \mu}{\sigma} = \frac{3500 - 3300}{357} = +0.56$$

→ P2=0.7123

P=P2-P1= 0.7123 - 0.201 = **0.511=51%**

Exercice 6.7

Utilisation des tables

Question 5. Les bornes de l'intervalle de confiance où se trouvent 95% des poids des nouveau-nés

*Si la variable **poids des nnés** suit une loi normale, 95% des nnés ont un poids compris entre $m \pm 1.96\sigma$*

Or $m=3300g$ et $\sigma=357g$

*IC = [3300 - 1.96*357 ; 3300 + 1.96*357] = **[2600; 4000]***

Exercice 6.7

6.7 Les réponses peut être facilement obtenues sur R, mais on peut utiliser les tables

Utilisation de R

1. *pnorm(q=2700, m=3300, sd=357, lower.tail=TRUE)* = 4.6%

2. *1-pnorm(q=3500, m=3300, sd=357, lower.tail=TRUE) = 28.8%*

Il y a 28.8% de chances qu'un nouveau-né pèse plus de 3500

3. qnorm(p=0.9, m=3300, sd=357, lower.tail = TRUE) = 3758

C'est la borne au dessous de laquelle on retrouve 90% des poids de nouveaux-nés

4. pnorm(q=3500, m=3300, sd=357, lower.tail=TRUE)- pnorm(q=3000, m=3300, sd=357, lower.tail=TRUE) = 0.512

51.5% des nouveaux-nés pèsent entre 3000 et 3500 g

5. Si la variable poids des nnés suit une loi normale, 95% des nnés ont un poids compris entre m ± 1.96σ

Exercice 9.1

L'écart type de la moyenne est $s_m = \frac{s}{\sqrt{n}} = \frac{0.4}{5} = 0.08$

1. La taille de l'échantillon est inférieure à 30

La formule à utiliser est : $m - t_{\alpha,n-1}\frac{s}{\sqrt{n}}; m + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$

$t_{\alpha,\ ddl}$ = $t_{5\%,\ 25\ \_\ 1}$ =2.064

IC $1.52 - 0.08 \times 2.064; m + 0.08 \times 2.064 = [1.355; 1.685]$

2. Pour $t_{1\%}$ IC [1.296; 1.744]

Plus on réduit le risque de se tromper (alpha faible), plus l'IC est large

Exercice 9.2

p=1050/3500=30% = proportion observée sur l'échantillon

$$\text{Ecart type}=Sp=\sqrt{\frac{f(1-f)}{n}}=\sqrt{\frac{0.3(1-0.3)}{3500}}=0.0077$$

$$IC=f-N_\alpha\sqrt{\frac{f(1-f)}{n}};f+N_\alpha\sqrt{\frac{f(1-f)}{n}}$$

Or alpha=5%,
Donc Zα = Nα=1.96

$$p=30\%\pm1.96\times0.0077$$

$$\left[28.5\%-31.5\%\right]$$

Exercice 9.3

$p = 8/12 = 66.7\%$ mais $n(1 - p) = 12 \times (1 - 0.667) = 4 < 5$

On ne peut pas utiliser l'approximation par la loi normale.

Test binomial exact

binom.test(x=1,n=12,p=0.5,alternative = "two.sided", conf.level =0.95)

CI [34.9% à 90.1%]

Exercice 9.4

p=10/70=0.143

$$Sp = \sqrt{\frac{0.143(1-0.143)}{70}} = 0.034$$

np>5 et nq>5

IC95%=0.143±1.96x0.034

[7.6% - 21.0%]

Exercice 9.5

Il s'agit de calculer la taille de la population nécessaire pour estimer une proportion sur un échantillon avec une précision souhaiter

$$n = P(1-P)\frac{Z_{\alpha}^{\ 2}}{i^2}$$

Il faut donc :

- Connaitre approximativement la fréquence attendue *P*
- *Choisir la précision souhaitée i (moitié de l'intervalle de confiance)*
- *Choisir le risque α consenti de se tromper*

Merci