

1^{ère} baccalauréat en Sciences Vétérinaires

Biostatistiques – Travaux dirigés

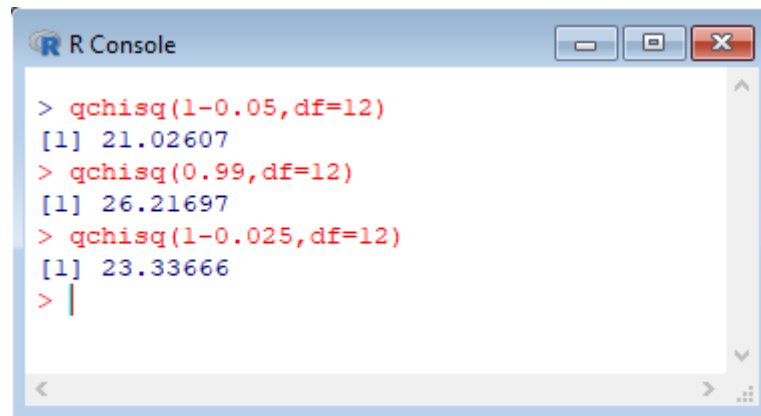
Séance de TD n°3 : Résolutions

1. Pour une distribution de chi-carré à 12 ddl, trouver la valeur de chi-carré telle que :

- a) L'aire à droite de cette valeur soit égale à 0,05
- b) L'aire à gauche de cette valeur soit égale à 0,99
- c) L'aire à droite de cette valeur soit égale à 0,025

Solution

C'est évidemment très facile avec R :



```

R Console
> qchisq(1-0.05,df=12)
[1] 21.02607
> qchisq(0.99,df=12)
[1] 26.21697
> qchisq(1-0.025,df=12)
[1] 23.33666
> |
  
```

Sans logiciel, il faut utiliser la table de χ^2 :

Table de Chi-carré								
ddl	$c^2_{0,005}$	$c^2_{0,025}$	$c^2_{0,05}$	$c^2_{0,90}$	$c^2_{0,95}$	$c^2_{0,975}$	$c^2_{0,99}$	$c^2_{0,995}$
1	0,0000393	0,000982	0,00393	2,706	3,841	5,024	6,635	7,879
2	0,01	0,051	0,103	4,605	5,991	7,378	9,21	10,597
3	0,072	0,216	0,352	6,251	7,815	9,348	11,345	12,838
4	0,207	0,484	0,711	7,779	9,488	11,143	13,277	14,86
5	0,412	0,831	1,145	9,236	11,07	12,832	15,086	16,75
6	0,676	1,237	1,635	10,645	12,592	14,449	16,812	18,548
7	0,989	1,69	2,167	12,017	14,067	16,013	18,475	20,278
8	1,344	2,18	2,733	13,362	15,507	17,535	20,09	21,955
9	1,735	2,7	3,325	14,684	16,919	19,023	21,666	23,589
10	2,156	3,247	3,94	15,987	18,307	20,483	23,209	25,188
11	2,603	3,816	4,575	17,275	19,675	21,92	24,725	26,757
12	3,074	4,404	5,226	18,549	21,026	23,337	26,217	28,3
13	3,565	5,009	5,892	19,812	22,362	24,736	27,688	29,819

Les valeurs recherchées ont été bleutées.

2. On procède à 200 parties de pile ou face avec une pièce de monnaie. On observe 115 faces et 85 piles. Tester si la pièce est parfaite au seuil 5%.

Solution

L'hypothèse nulle est que la pièce est parfaite, ce qu'on traduit par :

$$H_0 : \pi(\text{pile}) = \pi(\text{face}) = 0.50$$

Cette hypothèse nulle se traduit par des effectifs attendus valant 100 pour le nombre de faces et 100 pour le nombre de piles.

Les effectifs observés étant 115 et 85, respectivement, on peut les comparer aux attendus par un χ^2 :

$$\chi^2 = \frac{(115 - 100)^2}{100} + \frac{(85 - 100)^2}{100} = 4.5$$

On peut confronter cette valeur avec le seuil qu'un χ^2 avec un seul degré de liberté (rappelons que le nombre de degrés de liberté est égal au nombre de catégories moins un, soit $2 - 1 = 1$, ici) n'excède que dans $\alpha = 5\%$ des cas. Cette valeur s'obtient de la table de χ^2 (voir exercice 1) ou via R ou Excel. On trouve que $\chi_{seuil}^2(\alpha = 5\%) = 3.841$.

Comme la valeur observée (4.5) dépasse le seuil, la valeur observée a une probabilité inférieure à 5% sous l'hypothèse nulle, qui est dès lors rejetée à ce seuil : le nombre de faces de cette pièce est significativement plus élevé que le nombre de piles (au seuil $\alpha = 5\%$).

3. Afin de tester l'efficacité d'un régime, un expérimentateur souhaite connaître la distribution (à priori, c'est-à-dire sans régime) du "gain quotidien moyen (GQM)" dans une population bovine. Il fait les mesures sur 5 groupes de 20 animaux: le gain moyen (pour les 5 groupes) est de 1400 grammes, et la déviation standard de la moyenne pour les cinq mesures est de 100 grammes (erreur standard).

Quelle proportion de groupes de 20 bovins de cette population aura-t-elle un GQM > 2000 grammes (lire: dans quel pourcentage de cas la valeur moyenne dépasse-t-elle 2000 g) ?

Quelle est la déviation standard dans la population bovine (données individuelles) ?

En supposant la distribution des GQM individuels normale, quelle proportion de bovins aura un GQM > 2000 grammes ?

Solution

On supposera les moyennes normalement distribuées (rappelons que les moyennes tendent vers une distribution normale, en vertu du théorème de la limite centrale), avec une moyenne (des moyennes) $\mu_{\bar{x}} = 1400$ g et une déviation standard (des moyennes) valant $\sigma_{\bar{x}} = 100$ g.

a) On peut calculer la valeur de Z correspondant à la première question :

$$Z = \frac{2000 - 1400}{100} = 6$$

On en déduit que la probabilité que le GQM dépasse 2000 g est égale à la probabilité que Z dépasse 6. On voit dans la table de Z que cette probabilité est extrêmement faible (en fait, la table de Z s'arrête à $Z = 3 \dots$).

Si on veut la valeur, on peut utiliser un logiciel. Par exemple, avec R :

$1-\text{pnorm}(6) \Rightarrow 9.865877\text{e-}10$

b) L'erreur standard dépend de la taille des échantillons : $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$

On en déduit que : $\sigma_X = \sqrt{n} * \sigma_{\bar{X}}$

Dans notre problème, on a donc :

$$\sigma_X = \sqrt{20} * 100 = 447.2136 \text{ g}$$

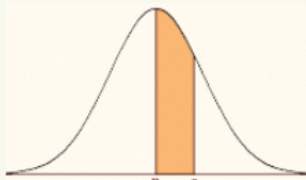
c) On peut faire le même calcul qu'au point a, mais en utilisant la déviation standard plutôt que l'erreur standard :

$$Z = \frac{2000 - 1400}{447.2136} = \frac{6}{\sqrt{20}} = 1.3416$$

La probabilité recherchée est donc égale à la probabilité que Z excède 1.3416, ce qu'on obtient facilement dans la table de Z :

$$P(Z > 1.34) = 0.5 - P(0 < Z < 1.34) = 0.5 - 0.4099 = 0.0901$$

Table des surfaces de la variable z



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4333	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441

4. Un chercheur désire savoir si une nouvelle molécule a un effet sur la pression sanguine systolique. Si l'effet sur la moyenne de la population est de 10 mm Hg (faisant par exemple passer la moyenne de 100 mm Hg à 90 mm Hg), quelle devra être la taille de l'échantillon à utiliser pour permettre de détecter cet effet avec une puissance de 80% ?

Solution

Le seuil de signification utilisé n'est pas donné, nous utiliserons donc $\alpha = 5\%$. Par ailleurs, la puissance valant $(1 - \beta)$, on en déduit la valeur de β souhaitée : $1 - \beta = 0.8 \Rightarrow \beta = 1 - 0.8 = 0.2$.

Une première équation vient de la contrainte sur α : si H_0 est vraie, la probabilité de descendre à une pression inférieure à la pression seuil P_s est 5%, ce qui correspondra à une valeur de Z de -1.645 :

$$Z_{s0} = \frac{P_s - \mu_0}{\sigma / \sqrt{n}} = -1.645 \Rightarrow P_s = \mu_0 - 1.645 * \sigma / \sqrt{n}$$

Une seconde équation vient de la contrainte sur β : si H_1 est vraie, la probabilité d'une pression supérieure à la pression seuil P_s est 20%, ce qui correspondra à une valeur de Z de 0.84 (voir la table de Z) :

$$Z_{s1} = \frac{P_s - \mu_1}{\sigma / \sqrt{n}} = 0.84 \Rightarrow P_s = \mu_1 + 0.84 * \sigma / \sqrt{n}$$

Egalant les deux expressions de P_s , on obtient après quelques manipulations :

$$\sqrt{n} = \frac{(1.645 + 0.84) * \sigma}{\mu_0 - \mu_1}$$

Ce qui donne la valeur de n en élevant au carré. On remarque que n dépend de α (via 1.645), de β (via 0.84), de la taille de l'effet (via $\mu_0 - \mu_1$) et de la variation intra-population (via σ).

5. Un nombre est tiré au hasard entre 0 et 10, et chaque valeur est à priori équiprobable (distribution uniforme). Si on tire des ensembles de 9 nombres et qu'on en calcule la moyenne, donnez la distribution de ces moyennes. (Rappel: $\mu = E[x]$, et $\sigma^2 = E[(x - \mu)^2]$)

Solution

Le théorème de la limite centrale nous dit que la distribution de moyennes d'échantillons provenant de distributions quelconques tend vers une distribution normale de même moyenne que la distribution dont sont pris les échantillons, et de variance égale à la variance de la population initiale divisée par la taille des échantillons.

Plus mathématiquement : $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$

Pour répondre à la question, il faut donc connaître μ_X et σ_X .

On utilise les définitions de la moyenne et de la variance pour calculer ces grandeurs :

$$\mu_X = E[X] = \int_0^{10} x * f(x) * dx$$

Une distribution uniforme est une fonction constante. Cette constante vaut 0.1 (rectangle de surface 1 et de coté 10).

On en déduit que :

$$\mu_X = 0.1 * \int_0^{10} x * dx = 0.05 * x^2|_0^{10} = 5$$

Pour la variance :

$$\sigma_X^2 = E[(X - \mu_X)^2] = 0.1 * \int_0^{10} (x - 5)^2 * dx = 0.1 * \frac{(x - 5)^3}{3} \Big|_0^{10} = \frac{25}{3}$$

On en déduit que la distribution des moyennes sera :

$$\bar{X} \sim N\left(5, \frac{5}{3 * \sqrt{3}}\right)$$

6. Sur 50 veaux culards du sexe mâle, 34 sont nés par césarienne; sur 40 veaux culards du sexe femelle, 11 sont nés par césarienne. Le sexe influence-t-il les circonstances du vêlage ? Comment exprimez-vous votre H0? Dans votre conclusion, quel type d'erreur commettez-vous et avec quelle probabilité ?

Solution

L'hypothèse nulle est que le sexe n'influence pas le taux de césarienne : ce taux est donc le même chez les mâles et chez les femelles.

On peut mettre les données sous la forme d'un tableau :

Observés	Mâles	Femelles	Totaux
Césariennes	34	11	45
Naturel	16	29	45
Total	50	40	90

(Les valeurs en gras ont été déduites par différences)

On peut facilement calculer les attendus si H0 est vraie :

Attendus	Mâles	Femelles	Totaux
Césariennes	25	20	45
Naturel	25	20	45
Total	50	40	90

On peut en déduire une valeur de χ^2 :

$$\chi^2 = \frac{(34 - 25)^2}{25} + \frac{(16 - 25)^2}{25} + \frac{(11 - 20)^2}{20} + \frac{(29 - 20)^2}{20} = 14.58$$

Cette valeur de χ^2 (avec 1 degré de liberté) est supérieure au seuil qu'on ne dépasse que dans 5% des cas, à savoir 3.841 (voir la table de χ^2 avec 1 ddl).

Par conséquent, la probabilité d'une telle valeur est très faible, et en tous cas inférieure à 5% (on peut calculer par exemple avec R que la probabilité d'atteindre ou dépasser une telle valeur vaut 0.000134), et conduit à rejeter l'hypothèse nulle.

L'erreur potentielle dans un rejet de H_0 est une erreur de type I, qui est très faible, ici ($p = 0.000134$).

7. L'examen à l'abattage de 1000 porcs a révélé que 300 avaient des lésions pulmonaires et 100 des lésions de la plèvre. 620 étaient absolument indemnes de toute lésion.

Y a-t-il indépendance entre les lésions du poumon et les lésions de la plèvre ? Ou encore, un individu ayant des lésions pulmonaires a-t-il plus de chance d'avoir en même temps des lésions de la plèvre qu'un individu n'en ayant pas ?

Solution

On peut représenter ces résultats à l'aide d'une table de contingences :

Observés	Lésion plèvre	Plèvre indemne	Totaux
Lésion pulmonaire	20	280	300
Poumon indemne	80	620	700
Total	100	900	1000

(Les valeurs en gras ont été déduites par différences)

L'hypothèse nulle d'indépendance entre les deux types de lésions conduit à la table des attendus :

Globalement, sur l'échantillon, les lésions de la plèvre sont de $100/1000 = 0,1$.

Attendus	Lésion plèvre	Plèvre indemne	Totaux
Lésion pulmonaire	$0,1 \cdot 300 = 30$	$300 - 30 = 270$	300
Poumon indemne	$0,1 \cdot 700 = 70$	$700 - 70 = 630$	700
Total	100	900	1000

Le calcul conduit à une valeur :

$$\chi^2 = \frac{(20 - 30)^2}{30} + \frac{(280 - 270)^2}{270} + \frac{(80 - 70)^2}{70} + \frac{(620 - 630)^2}{630} = 5.291$$

Cette valeur, supérieure à 3.841, conduit à rejeter l'hypothèse d'indépendance entre les deux types de lésions : il semble y avoir significativement (au seuil 5%) moins de lésion de la plèvre lorsqu'il y a lésion pulmonaire (et vice-versa).

8. On a comparé des mères primipares et des mères multipares pour la répartition des poids à la naissance, dans 3 classes :

Poids	Primipares	Multipares	Totaux
< 3 kg	26	20	46
3-4 kg	61	63	124
> 4 kg	8	22	30
Totaux	95	105	200

Testez l'indépendance entre ces deux variables, parité et poids à la naissance ? Justifiez la validité du test choisi. Formulez concrètement votre conclusion.

Solution

Ces résultats sont présentés à l'aide d'une table de contingences. On peut en déduire les effectifs attendus sous l'hypothèse nulle d'indépendance entre poids et parité considérant que $95/200 = 0,475$

Attendus	Primipares	Multipares	Totaux
< 3 kg	$0,475 \cdot 46 = 21.85$	$46 - 21.85 = 24.15$	46
3-4 kg	$0,475 \cdot 124 = 58.90$	$124 - 58.9 = 65.10$	124
> 4 kg	$0,475 \cdot 30 = 14.25$	$30 - 14.25 = 15.75$	30
Totaux	95	105	200

Puisque tous les attendus sont suffisamment grands (> 10), l'écart entre attendus et observés se calcule par un χ^2 :

$$\chi^2 = \frac{(26 - 21.85)^2}{21.85} + \dots + \frac{(22 - 15.75)^2}{15.75} = 6.8654$$

Pour cette table 3 x 2, il y a $(3-1) \times (2-1) = 2$ degrés de liberté. La valeur qu'on ne dépasse que dans 5% des cas dans la distribution de χ^2 avec 2 degrés de liberté est 5.99. La valeur obtenue est supérieure à ce seuil, indiquant un écart anormalement grand par rapport à ce que prévoyait l'hypothèse nulle, qui est donc rejetée : il y a donc dépendance entre la parité et le poids des produits dans cette espèce.

9. Trois préparations d'un même vaccin contre la maladie de Marek ont assuré la protection suivante :

- La 1ère: 16 volailles atteintes sur 54 vaccinées
- La 2ème: 25 volailles atteintes sur 55 vaccinées
- La 3ème: 4 volailles atteintes sur 38 vaccinées.

Ces 3 préparations ont elles des pouvoirs protecteurs différents ?

Solution :

On peut représenter ces résultats à l'aide d'une table de contingences :

Observés	Atteints	Indemnes	Totaux
Préparation I	16	38	54
Préparation 2	25	30	55
Préparation 3	4	34	38
Total	45	102	147

(Les valeurs en gras ont été déduites)

L'hypothèse nulle d'indépendance entre les deux types de lésions conduit à la table des attendus :

Attendus	Atteints	Indemnes	Totaux
Préparation I	16.531	37.469	54
Préparation 2	16.837	38.163	55
Préparation 3	11.633	26.367	38
Total	45	102	147

Le calcul conduit à une valeur :

$$\chi^2 = \frac{(16 - 16.531)^2}{16.531} + \dots + \frac{(34 - 26.367)^2}{26.367} = 12.946$$

Pour cette table 3 x 2, il y a (3-1) x (2-1) = 2 degrés de liberté. La valeur qu'on ne dépasse que dans 5% des cas dans la distribution de χ^2 avec 2 degrés de liberté est 5.99. La valeur obtenue est supérieure à ce seuil, indiquant un écart anormalement grand par rapport à ce que prévoyait l'hypothèse nulle, qui est donc rejetée : il y a donc des différences (très) significatives entre préparations (en termes de qualité, il semble que la préparation 3 est la meilleure, suivie de la préparation 1).

10. Cent patients ayant été victimes d'un infarctus du myocarde ont été suivis pendant 8 ans. La moitié d'entre eux ont été soumis à un régime strict (pauvre en matières grasses, riche en protéines et en vitamines), tandis que l'autre moitié gardait toute liberté dans le choix de ses aliments. Après 8 ans, on a observé 28 survivants dans le groupe à régime strict et 12 survivants dans le groupe à régime libre. Un régime strict est-il indiqué après infarctus du myocarde ?

Solution :

On peut représenter ces données sous la forme d'une table de contingences :

Observés	Survie	Mort	Totaux
Régime strict	28	22	50
Régime libre	12	38	50
Total	40	60	100

(Les valeurs en gras ont été déduites par différences)

L'hypothèse nulle d'indépendance entre régime et infarctus conduit à la table des attendus :

Attendus	Survie	Mort	Totaux
Régime strict	0.4 * 50 = 20	50 - 20 = 30	50
Régime libre	20	30	50
Total	50	50	100

Le calcul conduit à une valeur :

$$\chi^2 = \frac{(28 - 20)^2}{20} + \frac{(12 - 20)^2}{20} + \frac{(22 - 30)^2}{30} + \frac{(38 - 30)^2}{30} = 10.667$$

Cette valeur, supérieure à 3.841, conduit à rejeter l'hypothèse d'indépendance entre infarctus et régime : il semble y avoir significativement (au seuil 5%) plus d'infarctus parmi les individus ayant subi une régime « libre ».

11. Cuénot a obtenu une F2 de monohybridisme composée de 198 souris grises et 72 souris albinos, le croisement parental étant souris grise X souris albinos. Testez l'hypothèse 3/4 : 1/4 (le caractère blanc est récessif).

Si l'on fait le tableau des croisements, on obtient :

	G	G
b	Gb	Gb
b	Gb	bb

H0 : la répartition de la couleur des robes est de ¾ : ¼

H1 : la répartition de la couleur des robes est différente de ¾ : ¼

Si l'on reprend les effectifs observés et que l'on estime les effectifs attendus selon H_0 , on obtient :

<u>Effectifs observés</u>	
Blanches bb	72
Grises Gb	198
Total	270

<u>Effectifs attendus</u>	
Blanches bb	67,5 = 1/4
Grises Gb	202,5 = 3/4
Total	270

Les effectifs sont suffisants pour appliquer un test de χ^2 :

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Avec k le nombre de catégories -1 (= ddl). Ici, $k = 2 - 1 = 1$

$$\chi^2 = \frac{(72 - 67,5)^2}{67,5} + \frac{(198 - 202,5)^2}{202,5} = 0,4$$

Or, χ^2_{seuil} pour $\alpha = 5\%$ et 1 ddl est égal à 3,841.

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , la répartition est bien de $\frac{3}{4}$: $\frac{1}{4}$

12. Dans une région exposée à une maladie, une substance susceptible de protéger les individus contre la pathologie est testée et les résultats obtenus sont reportés dans la table suivante:

	Atteints	Non atteints	Total
Protégés	2	12	14
Non protégés	8	6	14
Total	10	18	28

Si la substance n'a aucun effet, on attendrait des proportions identiques d'atteints dans les deux groupes, "protégés" et "non protégés", à savoir 5 sur 14. Bien entendu, des fluctuations aléatoires d'échantillonnage pourraient faire dévier les effectifs réellement observés des valeurs moyennes attendues. On demande :

a) De calculer la probabilité de ce qui a été observé si l'hypothèse nulle (pas d'effet de la substance) est vraie, ce qui revient à calculer la probabilité, quand on prélève 10 individus dans un groupe total de 28, que 2 appartiennent à un sous-groupe de 14 et les 8 autres à un autre sous-groupe de 14.

b) De calculer les probabilités des situations qui s'écartent encore plus de la situation correspondant à l'hypothèse nulle (c'est à dire où le nombre d'atteints parmi les protégés est <2, le nombre total d'individus(28), de protégés(14), de non protégés(14), d'atteints(10) et d'indemnes (18) restant inchangés).

c) A quoi correspond la somme des probabilités calculées dans les deux points précédents et qu'en concluez-vous ?

a) Les effectifs attendus ne permettent pas d'appliquer un test de χ^2 puisqu'ils sont inférieurs à 10 dans plusieurs catégories.

En revanche, nous pouvons utiliser un test exact de Fisher en appliquant la loi hypergéométrique puisque nous avons 2 sous-ensembles contenant 2 catégories.

La loi hypergéométrique nous indique que, dans un ensemble de taille N, constitué de deux sous-ensembles de tailles N_1 et $N_2 = N - N_1$, la probabilité de prélever a_1 ($< N_1$) individus dans l'ensemble de taille N_1 quand on prélève un total de a ($< N$) individus vaut :

$$P(a_1; N_1; a; N) = \frac{C_{N_1}^{a_1} * C_{N_2}^{a_2}}{C_N^a} \text{ avec } a_1 + a_2 = a$$

$$\text{Et } C_N^a = \frac{N!}{a! * (N-a)!}$$

Ici, on définit par exemple :

- N_1 le nombre de protégés = 14
- a_1 le nombre d'atteints parmi les protégés = 2
- N_2 le nombre de non protégés = 14
- a_2 le nombre d'atteints parmi les non protégés = 8
- N correspond donc aux effectifs totaux = 28
- a correspond donc à $a_1 + a_2 = 10$

Ainsi, on a :

$$P(2; 14; 10; 28) = \frac{C_{14}^2 * C_{14}^8}{C_{28}^{10}} = \frac{\frac{14!}{2! * 12!} * \frac{14!}{8! * 6!}}{\frac{28!}{10! * 18!}} = 0,0208$$

b) On nous demande les probabilités que le nombre d'atteints parmi les protégés soit de moins de deux, c'est-à-dire de calculer $P(a_1 = 1)$ et $P(a_1 = 0)$.

Si nous réalisons les tables, on obtient :

	Atteints	Non atteints	Total
Protégés	1	13	14
Non protégés	9	5	14
Total	10	18	28

$$P(1; 14; 10; 28) = \frac{C_{14}^1 * C_{14}^9}{C_{28}^{10}} = \frac{\frac{14!}{1! * 13!} * \frac{14!}{9! * 5!}}{\frac{28!}{10! * 18!}} = 0,0021$$

	Atteints	Non atteints	Total
Protégés	0	14	14
Non protégés	10	4	14
Total	10	18	28

$$P(0; 14; 10; 28) = \frac{C_{14}^0 * C_{14}^{10}}{C_{28}^{10}} = \frac{\frac{14!}{0! * 14!} * \frac{14!}{10! * 4!}}{\frac{28!}{10! * 18!}} = 0,00007628$$

c) $P(a_1 = 0) + P(a_1 = 1) + P(a_1 = 2) = P(a_1 < 3)$ puisque nous sommes face à une distribution discrète.

Ainsi, $P(a_1 < 3) = P(a_1 = 0) + P(a_1 = 1) + P(a_1 = 2) = 0,00007628 + 0,0021 + 0,0208 = 0,02298$

La probabilité calculée est la probabilité de s'écarter aussi fort, voire plus fort, de ce qui était attendu si l'hypothèse nulle est vraie. Comme cette probabilité est inférieure à $\alpha = 5\%$, on rejette H_0 : il y a un effet de la substance.

13. Un généticien désire déterminer si un caractère a un déterminisme de type dominant. Pour réaliser son expérience, il effectue 100 croisements d'individus exhibant le caractère, et obtient 70 individus eux-mêmes porteurs du caractère. Expliquez la démarche, et les conclusions de l'expérience.

En cas de déterminisme de type dominant, on s'attend à obtenir une répartition de $\frac{3}{4} : \frac{1}{4}$ en cas de croisement entre deux hétérozygotes Ww (avec W le caractère dominant et w le récessif) :

	W	w
w	Ww	ww
W	WW	Ww

Avec WW ou Ww = caractère exprimé = $\frac{3}{4}$ et ww = caractère non exprimé = $\frac{1}{4}$

Ainsi, les hypothèses sont les suivantes :

H0 : la répartition sera de $\frac{3}{4}$: $\frac{1}{4}$

H1 : la répartition sera différente de $\frac{3}{4}$: $\frac{1}{4}$

Si l'on effectue 100 croisements et que l'on part du principe que chaque croisement a généré un individu, on obtient :

Effectifs observés	
Caractère	70
Absence caractère	30
Total	100

Effectifs attendus	
Caractère	75
Absence caractère	25
Total	100

$$\chi^2 = \frac{(70 - 75)^2}{75} + \frac{(30 - 25)^2}{25} = 1,333$$

Or, χ^2_{seuil} pour $\alpha = 5\%$ et 1 ddl est égal à 3,841.

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H0, la répartition est bien de $\frac{3}{4}$: $\frac{1}{4}$ donc le caractère est bien de type dominant

Rq : il n'est pas dit que les deux parents sont hétérozygotes, c'est donc une hypothèse incluse dans H0. En réalité, on pourrait avoir croisé des WW avec des WW, des WW avec des Ww et des Ww avec des Ww : si on ne connaît pas les proportions pour chaque type de croisement, il est en fait difficile de répondre à la question)

14. Une expérience portant sur l'effet d'une nouvelle molécule sensée augmenter la protection contre une pathologie a été menée. 50 individus ont reçu le vaccin, 50 ont reçu un placebo. 20 individus ont manifesté des signes de la maladie, parmi lesquels 8 avaient reçu la molécule. La molécule a-t-elle réellement un effet ?

On peut poser les hypothèses de la manière suivante :

H0 : il n'y a pas d'effet du vaccin : effectifs observés = effectifs attendus

H1 : il y a un effet du vaccin : effectifs observés $\neq$ effectifs attendus

Si l'on remplit les tables des effectifs observés, on obtient :

Effectifs observés			
	Signes cliniques	Absence signes	Total
Vaccin	8	42	50
Placebo	12	38	50
Total	20	80	100

Considérant que $50/100 = 0,5$ (effectifs vaccin/effectifs totaux et effectifs placebo/effectifs totaux), on calcule les effectifs théoriques :

Effectifs théoriques			
	Signes cliniques	Absence signes	Total
Vaccin	$20 \times 0,5 = 10$	$50 - 10 = 40$	50
Placebo	$20 \times 0,5 = 10$	$50 - 10 = 40$	50
Total	20	80	100

$$\chi^2 = \frac{(8 - 10)^2}{10} + \frac{(42 - 40)^2}{40} + \frac{(12 - 10)^2}{10} + \frac{(38 - 40)^2}{40} = 1$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl (puisque $(\text{lignes} - 1) \times (\text{colonnes} - 1) = (2-1) \times (2-1) = 1$)

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , il n'y a pas d'effet du vaccin.

15. Certains chercheurs ont émis l'hypothèse selon laquelle le développement du cancer du sein pourrait être lié à l'âge lors de la première naissance. Les données suivantes ont été collectées :

	Age à la première naissance					
Cas/Contrôle	<20	20-24	25-29	30-34	>34	Total
Cas	320	1206	1011	463	220	3220
Contrôle	1422	4432	2893	1092	406	10245
Total	1742	5638	3904	1555	626	13465

Y a-t-il réellement un effet de l'âge ? Si oui, dans quel sens s'exerce-t-il ?

On peut poser les hypothèses de la manière suivante :

H_0 : il n'y a pas d'effet de l'âge à la première naissance sur le développement du cancer du sein

H_1 : il y a un effet de l'âge à la première naissance sur le développement du cancer du sein (→ test bilatéral puisque le sens de l'éventuel lien n'est pas précisé)

On pourrait postuler, considérant les données du tableau, que l'hypothèse alternative est (mais, en principe, la question posée ne fait pas intervenir les données, dont la récolte est ultérieure...)

H1 : il y a une augmentation du développement du cancer du sein avec l'augmentation de l'âge à la première naissance → test unilatéral

Pour obtenir les effectifs observés, on peut considérer les facteurs de multiplication suivants pour les effectifs totaux de chaque catégorie d'âge à la première naissance :

Classes	<20	20-24	25-29	30-34	>34
Calcul	1742/13465	5638/13465	3904/13465	1555/13465	656/13465
Résultat	0,1293724	0,4187152	0,2899369	0,1154846	0,0464909

Ainsi, on obtient les effectifs théoriques suivants :

Effectifs attendus						
	Age à la première naissance					
Cas/Contrôle	<20	20-24	25-29	30-34	>34	Total
Cas	3220* 0,1293	3220* 0,4187	3220* 0,2899	3220*0,1155	3220* 0,0465	3220
Contrôle	10245* 0,1293	10245* 0,4187	10245* 0,2899	10245*0,1155	10245* 0,0465	10245
Total	1742	5638	3904	1555	626	13465

Effectués sur excel, on obtient les résultats arrondis :

	Age à la première naissance					
Cas/Contrôle	<20	20-24	25-29	30-34	>34	Total
Cas	416,58	1348,26	933,60	371,86	149,70	3220
Contrôle	1325,42	4289,74	2970,40	1183,14	476,30	10245
Total	1742	5638	3904	1555	626	13465

A noter que, pour la dernière classe (> 34), on peut aussi faire :

$$Cas = 3220 - (416,58 + 1348,26 + 933,6 + 371,86) = 149,7$$

$$Contrôle = 10245 - (41324,42 + 4289,74 + 2970,40 + 1183,14) = 476,3$$

On peut maintenant calculer le χ^2 observé :

$$\chi^2_{\text{observé}} = \frac{(320 - 416,58)^2}{416,58} + \frac{(1206 - 1348,26)^2}{1348,26} + \dots + \frac{(1092 - 1183,14)^2}{1183,14} + \frac{(406 - 476,30)^2}{476,30}$$

$$\chi^2_{\text{observé}} = 130,34$$

Or, si l'on fait un test unilatéral, $\chi^2_{\text{seuil}} = 9,488$ pour $\alpha = 5\%$ et 4 ddl (puisque (lignes - 1)* (colonnes - 1) = (2-1)*(5-1) = 4)

Ainsi $\chi^2_{\text{observé}} \gg \chi^2_{\text{seuil}}$: on rejette H_0 , l'augmentation de l'âge à la première naissance a un effet sur le nombre de cancers du sein. On peut voir dans les tableaux que, parmi les cas, les attendus pour les classes d'âge les plus jeunes sont plus élevés que les effectifs observés (âge < 24 ans), et que c'est l'inverse pour les classes les plus âgées, ce qui semble indiquer un effet délétère de l'âge à la naissance du premier enfant sur l'apparition de cancers du sein.

16. Dans un sondage sur 1000 personnes, combien de personnes doivent se déclarer "pour" pour pouvoir conclure à une majorité significative de "pour" (au seuil 5%). Effectuez le calcul avec trois méthodes.

H_0 : il n'y a pas de différence entre le nombre de « pour » et de « contre »

H_1 : il y a une différence entre le nombre de « pour » et de « contre »

Trois possibilités :

- 1) Utilisation du χ^2 (test bilatéral)
- 2) Utilisation de la loi binomiale (test unilatéral ou bilatéral) car distribution discrète avec 2 possibilités : oui ou non
- 3) Utilisation de l'approximation normale (test unilatéral ou bilatéral) comme n est grand

1) Utilisation du χ^2

On peut représenter les effectifs observés et attendus de la manière suivante :

<u>Effectifs observés</u>	
Pour	n
Contre	1000-n
Total	1000

<u>Effectifs attendus</u>	
Pour	500
Contre	500
Total	1000

Ainsi,

$$\chi^2_{\text{obs}} = \sum_{i=1}^1 \frac{(O_i - 500)^2}{500} = \frac{(n - 500)^2}{500} + \frac{((1000 - n) - 500)^2}{500}$$

$$\chi_{obs}^2 = \frac{(n-500)^2}{500} + \frac{(500-n)^2}{500} \Leftrightarrow \chi_{obs}^2 = 2 * \frac{(n-500)^2}{500}$$

Puis l'on remplace la valeur de χ_{obs}^2 par $\chi_{théo}^2$ afin de déterminer n. $\chi_{théo}^2 = 3,841$ pour 1 ddl et

$$\alpha = 5\% \quad \chi_{théo}^2 = 2 * \frac{(n-500)^2}{500} \Leftrightarrow \frac{\chi_{obs}^2}{2} = \frac{(n-500)^2}{500} \Leftrightarrow \frac{3,841}{2} = \frac{(n-500)^2}{500}$$

$$1,921 * 500 = (n-500)^2 \Rightarrow \leq 960,5 = (n-500)^2 \Rightarrow \leq$$

$$(n-500)^2 - \sqrt{960,5}^2 = (n-500 - \sqrt{960,5}) * (n-500 + \sqrt{960,5}) = 0$$

$$\Rightarrow (n-530,99) * (n-469,01) = 0$$

$$n = 530,99 \quad \text{ou} \quad n = 469,01$$

Il faudrait donc soit plus de 530 voix « pour » afin d'avoir une majorité significative de « pour » au seuil 5%, soit moins de 470 afin d'avoir une majorité significative de « contre » au seuil 5%.

2) Utilisation de la loi binomiale

On connaît, selon H0 :

- n = 1000
- p = 0,5

Il faut donc déterminer pour quelle valeur seuil r_s de r il y aura une majorité de « pour » au seuil 5%, soit $P(r \geq r_s) \leq 0.05$

Cela serait extrêmement long à calculer à la main si l'on doit tester toutes les valeurs de r...

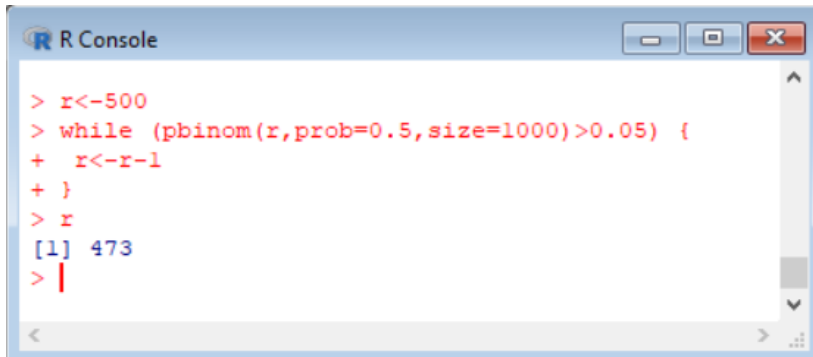
On peut cependant faire le calcul à l'aide d'un logiciel tel qu'excel.

=1-LOI.BINOMIALE.N(D557;1000;0,5; VRAI)	
D	E
valeur r	1- Probabilité cumulée
520	0,097383164
521	0,086934658
522	0,077346853
523	0,068584003
524	0,060607133
525	0,053374771
526	0,046843646
527	0,040969351
528	0,035706962
529	0,031011598
530	0,026838925

$P(r > 526) < 0,05$ (ou inversement $P(r \leq 526) > 0,95$).

Il faut donc au moins 526 voix « pour » pour pouvoir conclure à une majorité significative. On pourrait faire le calcul unilatéral permettant de trouver le seuil en dessous duquel on décrèterait une majorité de contre de manière similaire : $P(r \leq r_s) \leq 0.05$

Faisons le calcul avec R, cette fois (pour changer...) :



```

R Console
> r<-500
> while (pbinom(r,prob=0.5,size=1000)>0.05) {
+   r<-r-1
+ }
> r
[1] 473
> |

```

Il y aura donc une majorité significative de « contre » (au seuil 5%) si le nombre de « pour » est inférieur ou égal à 473.

3) Utilisation de l'approximation normale de la loi binomiale (test unilatéral)

$$Z = \frac{r - n * p}{\sqrt{n * p * q}}$$

et

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

$$\mu = n * p = 1000 * 0,5 = 500$$

Pour tester s'il y a une majorité de « pour » :

$P(Z > z_1) = 0,05$ donc Z sera positif

et $P(0 < Z < z_1) = P(Z > 0) - P(Z > z_1) = 0,5 - 0,05 = 0,45$

Si on lit la table à l'envers, $Z = 1,645$.

$$1,645 = \frac{X - 500}{\sigma} \Leftrightarrow 1,645 = \frac{r - 500}{\sqrt{1000 * 0,5 * 0,5}}$$

$$1,645 * \sqrt{250} = r - 500 \Leftrightarrow r = 1,645 * 15,811 + 500$$

$$r = 526,009$$

Il faut au moins que 526 personnes se déclarent « pour » afin d'obtenir une majorité significative au seuil 5%.

Les calculs sont similaires pour tester s'il y a une majorité significative de « contre ».

17. Dans une expérience d'inoculation d'un agent pathogène, 3 animaux ayant absorbé une substance protectrice n'ont montré aucun signe de maladie au terme de l'expérience, alors que deux ont été atteints. Parallèlement, 4 autres animaux exposés sans protection sont tous atteints. La protection est-elle réellement efficace ?

H_0 : il n'y a pas de différence entre les animaux qui ont absorbé la substance protectrice et ceux qui ne l'ont pas absorbé : la substance n'a pas d'effet.

H_1 : il y a une différence entre les animaux qui ont absorbé la substance protectrice et ceux qui ne l'ont pas absorbé : la substance a un effet.

Les effectifs attendus ne permettent pas d'appliquer un test de χ^2 puisqu'ils sont inférieurs à 10 dans plusieurs catégories.

En revanche, nous pouvons utiliser un test exact de Fisher en appliquant la loi hypergéométrique.

On obtient la table suivante :

	Atteints	Sains	Total
Substance protectrice	2	3	5
Pas de protection	4	0	4
Total	6	3	9

Comme il y a, au total, 2/3 d'atteints, si l'hypothèse nulle est vraie, on attend 2/3 d'atteints parmi ceux qui ont absorbé la substance potentiellement protectrice, soit 3.33... individus. On constate qu'il y a moins d'atteints que prévu, ce qui va dans le sens de l'effet protecteur supposé.

Pour savoir, si l'écart entre ce qui était prévu et ce qui est observé est significatif, on calcule la probabilité de s'écarter aussi fort (table observée), voire plus fort (tables présentant des nombres d'atteints encore plus faibles parmi ceux ayant absorbé la substance) est grande ou petite.

La loi hypergéométrique nous indique que, dans un ensemble de taille N , constitué de deux sous-ensembles de tailles N_1 et $N_2 = N - N_1$, la probabilité de prélever a_1 ($< N_1$) individus dans l'ensemble de taille N_1 quand on prélève un total de a ($< N$) individus vaut :

$$P(a_1; N_1; a; N) = \frac{C_{N_1}^{a_1} * C_{N_2}^{a_2}}{C_N^a} \text{ avec } a_1 + a_2 = a$$

$$\text{Et } C_N^a = \frac{N!}{a! * (N-a)!}$$

Ici, on définit par exemple :

- N_1 le nombre de protégés = 5
- a_1 le nombre d'atteints parmi les protégés = 2
- N_2 le nombre de non protégés = 4
- a_2 le nombre d'atteints parmi les non protégés = 4
- N correspond donc aux effectifs totaux = 9
- a correspond donc à $a_1 + a_2 = 6$

Ainsi, on a :

$$P(2; 5; 6; 9) = \frac{C_5^2 * C_4^4}{C_9^6} = \frac{\frac{5!}{2! * 3!} * \frac{4!}{4! * 0!}}{\frac{9!}{6! * 3!}} = 0,119$$

Comme la probabilité de s'écarter aussi fort est supérieure à 5%, il n'est pas vraiment nécessaire de lui ajouter la probabilité des tables qui s'écarteraient encore plus fort, avec les mêmes effectifs marginaux, de ce qui était attendu. En fait, la table de l'exercice est en réalité la table la plus extrême dans le sens d'un effet protecteur : avec ces totaux marginaux, on ne peut pas avoir moins de 2 atteints parmi les individus ayant ingéré la substance.

$P(2;5;6;9) > 0,05$ donc on accepte H_0 : il n'y a pas d'effet protecteur de la substance démontré dans cette petite (donc, peu puissante) expérience.

18. Un élevage de 50 chiens est visité par un vétérinaire, qui y détecte 10 cas d'une affection. Le vétérinaire conclut à un problème alimentaire, modifie les rations et revisite l'élevage deux mois plus tard. Dans les dix chiens atteints, 6 sont rétablis. Par contre, 2 nouveaux cas sont apparus. L'alimentation a-t-elle un effet significatif ?

Nous devons appliquer un test de McNemar puisque le changement d'alimentation est testé sur les mêmes chiens (entre la visite 1 et la visite 2 deux mois plus tard).

H_0 : Les changements de sains vers malades sont égaux aux changements de malades vers sains : il n'y a pas d'effet de la modification de l'aliment.

H_1 : les changements de sains vers malades sont différents des changements de malades vers sains : il y a un effet de la modification de l'aliment.

La table peut être représentée de la manière suivante :

		Ancien aliment (visite 1)		Total
		Malades	Sains	
Nouvel aliment (visite 2)	Malades	4	2	6
	Sains	6	38	44
	Total	10	40	50

$$\square \chi^2 = \left(a - \frac{a+b}{2}\right)^2 / \frac{a+b}{2} + \left(b - \frac{a+b}{2}\right)^2 / \frac{a+b}{2}$$

$$\Rightarrow \chi^2 = \frac{(a-b)^2}{a+b}$$

Avec:

- a : transition de malades → sains = 6
- b : transition de sains → malades = 2

$$\chi^2 = \frac{(6 - 2)^2}{6 + 2} = 2$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl (puisque (lignes - 1) * (colonnes - 1) = (2-1)*(2-1) = 1)

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , il n'y a pas d'effet du changement d'aliment.

Il y a toutefois un petit problème avec ce test : les attendus sont trop petits pour employer un test de Mac-Nemar...

L'alternative est de raisonner en termes binomiaux : si H_0 est vraie, la probabilité de passer de malade à sain (ou de sain à malade) est $p = 0.5$, et le nombre d'individus concernés est $n = 8$. On peut donc calculer la probabilité binomiale d'avoir 2 sains devenus malades, ou moins (ou, alternativement d'avoir 6 malades devenus sains, ou plus) :

$$P = C_8^0 * 0.5^0 * (1 - 0.5)^8 + C_8^1 * 0.5^1 * (1 - 0.5)^7 + C_8^2 * 0.5^2 * (1 - 0.5)^6 = 37 * 0.5^8 = 0.144$$

H_0 est donc acceptée, puisque $P > \alpha$

19. Une maladie a une prévalence de 10% dans une population donnée. Un échantillon de 200 individus est constitué dans une sous-population, et les animaux sont soumis à un test de détection de la maladie. Le résultat du test est que 32 animaux sont atteints. Peut-on conclure que la sous-population considérée est plus exposée à la maladie ? Comment exprimez-vous votre hypothèse nulle ? Expliquez les types d'erreurs que vous pouvez commettre dans vos conclusions, et donnez leur probabilité.

H_0 : la prévalence de la sous-population est égale à la prévalence de la population : la sous-population n'est pas plus exposée à la maladie.

H_1 : la prévalence de la population est inférieure à la prévalence de la sous-population : la sous-population est plus exposée à la maladie → test unilatéral.

1) Utilisation du χ^2

La prévalence est de 10%, soit 0,1. Ainsi, l'effectif attendu d'atteints = $0,1 * \text{effectif total} = 0,1 * 200 = 20$.

Les effectifs observés et attendus sont donc les suivants :

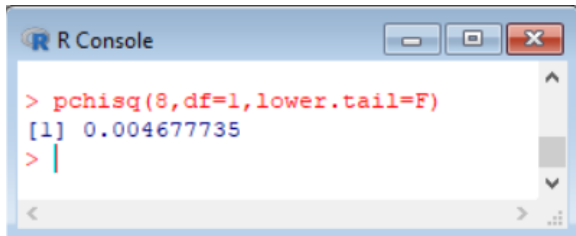
	Sous-population = effectifs observés	Population = effectifs attendus
Atteints	32	$200 * 0,1 = 20$
Sains	$200 - 32 = 168$	$200 - 20 = 180$
Total	200	200

$$\chi^2 = \frac{(32 - 20)^2}{20} + \frac{(168 - 180)^2}{180} = 8$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl.

Ainsi $\chi^2_{\text{observé}} > \chi^2_{\text{seuil}}$: on rejette H_0 , la sous-population est plus exposée à la maladie.

La probabilité (bilatérale) associée à ce χ^2 peut se calculer avec un logiciel :



```
> pchisq(8,df=1,lower.tail=F)
[1] 0.004677735
>
```

En première approximation, la probabilité (unilatérale) correspondant à H_1 vaut la moitié de cette probabilité, soit 0.0023, correspondant à l'erreur de type I pour ce problème (probabilité de rejeter H_0 alors qu'elle est vraie).

20. Dans la population Blanc-Bleu Belge, la couleur de la robe est donnée essentiellement par le génotype au niveau d'un gène. Deux variants (allèles) sont possibles au niveau de ce gène: N, qui a une fréquence de 0,7 dans la population et n qui a une fréquence de 0,3. Il y a donc trois génotypes possibles, qui sont NN, Nn et nn, donnant les robes blanche, bleue et noire respectivement. Dans un échantillon de 1000 individus, 80 sont noirs, 502 sont blancs et 418 sont bleus. Montrez que ces observations sont en accord avec une situation d'indépendance des allèles (situation connue sous le nom d'équilibre de Hardy-Weinberg).

Selon l'équilibre de Hardy-Weinberg, si deux allèles N et n ont des fréquences respectives p et q = (1 - p), les génotypes NN, Nn et nn ont des fréquences p^2 , q^2 et $2*p*q$, respectivement.

$$p^2 + q^2 + 2pq = 1$$

Ici, p = 0.7 et q = 0.3 :

- p^2 = la fréquence du génotype homozygote NN = $0,7^2 = 0,49$
- q^2 = la fréquence du génotype homozygote nn = $0,3^2 = 0,09$
- $2pq$ = la fréquence des génotypes hétérozygotes Nn et nN = $2*0,3*0,7 = 0,42$

H_0 : les fréquences observées suivent l'équilibre de Hardy-Weinberg : il y a indépendance des allèles.

H_1 : les fréquences observées ne suivent pas l'équilibre de Hardy-Weinberg : les allèles ne sont pas indépendants.

Nous pouvons réaliser les tables des effectifs observés et attendus selon l'équilibre de Hardy-Weinberg de la manière suivante :

Effectifs observés	
NN (blanc)	502
Nn (bleu)	418
nn (noir)	80
Total	1000

Effectifs attendus	
NN (blanc)	$0,49 * 1000 = 490$
Nn (bleu)	$0,42 * 1000 = 420$
nn (noir)	$0,09 * 1000 = 90$
Total	1000

Ainsi,

$$\chi^2 = \frac{(502 - 490)^2}{490} + \frac{(418 - 420)^2}{420} + \frac{(80 - 90)^2}{90} = 1,4145$$

Or, $\chi^2_{\text{seuil}} = 5,991$ pour $\alpha = 5\%$ et 2 ddl.

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , les fréquences respectent l'équilibre de Hardy-Weinberg ; il y a bien indépendance allélique.

21. On estime la proportion d'animaux atteints d'un parasite donné à 20%. Dans une ferme, 5 animaux sont testés, et aucun n'est parasité. Pourquoi le test du χ^2 n'est-il pas adapté dans cette situation pour vérifier si la ferme est plus épargnée que la moyenne de la population? Quelle solution alternative proposez-vous ? Que peut-on conclure de cette expérience ?

H_0 : la proportion d'animaux parasités dans la ferme est identique à celle de la population.

H_1 : la proportion d'animaux parasités dans la ferme est différente de celle de la population.

Le test de chi-carré n'est pas adapté car l'effectif est trop faible.

Dans cette situation, la distribution est discrète, avec 2 modalités (parasitisme ou absence de parasitisme) et un nombre de répétitions faibles (nombre d'animaux testés). Nous pouvons donc appliquer la loi binomiale.

$$P(r) = B(r|n, p) = C_n^r * p^r * q^{n-r}$$

Avec :

- n le nombre d'animaux testés = 5
- r le nombre d'animaux parasités = 0
- p la probabilité d'animaux parasités (ou prévalence) = 0,2
- q = 1-p = la probabilité d'animaux non parasités = 0,8

Ainsi,

$$P(r = 0) = B(0|5,0.2) = C_5^0 * p^0 * q^{5-0} = \frac{5!}{0!5!} * 0,2^0 * 0,8^5 = 0,32768$$

$P(r=0) > 0,05$ donc on accepte H_0 : le résultat de notre expérience est compatible avec une fréquence de 20% de parasités.

22. Un zootechnicien a réalisé une expérience de pesée sur un ensemble de 250 bovins d'âges similaires. Il a réparti ses observations par classes de poids. Ses observations se répartissent comme suit:

Classes de poids	Fréquence observées
< 180	2
180- 200	13
200- 220	26
220- 240	51
240- 260	71
260- 280	60
280- 300	23
300- 320	3
Total	249

Il aimerait s'assurer que la distribution de poids est gaussienne. Expliquez la démarche.

H_0 : la distribution des poids des bovins suit une distribution gaussienne

H_1 : la distribution des poids des bovins ne suit pas une distribution gaussienne

Afin d'effectuer un test de χ^2 , nous avons besoin des fréquences attendues. Néanmoins, sans logiciel, nous devons passer par la loi normale centrée réduite afin de se référer à la table des Z par la suite.

$$Z_i = \frac{X_i - \bar{X}_l}{S}$$

Il est donc nécessaire de calculer la moyenne et l'écart-type pour données groupées puisque chaque classe comprend un certain effectif.

Moyenne μ pour données groupées :

$$\bar{X}_i = \frac{\sum f_i * X_i}{\sum f_i}$$

avec :

- X_i les classes : 180 pour la première classe, valeur centrale pour les autres classes
- f_i les fréquences correspondant à chaque classe.

$$\mu = \frac{180 * 2 + 190 * 13 + 210 * 26 + 230 * 51 + 250 * 71 + 270 * 60 + 290 * 23 + 310 * 3}{249}$$

$$\bar{X} = \frac{61570}{249} = 247,269$$

Variance S^2 pour données groupées :

$$S^2 = \frac{\sum f_i * X_i^2 - \frac{(\sum f_i * X_i)^2}{\sum f_i}}{\sum f_i - 1}$$

$$S^2 = \frac{15412700 - \frac{61570^2}{249}}{248} = 759,447$$

Or, l'écart-type $S = \sqrt{S^2} = \sqrt{759,447} = 27,558$

Exemples des calculs pour les classes <180 (classe 1) et 180-200 kg (classe 2)

Conversion de la valeur supérieure de la classe X_i en valeur supérieure de la classe Z_i

$$Z_1 = \frac{180 - 247,269}{27,558} = -2,44$$

$$Z_2 = \frac{200 - 247,269}{27,558} = -1,72$$

Probabilités cumulées pour les classes 1 et 2

$$P(Z_1 < -2,44) = P(Z_1 < 0) - P(-2,44 < Z_1 < 0) = 0,5 - 0,4927 = 0,0073$$

$$P(Z_2 < -1,72) = P(Z_2 < 0) - P(-1,72 < Z_2 < 0) = 0,5 - 0,4573 = 0,0427$$

Probabilités théoriques non cumulées pour les classes 1 et 2 :

$P(Z_1 < -2,44) = 0,0073 \rightarrow$ Nous conservons cette valeur comme il s'agit de la première classe

$P(-2,44 < Z_2 < -1,72) = P(Z_2 < -1,72) - P(Z_1 < -2,44) = 0,0427 - 0,0073 = 0,0354$

Fréquences théoriques pour les classes 1 et 2 :

$$\text{Fréquence}_{\text{théo}} = \text{Probabilité}_{\text{théo}} * \text{Effectifs totaux}$$

$$\text{Fréquence}_{\text{théo classe 1}} = 0,0073 * 249 = 1,8177$$

$$\text{Fréquence}_{\text{théo classe 2}} = 0,0354 * 249 = 8,815$$

Sur excel, on obtient les valeurs suivantes pour l'ensemble des calculs :

Classes	Valeurs supérieures Classes Xi	Valeurs supérieures classes Zi	Probabilités théoriques cumulées	Probabilités théoriques	Fréquences théoriques
1	180	-2,44	0,0073	0,0073	1,8309
2	200	-1,72	0,0431	0,0358	8,9565
3	220	-0,99	0,1612	0,1181	29,5142
4	240	-0,26	0,3960	0,2348	58,6928
5	260	0,46	0,6779	0,2820	70,4926
6	280	1,19	0,8825	0,2046	51,1445
7	300	1,91	0,9722	0,0896	22,4072
8	320	2,64	0,9958	0,0237	5,9225

Remarque : les résultats diffèrent des calculs manuels puisqu'excel travaille sur les valeurs exactes.

On obtient donc :

Classes de poids	Fréquence observées	Fréquences théoriques
< 180	2	1,83
180- 200	13	8,96
200- 220	26	29,51
220- 240	51	58,69
240- 260	71	70,49
260- 280	60	51,14
280- 300	23	22,41
300- 320	3	5,92
Total	249	249

Or, on constate que certaines fréquences théoriques sont inférieures à 10, ce qui signifie que nous ne pouvons pas réaliser de χ^2 dans les conditions actuelles. Afin de pallier à cette problématique, il est possible de rassembler les classes de fréquence < 10 avec la classe suivante ou précédente.

Classes de poids	Fréquence observées	Fréquences théoriques
< 200	15	10,79
200- 220	26	29,51
220- 240	51	58,69
240- 260	71	70,49
260- 280	60	51,14
280- 320	26	28,33
Total	249	249

Ainsi,

$$\chi^2 = \frac{(15 - 10,79)^2}{10,79} + \frac{(26 - 29,51)^2}{29,51} + \dots + \frac{(26 - 28,33)^2}{28,33} = 4,8$$

Or, $\chi^2_{\text{seuil}} = 11,07$ pour $\alpha = 5\%$ et 5 ddl.

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , la distribution des poids de bovin suit la loi normale.

23. Un généticien prétend que les individus porteurs d'un génotype particulier au niveau d'un gène qu'il a identifié ont tendance à être protégés contre une maladie. Pour démontrer ses dires, il collectionne un échantillon de 100 personnes porteuses du génotype et 100 personnes porteuses d'autres génotypes. Parmi ces dernières, 30 sont atteintes. Combien peuvent être atteintes dans son échantillon de porteurs pour que l'effet soit démontré (au seuil $\alpha = 5\%$) ? Comment s'appelle ce type d'expérience ? Quelles proportions n'ont pas de sens biologique particulier ?

H_0 : il y a indépendance entre le génotype et le fait d'être atteint : il y a autant de personnes atteintes dans le groupe portant un génotype particulier que dans le groupe portant d'autres génotypes.

H_1 : il y a moins de personnes atteintes dans le groupe portant un génotype particulier que dans le groupe portant d'autres génotypes.

On déduit la table des effectifs observés suivante :

<u>Effectifs observés</u>	Atteintes	Saines	Total
Génotype particulier	A	100-A	100
Autres génotypes	30	70	100
Total	30+A	200-(30+A)	200

Ainsi, pour les effectifs attendus, on calcule :

<u>Effectifs attendus</u>	Atteintes	Saines	Total
Génotype particulier	$\frac{(30+A)*100}{200} = \frac{(30+A)}{2} = 15 + \frac{A}{2}$	$\frac{(170-A)*100}{200} = 85 - \frac{A}{2}$	100
Autre génotype	$\frac{(30+A)*100}{200} = \frac{(30+A)}{2} = 15 + \frac{A}{2}$	$\frac{(170-A)*100}{200} = 85 - \frac{A}{2}$	100
Total	30+A	170-A	200

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(A - (15 + \frac{A}{2}))^2}{15 + \frac{A}{2}} + \frac{((100 - A) - (85 - \frac{A}{2}))^2}{85 - \frac{A}{2}} + \frac{(30 - (15 + \frac{A}{2}))^2}{15 + \frac{A}{2}} + \frac{(70 - (85 - \frac{A}{2}))^2}{85 - \frac{A}{2}}$$

$$\chi^2 = \frac{(\frac{A}{2} - 15)^2}{15 + \frac{A}{2}} + \frac{(15 - \frac{A}{2})^2}{85 - \frac{A}{2}} + \frac{(15 - \frac{A}{2})^2}{15 + \frac{A}{2}} + \frac{(\frac{A}{2} - 15)^2}{85 - \frac{A}{2}}$$

Considérant que :

$$\left(15 - \frac{A}{2}\right)^2 = \left(\frac{A}{2} - 15\right)^2$$

On obtient :

$$\chi^2 = 2 * \frac{\left(15 - \frac{A}{2}\right)^2}{85 - \frac{A}{2}} + 2 * \frac{\left(15 - \frac{A}{2}\right)^2}{15 + \frac{A}{2}}$$

On veut que $\chi^2_{\text{observé}} > \chi^2_{\text{seuil}}$

Donc $\chi^2_{\text{observé}} > 3,84$ pour $\alpha = 5\%$ et 1 ddl

$$\chi^2_{\text{observé}} > 3,84 \Leftrightarrow 2 * \frac{\left(15 - \frac{A}{2}\right)^2}{85 - \frac{A}{2}} + 2 * \frac{\left(15 - \frac{A}{2}\right)^2}{15 + \frac{A}{2}} > 3,84$$

$$3,84 * \left(15 + \frac{A}{2}\right) * \left(85 - \frac{A}{2}\right) < 2 * \left(15 - \frac{A}{2}\right)^2 * \left(85 - \frac{A}{2}\right) + 2 * \left(15 - \frac{A}{2}\right)^2 * \left(15 + \frac{A}{2}\right)$$

$$3,84 * \left(1275 - 15 * \frac{A}{2} + 85 * \frac{A}{2} - \frac{A^2}{4} \right) < 2 * \left(15 - \frac{A}{2} \right)^2 * \left(85 - \frac{A}{2} + 15 + \frac{A}{2} \right)$$

$$4896 + 134,4 * A - 0,96A^2 < 100 * 2 * \left(15 - \frac{A}{2} \right)^2$$

$$4896 + 134,4 * A - 0,96A^2 < 200 * \left(225 + \frac{A^2}{4} - 15A \right)$$

$$4896 + 134,4 * A - 0,96A^2 < 45000 + 50A^2 - 3000 A$$

$$50,96A^2 - 3134,4 A + 40104 > 0$$

Résolution d'une équation du second degré

$$\delta = b^2 - 4ac = (3134,4)^2 - 4 * 50,96 * 40104 = 1649664$$

$$A = \frac{-b \pm \sqrt{\delta}}{2a} = \frac{-3134,4 \pm \sqrt{1649664}}{2 * 50,96} = 43,3555 \text{ et } 18,1516$$

Comme l'hypothèse alternative est qu'il y a moins de personnes atteintes dans le groupe portant un génotype particulier que dans le groupe portant d'autres génotypes, on doit choisir la valeur de A la plus petite des deux. Le nombre d'atteintes parmi les génotypes particulier doit donc être de 18 ou moins.

Si l'on veut appliquer le test du χ^2 (comme vérification), on travaillera sur les tables suivantes :

<u>Effectifs observés</u>	Atteintes	Saines	Total
Génotype particulier	18	82	100
Autre génotype	30	70	100
Total	48	152	200

<u>Effectifs attendus</u>	Atteintes	Saines	Total
Génotype particulier	24	76	100
Autre génotype	24	76	100
Total	48	152	

$$\chi^2 = \frac{(18 - 24)^2}{24} + \frac{(82 - 76)^2}{76} + \frac{(30 - 24)^2}{24} + \frac{(70 - 76)^2}{76} = 3,95$$

Ainsi $\chi^2_{\text{observé}}$ est bien supérieur à χ^2_{seuil} pour $A = 18$ (et valeurs inférieures) : on rejette H_0 au seuil 5%.

Il s'agit d'une étude prospective (dite également « de cohorte ») puisqu'on collecte des individus sur base d'un facteur (génotype présent ou absent) en constituant deux cohortes de tailles arbitraires 100 et 100. Puis on regarde "l'apparition" de la maladie dans ces 2 cohortes.

Remarque : il est indiqué dans le correctif que l'étude est rétrospective. On pourrait en effet la considérer rétrospective si la maladie était apparue avant la formation des groupes, auquel cas on aurait observé cet élément dans le passé des individus collectés.

Si l'on considère que l'étude est prospective, la dernière question (concernant les mesures d'association) ne s'applique plus. Il n'est en revanche pas possible de calculer la fréquence du génotype favorable dans la population avec un échantillon de type « cohortes ». Néanmoins, si l'on considère l'étude rétrospective, pour laquelle on choisit de manière arbitraire le nombre de cas et de contrôles, les mesures RR, O+ et O- dépendent de ce choix arbitraire (changer les proportions relatives de cas et de contrôles modifie ces mesures). Calculer les RR, O+ et O- n'a donc pas de sens pour ce type d'étude. Par contre, OR est insensible aux changements de tailles des groupes de cas et de contrôles, ce qui en fait une mesure intéressante.

24. Une firme pharmaceutique souhaite commercialiser un médicament utile pour traiter l'arthrite chez les chiens âgés. Pour cela, elle envoie à 100 vétérinaires deux colis à un mois d'intervalle. Un des colis contient effectivement le médicament, et l'autre un placebo (évidemment, les destinataires ne sont pas prévenus qu'une des substances servira de témoin). La firme demande ensuite aux vétérinaires quelle substance ils jugent plus efficace. 16 d'entre eux les ont trouvées efficaces toutes les deux. 38 les ont jugées inopérantes. Finalement, 27 ont préféré le traitement 1 (correspondant au médicament), et les autres ont eu l'impression que le traitement 2 (correspondant au placebo) avait un effet plus favorable. Comment analysez-vous ces données ?

Concluez sur l'efficacité du traitement proposé.

Les données sont ici paires puisqu'il s'agit du même échantillon de vétérinaires qui teste les deux produits. Par conséquent, nous devons réaliser un test de McNemar pour comparer les effectifs attendus aux effectifs observés.

La table des effectifs observés est la suivante :

		Placebo		
		Efficace	Non efficace	
Médicament	Efficace	16	27	43
	Non efficace	19	38	57
	Total	35	65	100

Nos hypothèses seront les suivantes :

H0 : il n'y a pas de différence entre le médicament et le placebo.

H1 : il y a une différence entre le médicament et le placebo.

Ce qui nous intéresse ici sont donc les situations pour lesquelles des changements ont été observés:

Les changements de non efficace (médicament) vers efficace (placebo) ou d'efficace (placebo) vers non efficace (médicament) ; a= 19

Les changements d'efficace (médicament) vers non efficace (placebo) (ou l'inverse) ; b= 27

$$\chi^2 = \frac{(a - b)^2}{a + b} = \frac{(19 - 27)^2}{19 + 27} = 1,39$$

$\chi^2_{\text{seuil}} = 3,841$ pour 1 ddl et un seuil de 5% donc $\chi^2_{\text{obs}} < \chi^2_{\text{seuil}}$: on accepte H0, il n'y a pas de différence d'efficacité démontrée dans cette expérience entre le médicament et le placebo.

25. Un petit test a été réalisé pour tester l'efficacité d'un vaccin: 5 individus ont été vaccinés et 5 autres pas. Les résultats du test sont dans la table suivante :

	Vaccinés	Non vaccinés	Total
Malades	1	3	4
Sains	4	2	6
Total	5	5	10

De quel genre d'analyse s'agit-il ? Pourquoi n'est-on pas intéressé par toutes les conclusions que l'on pourrait tirer de cette analyse ?

Comment tester l'efficacité du vaccin ?

Il s'agit d'une étude prospective ou « de cohorte » puisque l'on constitue 2 groupes auxquels on applique deux traitements différents (puis on regarde lesquels tombent malades et lesquels sont sains au sein de chaque groupe).

Les mesures globales sont biaisées car la petite taille de l'échantillon participe à une puissance de test très (trop) faible pour démontrer un effet dans la plupart des cas. Remarque : Dans les études prospectives, l'exposition au facteur n'est pas représentatif de la population : le fait de vacciner 5 personnes et de ne pas vacciner 5 autres ne signifie pas que 50% de la population est vaccinée et 50% ne l'est pas. Par contre, le taux de « protection » dans chaque cohorte est supposé représentatif des taux correspondants dans les sous-populations de vaccinés et non-vaccinés.

Les hypothèses sont les suivantes :

H_0 : il n'y a pas de différence du nombre de malades entre les vaccinés et les non vaccinés : le vaccin n'est pas efficace.

H_1 : il y a moins de malades parmi les vaccinés par rapport aux non vaccinés : le vaccin est efficace.

On constate que les effectifs observés sont faibles avec un effectif total < 25 (et des effectifs attendus qui seront inférieurs à 10). Il n'est donc pas possible de faire un χ^2 pour répondre à l'hypothèse.

Comme il existe 2 sous-populations contenant 2 catégories, on peut appliquer un test exact de Fisher selon la loi hypergéométrique.

Ainsi,

$$P(a_1; N_1; a; N) = \frac{C_{N_1}^{a_1} * C_{N_2}^{a_2}}{C_N^a} \text{ avec } a_1 + a_2 = a$$

Avec :

- N_1 le nombre de vaccinés = 5
- a_1 le nombre de malades parmi les vaccinés = 1
- N_2 le nombre de non vaccinés = 5
- a_2 le nombre de malades parmi les non vaccinés = 3
- N correspond donc aux effectifs totaux = 10
- a correspond donc à $a_1 + a_2 = 4$

$$P(1; 5; 4; 10) = \frac{C_5^1 * C_5^3}{C_{10}^4} = \frac{\frac{5!}{1! * 4!} * \frac{5!}{3! * 2!}}{\frac{10!}{4! * 6!}} = 0,2381$$

$P(1; 5; 4; 10) > 0,05$ donc on sait que l'on peut déjà accepter H_0 : le vaccin n'est pas efficace.

En principe, la valeur p est la somme des probabilités de la table observée et des tables plus extrêmes. On doit donc calculer ces tables plus extrêmes si l'on veut connaître la valeur p exacte. Si on souhaite juste savoir si la valeur p est > 0.05 , le fait que la probabilité de la table observée soit > 0.05 suffit déjà pour accepter H_0 à ce seuil.

Sachant que $a_1=1$, la seule table extrême aurait été pour une valeur de $a_1 = 0$.

On peut par exemple demander à R d'effectuer le calcul :

```
> dhyper(0, 4, 6, 5)
[1] 0.02380952
> |
```

$P(0; 5; 4; 10) = 0,0238$ donc la probabilité cumulée pour $a_1=1$ correspond à $P(1;5;4;10) + P(0; 5; 4;10) = 0,2381 + 0,02381 = 0,2619$.

Remarque : nous aurions pu aussi demander directement sur R la probabilité cumulée correspondant à $a_1=1$.

```
> phyper(1,4,6,5)
[1] 0.2619048
```

26. Générez une table de contingence 3x2 et testez-y l'indépendance. Commentez.

H_0 : les données sont indépendantes.

H_1 : les données sont dépendantes.

Si l'on génère la table d'effectifs observés suivante, pour laquelle on calcule les totaux marginaux :

<u>Effectifs</u> <u>« observés »</u>		Facteur X			Total
		X1	X2	X3	
Facteur Y	Y1	86	63	97	246
	Y2	85	57	81	223
Total		171	120	178	469

On peut en déduire la table des effectifs attendus :

		Facteur X			Total
		X1	X2	X3	
Facteur Y	Y1	89,69	62,94	93,36	246
	Y2	81,31	57,06	84,64	223
Total		171	120	178	469

Ainsi,

$$\chi^2 = \frac{(86 - 89,69)^2}{89,69} + \frac{(63 - 62,94)^2}{62,94} + \dots + \frac{(57 - 57,06)^2}{57,06} + \frac{(81 - 84,64)^2}{84,64} = 0,618$$

$\chi^2_{\text{seuil}} = 35,991$ pour 2 ddl et un seuil de 5% donc $\chi^2_{\text{obs}} < \chi^2_{\text{seuil}}$: on accepte H_0 , les données sont indépendantes.

27. On a développé une race de poule « Supercoq » dont la coquille des œufs serait plus solide et on a comparé cette solidité avec celle d'une race traditionnelle :

	Solidité +	Solidité -	Total
Poules supercoq	209	42	251
Poules traditionnelles	54	95	149
Total	263	137	400

a) La race « Supercoq » possède-t-elle des œufs avec des coquilles plus solides ?

b) Calculez l'Odds Ratio et interprétez-le.

c) Pouvez-vous interpréter le rapport 209/263 ?

a)

H0 : il n'y a pas d'effet de la race de poule sur la solidité de l'œuf.

H1 : il y a un effet de la race de poule sur la solidité de l'œuf.

La table des effectifs attendus est :

	Solidité +	Solidité -	Total
Poules supercoq	165,03	85,97	251
Poules traditionnelles	97,97	51,03	149
Total	263	137	400

$$\chi^2 = \frac{(209 - 165,03)^2}{165,03} + \frac{(42 - 85,97)^2}{85,97} + \frac{(54 - 97,97)^2}{97,97} + \frac{(95 - 51,03)^2}{51,03} = 91,81$$

$\chi^2_{\text{seuil}} = 3,841$ pour 1 ddl et un seuil de 5% donc $\chi^2_{\text{obs}} \gg \chi^2_{\text{seuil}}$: on rejette H0, il y a bien un effet de la race sur la solidité des coquilles.

b) On peut calculer les Odds, qui vont ici comparer la probabilité d'avoir un œuf plus solide si la race est supercoq (O+) ou traditionnelle (O-). Un œuf plus solide est noté S+ et un œuf moins solide est noté S-.

$O+ = P(S+|\text{supercoq})/P(S-|\text{supercoq}) = 209/42 = 4,976 \rightarrow$ cela signifie que l'œuf a 4,976 plus de chances d'être plus solide s'il est issu de la race supercoq.

$O- = P(S+|\text{traditionnelle})/P(S-|\text{traditionnelle}) = 54/95 = 0,568$

$$OR = \frac{O +}{O -} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a * d}{b * c}$$

Si l'on remplace par nos données :

$$OR = \frac{O +}{O -} = \frac{P(S + | \text{supercoq}) / P(S - | \text{supercoq})}{P(S + | \text{traditionnelle}) / P(S - | \text{traditionnelle})} = \frac{\frac{209}{42}}{\frac{54}{95}} = \frac{209 * 95}{42 * 54} = 8,75$$

Les « odds » associés à la présence du facteur (race supercoq) sont 8,75 fois plus élevés que ceux associés à l'absence de facteur (race traditionnelle). La race supercoq a un effet positif sur la solidité des œufs dans cet échantillon, et cette association est très significative, comme montré plus haut.

Remarque : une valeur d'OR élevée indique une association forte dans l'échantillon, mais ne veut pas nécessairement dire que cette association est réelle (lire « significative »). Pour déterminer si elle est réelle, il est nécessaire de réaliser le test statistique adéquat. C'est ce que nous avons réalisé ici et nous avons prouvé que cette association était bien significative.

Petite astuce visuelle pour le calcul de l'OR : les produits de l'OR correspondent aux « croisements » des valeurs dans un tableau de contingence 2*2 :

	Solidité +	Solidité -	
Poules supercoq	209	42	251
Poules traditionnelles	54	95	149
	263	137	400

$$OR = \frac{209 * 95}{42 * 54} = 8,75$$

A noter que nous aurions pu considérer le problème sous l'angle inverse, c'est-à-dire inverser la manière dont on considère le facteur, ce qui revient à inverser la valeur de l'OR, comme montré ci-dessous. Encore une fois, le test de χ^2 utilisé ci-dessus testait le fait qu'il n'y avait pas d'association, ce qui se traduirait par un OR = 1. Quelle que soit la manière dont on a calculé l'OR, les valeurs obtenues s'écartent de manière très significative de 1.

$$O+ = P(S+|traditionnelle)/P(S-|traditionnelle) = 42/209 = 0,201$$

$$O- = P(S+|supercoq)/P(S-|supercoq) = 95/54 = 1,759$$

$$OR = \frac{O +}{O -} = \frac{P(S + | \text{traditionnelle}) / P(S - | \text{traditionnelle})}{P(S + | \text{supercoq}) / P(S - | \text{supercoq})} = \frac{\frac{54}{95}}{\frac{209}{42}} = \frac{54 * 42}{95 * 209} = 0,114$$

c) La proportion $\frac{209}{263} = 0,79$ correspond à la probabilité d'être un supercoq si l'œuf est plus solide.

De la même manière, la proportion $\frac{54}{263} = 0,205$ correspond à la probabilité d'être une poule traditionnelle si l'œuf est plus solide (et ainsi de suite).

Ceci dit, il y a fort à parier que les tailles respectives des 2 cohortes de poules ne représentent pas une population particulière. Et donc ces probabilités ne sont valables que pour l'échantillon et ne représente pas une réalité plus large.

28. Quelle est la distance entre les moyennes de deux distributions normales de même variance 10^2 sachant que l'hypothèse nulle d'égalité des moyennes est rejetée 95 fois sur 100 au seuil $\alpha = 5\%$ bilatéral ?

Les données sont normales. Nous allons donc utiliser la distribution de Z.

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

Si nous nous intéressons à la première distribution de moyenne μ_1 et considérée comme correspondant à H_0 :

Puisque nous réalisons un test bilatéral pour $\alpha = 5\%$, nous cherchons les valeurs Z telles que $P(Z_{H0} < Z_1) = P(Z_{H0} > Z_2) = 0,025$.

Pour cela, nous avons besoin de déterminer $P(Z_1 < Z_{H0} < 0) = P(0 < Z_{H0} < Z_2) = 0,5 - 0,025 = 0,475$

Dans la table, nous obtenons $Z_{H0} = -1,96$ ou $Z_{H0} = 1,96$.

De ce fait, si nous considérons $Z_{H0} = 1,96$:

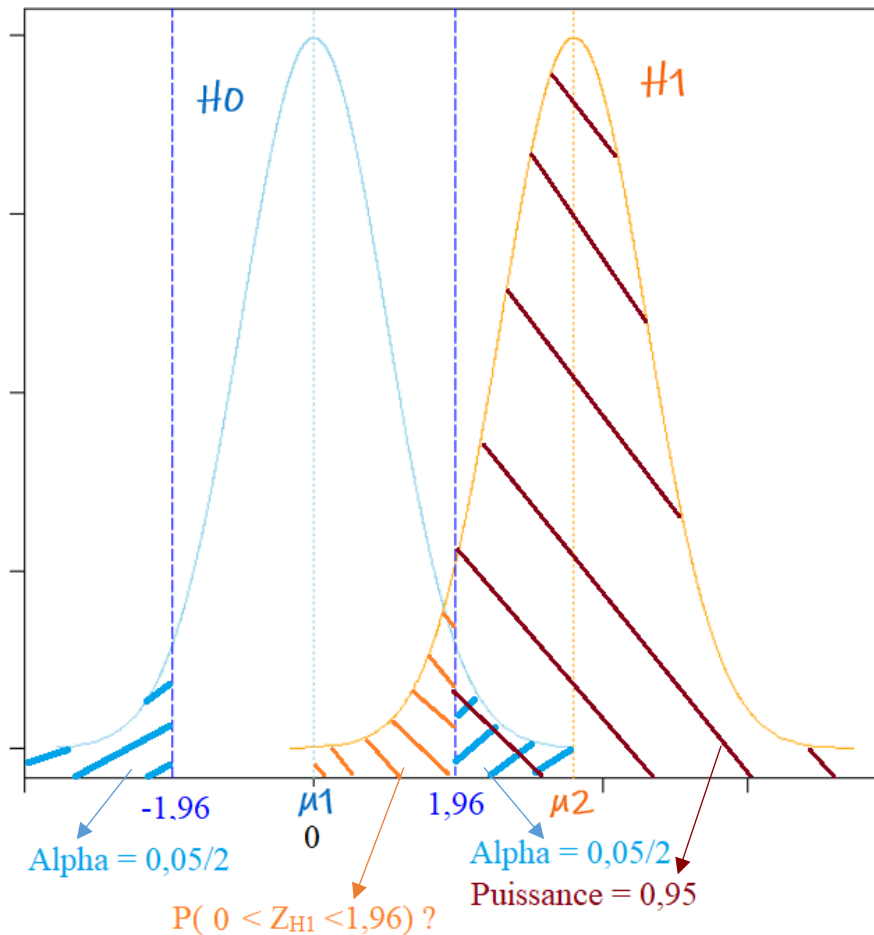
$$Z_{H0} = \frac{X - \mu_1}{\sigma} \Leftrightarrow 1,96 = \frac{X - \mu_1}{\sqrt{10^2}} \Leftrightarrow 1,96 * 10 = X - \mu_1$$

$$X = 1,96 * 10 + \mu_1$$

Si nous considérons un $Z_{H0} = 1,96$ et que H_0 est rejetée 95 fois sur 100, on sait que, pour une distribution H_1 , de moyenne μ_2 :

$P(Z_{H1} > 1,96) = 0,950$ (probabilité correspondant à la puissance).

Nous pouvons représenter les éléments mentionnés graphiquement :



Comme $P(Z_{H1} > 1,96)$ est supérieure à 0,5, nous savons déjà que Z_{H1} sera négatif (si l'on considère que la distribution a pour moyenne $\mu_2 = 0$).

Pour obtenir Z_{H1} , on calcule $P(0 < Z_{H1} < 1,96) = P(Z_{H1} > 1,96) - P(Z_{H1} > 0) = 0,950 - 0,5 = 0,450$. Si on lit la table à l'envers (et en se souvenant que la valeur sera négative), on obtient $Z_{H1} = -1,645$.

Ainsi,

$$Z_{H1} = \frac{X - \mu_2}{\sigma} \Leftrightarrow -1,645 = \frac{X - \mu_2}{\sqrt{10^2}} \Leftrightarrow -1,645 * 10 = X - \mu_2$$

$$X = -1,645 * 10 + \mu_2$$

Par conséquent,

$$X = -1,645 * 10 + \mu_2 = 1,96 * 10 + \mu_1$$

$$\mu_2 - \mu_1 = 1,96 * 10 + 1,645 * 10$$

$$\mu_2 - \mu_1 = 36$$

La distance entre μ_1 et μ_2 est de 36.

29. Une équipe scientifique a étudié les taux sanguins de cholestérol chez des chiennes hypothyroïdiennes. L'étude a révélé que le taux de cholestérol était normalement distribué avec une moyenne de 208 mg/dl et une déviation standard de 25 mg/dl. Une étude précédente avait montré que des taux inférieurs à 200 mg/dl pouvaient quasiment exclure une hypothyroïdie.

a) Quelle est la probabilité d'observer une chienne hypothyroïdienne avec un cholestérol en dessous de 200 mg/dl ?

b) Pour trois chiennes hypothyroïdiennes prises au hasard, quelle est la probabilité qu'au moins une des chiennes ait un taux de cholestérol sanguin inférieur à 200 mg/dl ?

a) La question est $P(X < 200) = ?$

Comme nous ne pouvons pas faire les calculs de probabilité sans logiciel avec les données telles quelles, nous allons passer par la loi normale centrée réduite.

$$Z = \frac{X - \mu}{\sigma} = \frac{200 - 208}{25} = -0,32$$

Dans la table Z, $P(-0,32 < Z < 0) = 0,1255$

Donc $P(Z < -0,32) = P(Z < 0) - P(-0,32 < Z < 0) = 0,5 - 0,1255 = 0,3745$

La probabilité d'observer une chienne hypothyroïdienne avec un cholestérol en dessous de 200 mg/dL est donc $P(X < 200) = 0,3745$.

b) Nous pouvons appliquer la loi binomiale puisque nous avons 2 modalités : cholestérol > 200 mg/dL ou cholestérol < 200 mg/dL.

$$P(r) = B(r|n, p) = C_n^r * p^r * q^{n-r}$$

Avec

- $n = 3$
- $p = P(X < 200) = 0,3745$
- $q = 1 - P(X < 200) = 0,6255$

Ce qui est demandé est $P(r \geq 1)$, ce qui peut se traduire par $1 - P(r=0)$ puisque nous travaillons sur une distribution discrète.

$$\begin{aligned} P(0) &= B(0|3, 0,3745) = C_3^0 * 0,3745^0 * 0,6255^{3-0} = \frac{3!}{0! 3!} * 0,3745^0 * 0,6255^3 \\ &= 0,2447 \end{aligned}$$

Et $P(r \geq 1) = 1 - 0,2447 = 0,7553$

30. Un éleveur de chiens, dont l'élevage est infecté par l'herpès virose canine, souhaite savoir si la vaccination de ses nouvelles chiennes a été efficace. Pour cela, il regarde la mortalité des chiots dans chacune des portées. Sur 11 portées en 2 ans, 6 ont été observées chez les nouvelles chiennes vaccinées et 1 a perdu sa portée d'herpès virose. Chez les « anciennes » chiennes de son élevage, pas toutes infectées, il a perdu 3 portées d'herpès virose. Quelle est la probabilité de cette situation si on suppose que la vaccination n'a pas eu d'effet (H0)? L'hypothèse nulle est-elle acceptable ?

H0 : il n'y a pas de différence de mortalité des chiots entre les mères vaccinées et non vaccinées : la vaccination n'a pas d'effet.

H1 : il y a une différence de mortalité des chiots entre les mères vaccinées et non vaccinées : la vaccination a un effet.

Selon l'énoncé, les effectifs observés sont les suivants :

<u>Effectifs observés</u>	Vaccinées	Non vaccinées	Total
Mortalité	1	3	4
Survie	5	2	7
Total	6	5	11

Ces effectifs sont faibles (effectifs totaux < 25 et effectifs attendus < 5), nous ne pouvons donc pas appliquer de test de χ^2 .

Nous pouvons en revanche faire un test exact de Fisher en appliquant la loi hypergéométrique.

La loi hypergéométrique nous indique que, dans un ensemble de taille N, constitué de deux sous-ensembles de tailles N_1 et $N_2 = N - N_1$, la probabilité de prélever a_1 ($< N_1$) individus dans l'ensemble de taille N_1 quand on prélève un total de a ($< N$) individus vaut :

$$P(a_1; N_1; a; N) = \frac{C_{N_1}^{a_1} * C_{N_2}^{a_2}}{C_N^a} \text{ avec } a_1 + a_2 = a$$

$$\text{Et } C_N^a = \frac{N!}{a!(N-a)!}$$

Ici, on définit par exemple :

- N_1 le nombre de femelles vaccinées = 6
- a_1 le nombre de portées ayant survécu chez les femelles vaccinées = 5
- N_2 le nombre de femelles non vaccinées = 5
- a_2 le nombre de portées ayant survécu chez les femelles non vaccinées = 2
- N correspond donc aux effectifs totaux = 11
- a correspond donc à $a_1 + a_2 = 7$

Ainsi, on a :

$$P(5; 6; 7; 11) = \frac{C_6^5 * C_5^2}{C_{11}^7} = \frac{\frac{6!}{5! * 1!} * \frac{5!}{2! * 3!}}{\frac{11!}{7! * 4!}} = 0,1818$$

Remarque : dans cette table de contingence de 2*2, nous pouvons calculer cette probabilité simple de 4 manières différentes en fonction du choix des valeurs correspondant à a1, a2, N1 et N2 (et donc de la façon dont on se pose la question).

Ainsi, on peut faire $P(5; 6; 7; 11) = P(1; 4; 6; 11) = P(1; 6; 4; 11) = P(3; 4; 5; 11)$

$$= \frac{C_6^5 * C_5^2}{C_{11}^7} = \frac{C_4^1 * C_7^5}{C_{11}^6} = \frac{C_6^1 * C_5^3}{C_{11}^4} = \frac{C_4^3 * C_7^2}{C_{11}^5} = 0,1818$$

Nous constatons que la probabilité simple pour $P(5; 6; 7; 11)$ -et équivalents selon le choix des valeurs- est déjà > 5%, ce qui rend inutile le calcul des probabilités pour les tables extrêmes (afin d'obtenir la probabilité cumulée).

Si l'on souhaite malgré tout faire le calcul, la seule table extrême dans le sens qui nous intéresse est :

	Vaccinées	Non vaccinées	Total
Mortalité	0	4	4
Survie	6	1	7
Total	6	5	11

Ce qui nous permet de calculer $P(6; 6; 7; 11) = 0,01515$

Ainsi, la probabilité cumulée est égale à $P(6; 6; 7; 11) + P(5; 6; 7; 11)$ si l'on considère la survie des portées = $0,1818 + 0,01515 = 0,197$

31. Une étude scientifique analysant l'apparition d'une fièvre de lait après le vêlage dans l'espèce bovine a permis de constater que sur un total de 450 vaches, 89 avaient présenté une fièvre de lait lors de leurs deux premières lactations, 152 n'avaient jamais présenté de fièvre de lait et 113 ont présenté une fièvre de lait uniquement lors de la première lactation. Énoncez l'hypothèse nulle et testez si le numéro de lactation a un impact sur l'incidence d'apparition de la fièvre de lait. Interprétez vos résultats.

H0 : le numéro de lactation n'a pas d'effet sur l'apparition de la fièvre de lait.

H1 : le numéro de lactation a un effet sur l'apparition de la fièvre de lait.

Les effectifs observés peuvent être représentés de la manière suivante :

		2ème lactation		
		Fièvre	Absence fièvre	Total
1ère lactation	Fièvre	89	113	202
	Absence fièvre	96	152	248
Total		185	265	450

Les données sont paires puisqu'il s'agit d'analyser l'apparition de la fièvre de lait sur les mêmes vaches lors de 2 lactations. Nous devons donc réaliser un test de McNemar.

Ainsi,

$$\chi^2 = \left(a - \frac{a+b}{2}\right)^2 / \frac{a+b}{2} + \left(b - \frac{a+b}{2}\right)^2 / \frac{a+b}{2}$$

$$\Rightarrow \chi^2 = \frac{(a-b)^2}{a+b}$$

Avec:

- a : transition de fièvre en première lactation → Absence de fièvre en deuxième lactation = 113
- b : transition d'absence de fièvre en première lactation → fièvre en deuxième lactation = 96

$$\chi^2 = \frac{(113 - 96)^2}{113 + 96} = 1,383$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl (puisque (lignes - 1) * (colonnes - 1) = (2-1)*(2-1) = 1)

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H0, il n'y a pas d'effet du numéro de lactation sur l'apparition des fièvres de lait.

32. Chez les Yorkshire Terriers, la mise-bas pose souvent des problèmes en raison de la conformation de la race. Dans une étude effectuée sur 160 chiennes, on a essayé de voir si le fait de présenter une dystocie (accouchement difficile) lors de la première portée était susceptible de se représenter lors de la seconde ou non. Les résultats sont donnés dans la table suivante:

		Nichée 2		Total
		Dystocie	Pas de dystocie	
Nichée 1	Dystocie	21	31	52
	Pas de dystocie	37	71	108
	Total	58	102	160

- a) Quelle hypothèse (nulle) veut-on tester dans ce test ? Quelle procédure utiliser pour mettre en évidence l'éventuelle relation ?
- b) Si l'hypothèse nulle est vraie, combien de chiennes, non dystociques en première portée, attend-on devenir dystociques en deuxième portée ?
- c) Combien vaut la statistique utilisée pour tester cet effet ?
- d) Est-ce que le test est significatif (càd. permet de rejeter H_0) ?

a) H_0 : le numéro de la portée n'a pas d'effet sur la présence de dystocie.

H_1 : le numéro de la portée a un d'effet sur la présence de dystocie.

Les données sont paires (mêmes chiennes sur 2 portées). Nous devons donc réaliser un test de McNemar.

b) On relève :

- a : transition de dystocie en nichée 1 $\rightarrow$ absence de dystocie en nichée 2 = 31
- b : transition d'absence de dystocie en nichée 1 $\rightarrow$ dystocie en nichée 2 = 37

Sous H_0 , on s'attend à ce que $a = b$ donc on réalise la moyenne de ces 2 valeurs pour connaître la valeur attendue :

$$\text{Valeur attendue sous } H_0 = \frac{a + b}{2} = \frac{31 + 37}{2} = 34$$

c) d)

$$\chi^2 = \frac{(31 - 37)^2}{31 + 37} = 0,5294$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl (puisque $(\text{lignes} - 1) * (\text{colonnes} - 1) = (2-1)*(2-1) = 1$)

Ainsi $\chi^2_{\text{observé}} < \chi^2_{\text{seuil}}$: on accepte H_0 , il n'y a pas d'effet du numéro de nichée sur la présence de dystocie.

On pourrait calculer la p-valeur avec un logiciel, par exemple R :

```
> pchisq(0.5294,1)
[1] 0.5331408
```

Cette probabilité est $> 5\%$, ce qui confirme que l'on accepte H_0 .

33. Un référendum est effectué sur 500 personnes. Déterminez de deux manières la majorité significative au seuil 5%.

- H_0 : il y a autant de pour que de contre
- H_1 : il y a un nombre différent de pour et de contre

Il est possible de calculer la probabilité en utilisant :

- 1) Un test de Z sur une proportion
- 2) Un test de Z avec l'approximation normale de la loi binomiale
- 3) Un test de χ^2
- 4) Un test exact de Fisher (avec logiciel)

1) Test de Z sur proportion (bilatéral)

Puisque $p=0,5$ sous H_0 et $n=500$, nous pouvons appliquer la formule de comparaison du pourcentage obtenu $\hat{p}$ (que nous cherchons ici à déterminer) par rapport à un pourcentage prévu p .

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p * q}{n}}} = \frac{\hat{p} - 0,5}{\sqrt{\frac{0,5 * (1 - 0,5)}{500}}}$$

Or, puisque nous réalisons un test bilatéral pour $\alpha = 5\%$, nous cherchons les valeurs Z_1 et Z_2 telles que $P(Z < Z_1) = P(Z > Z_2) = 0,025$.

Pour cela, nous avons besoin de déterminer $P(Z_1 < Z < 0) = P(0 < Z < Z_2) = 0,5 - 0,025 = 0,475$

Dans la table, nous obtenons $Z_1 = -1,96$ ou $Z_2 = 1,96$.

Ainsi, si l'on considère $Z_1 = 1,96$:

$$1,96 = \frac{\hat{p} - 0,5}{\sqrt{\frac{0,5 * (1 - 0,5)}{500}}} \Leftrightarrow 1,96 * \sqrt{\frac{0,5 * 0,5}{500}} = \hat{p} - 0,5$$

$$\hat{p} = 0,043827 + 0,5 = 0,543827$$

Si l'on rapporte cette proportion au nombre de votants, on obtient :

$$\hat{p} * n = 0,543827 * 500 = 271,91$$

Il faut donc au moins 272 voix « pour » ou « contre » pour atteindre une majorité significative au seuil 5% dans un sens ou dans l'autre (pour un test bilatéral).

2) Test de Z avec l'approximation normale de la loi binomiale (bilatéral)

$$Z = \frac{r - n * p}{\sqrt{npq}} \Leftrightarrow 1,96 = \frac{r - 500 * 0,5}{\sqrt{500 * 0,5 * 0,5}} \Leftrightarrow 1,96 * \sqrt{500 * 0,5 * 0,5} = r - 500 * 0,5$$

$$r = 21,9135 + 500 * 0,5 = 271,91$$

3) Test de χ^2

Effectifs observés

Pour	n
Contre	500-n
Total	500

Effectifs attendus

Pour	250
Contre	250
Total	500

Ainsi,

$$\chi^2 = \frac{(n - 250)^2}{250} + \frac{((500 - n) - 250)^2}{250} = \frac{(n - 250)^2}{250} + \frac{(-n + 250)^2}{250} = 2 * \frac{(n - 250)^2}{250}$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl

Ainsi,

$$\chi^2_{\text{seuil}} = 2 * \frac{(n - 250)^2}{250} \Leftrightarrow 3,841 = 2 * \frac{(n - 250)^2}{250} \Leftrightarrow 3,841 * \frac{250}{2} = (n - 250)^2$$

$$480,125 = (n - 250)^2 \Leftrightarrow \sqrt{480,125} = n - 250 \Leftrightarrow n = 21,9118 + 250$$

$$n = 271,91$$

Il faut au moins 272 voix « pour » ou « contre » pour atteindre une majorité significative au seuil 5% dans un sens ou dans l'autre (le test du χ^2 est toujours bilatéral).

4) Test exact

On aurait aussi pu faire un calcul exact selon la loi binomiale à l'aide d'un logiciel :

`=LOI.BINOMIALE.INVERSE(500;0,5;0,975)`

	D	E	F
	Loi binomiale		
	272		

34. On souhaite tester l'efficacité d'une vaccination. Pour cela, 300 chiens sont choisis dont 2/3 ont été vaccinés et on constate que 18 chiens vaccinés sont tombés malades contre 32 chez les non vaccinés après exposition au virus. La vaccination est-elle efficace ? Calculez l'Odds Ratio et interprétez-le. Sachant que sur les 50 chiens tombés malades, 18 étaient vaccinés, pouvez-vous interpréter la probabilité de 18/50 (36%) ?

H_0 : Il n'y a pas de différence dans l'occurrence de la maladie entre les chiennes vaccinées et non vaccinées : il n'y a pas d'effet du vaccin.

H_1 : Il y a une différence dans l'occurrence de la maladie entre les chiennes vaccinées et non vaccinées : le vaccin a un effet.

Suivant l'énoncé, on obtient les tableaux d'effectifs observés et attendus suivants :

<u>Effectifs observés</u>	Vaccin	Pas de vaccin	Total
Malade	18	32	50
Sain	182	68	250
Total	200	100	300

<u>Effectifs théo</u>	Vaccin	Pas de vaccin	Total
Malade	33,333	16,667	50
Sain	166,667	83,333	250
Total	200	100	300

$$\chi^2 = \frac{(18 - 33,333)^2}{33,333} + \frac{(32 - 16,667)^2}{16,667} + \frac{(182 - 166,667)^2}{166,667} + \frac{(68 - 83,333)^2}{83,333}$$

$$= 25,392$$

Or, $\chi^2_{\text{seuil}} = 3,841$ pour $\alpha = 5\%$ et 1 ddl (puisque (lignes - 1) * (colonnes - 1) = (2-1)*(2-1) = 1)

Ainsi $\chi^2_{\text{observé}} \gg \chi^2_{\text{seuil}}$: on rejette H_0 , la vaccination est efficace.

Calcul de l'Odds ratio

On peut commencer par calculer les Odds, qui vont ici comparer la probabilité d'être malade si l'on est vacciné (O+) ou non (O-).

On notera M+ pour les malades, M- pour les sains, V+ pour les vaccinés et V- pour les non vaccinés.

$O+ = P(M+ | V+)/P(M- | V+) = 18/182 = 0,0989 \rightarrow$ Cela signifie que la probabilité d'être malade est à peu près 10 fois moindre que celle de ne pas être malade si on est vacciné. Ce qui ne veut pas nécessairement dire que le vaccin est efficace : on pourrait avoir le même rapport parmi les non vaccinés ! D'où la nécessité de comparer ces deux rapports (odds) en formant un « odds-ratio ».

$O- = P(M+ | V-)/P(M- | V-) = 32/68 = 0,4706 \rightarrow$ Cela signifie que la probabilité d'être malade ne vaut que 47% de la probabilité de ne pas être malade si on est pas vacciné.

Nous allons donc calculer l'Odds ratio :

$$OR = \frac{O+}{O-} = \frac{\frac{a}{c}}{\frac{b}{d}} = \frac{a * d}{b * c}$$

Si l'on remplace par nos données :

$$OR = \frac{O+}{O-} = \frac{P(M+ | V+)/P(M- | V+)}{P(M+ | V-)/P(M- | V-)} = \frac{\frac{18}{182}}{\frac{32}{68}} = \frac{18 * 68}{182 * 32} = 0,2102$$

Il y a bien une association entre la vaccination et la maladie. Les « odds » associés à la présence du facteur (vaccination) sont 0,21 fois plus faibles que ceux associés à l'absence de facteur (pas de vaccination). La vaccination a un effet protecteur contre la maladie.

c) La proportion $\frac{18}{50} = 0,36$ correspond à la probabilité que le chien soit vacciné s'il est malade dans notre expérience, mais ne correspond à la même probabilité dans la population, la taille des cohortes « vaccinés » et « non-vaccinés » étant arbitraire, et donc non représentative de la population.

35. On a développé une race de poule « Supercoq » dont la coquille des œufs serait plus solide et on a comparé cette solidité avec celle d'une race traditionnelle :

	Solidité +	Solidité -	Total
Poules supercoq	7	2	9
Poules traditionnelles	2	5	7
Total	9	7	16

a) Quelle est la probabilité de cette situation ?

b) L'hypothèse nulle voulant une égalité de « solidité + » entre les deux races est-elle vraisemblable ? Interprétez la zone de rejet.

H_0 : il n'y a pas de différence de solidité des œufs entre les poules supercoq et traditionnelles.

H_1 : il y a une différence de solidité des œufs entre les poules supercoq et traditionnelles.

Les effectifs totaux et les effectifs attendus sont trop faibles pour réaliser un test de χ^2 . Nous devons donc réaliser un test exact de Fisher en utilisant la loi hypergéométrique.

La loi hypergéométrique nous indique que, dans un ensemble de taille N , constitué de deux sous-ensembles de tailles N_1 et $N_2 = N - N_1$, la probabilité de prélever a_1 ($< N_1$) individus dans l'ensemble de taille N_1 quand on prélève un total de a ($< N$) individus vaut :

$$P(a_1; N_1; a; N) = \frac{C_{N_1}^{a_1} * C_{N_2}^{a_2}}{C_N^a} \text{ avec } a_1 + a_2 = a$$

Et $C_N^a = \frac{N!}{a!(N-a)!}$

Ici, on définit par exemple :

- N_1 les poules supercoq = 9
- a_1 les œufs plus solides parmi les poules supercoq = 7
- N_2 les poules traditionnelles = 7
- a_2 les œufs plus solides parmi les poules traditionnelles = 2
- N correspond donc aux effectifs totaux = 16
- a correspond donc à $a_1 + a_2 = 9$

Ainsi, on pourrait calculer :

$$P(7; 9; 9; 16) = \frac{C_9^7 * C_7^2}{C_{16}^9} = \frac{\frac{9!}{7! * 2!} * \frac{7!}{2! * 5!}}{\frac{16!}{9! * 7!}} = 0,0661$$

Il est également nécessaire de calculer les probabilités pour les tables extrêmes dans le sens d'intérêt c'est-à-dire dans la situation où il y a plus d'œufs plus solides dans la race supercoq. Nous cherchons donc à calculer $P(a_1 = 8)$ et $P(a_1 = 9)$.

	Solidité +	Solidité -	Total
Poules supercoq	8	1	9
Poules traditionnelles	1	6	7
Total	9	7	16

$$P(8; 9; 9; 16) = \frac{C_9^8 * C_7^1}{C_{16}^9} = \frac{\frac{9!}{8! * 1!} * \frac{7!}{1! * 6!}}{\frac{16!}{9! * 7!}} = 0,00551$$

	Solidité +	Solidité -	Total
Poules supercoq	9	0	9
Poules traditionnelles	0	7	7
Total	9	7	16

$$P(9; 9; 9; 16) = \frac{C_9^9 * C_7^0}{C_{16}^9} = \frac{\frac{9!}{9! * 0!} * \frac{7!}{0! * 7!}}{\frac{16!}{9! * 7!}} = 0,00008741$$

Ainsi, $P(a_1 > 6) = P(a_1 = 9) + P(a_1 = 8) + P(a_1 = 7) = 0,00008741 + 0,00551 + 0,0661$

$$= 0,0716$$

$P(a_1 > 6) > 5\%$ donc on accepte H_0 : la race de poule n'a pas d'effet sur la solidité de la coquille.

Néanmoins, la probabilité calculée est relativement proche du seuil des 5%.

b) La zone (unilatérale) où il y a rejet au seuil 5% est donc celle où $a_1 > 7$: il faut qu'au moins 8 des 9 œufs solides proviennent de poules Supercoq.

$$P(a_1 > 7) = P(a_1 = 9) + P(a_1 = 8) = 0,00008741 + 0,00551 = 0,005597$$