Analyse des données « des Données brutes aux Connaissances exploitables »

Animée par: **FANGACHI Oussama**

Plan

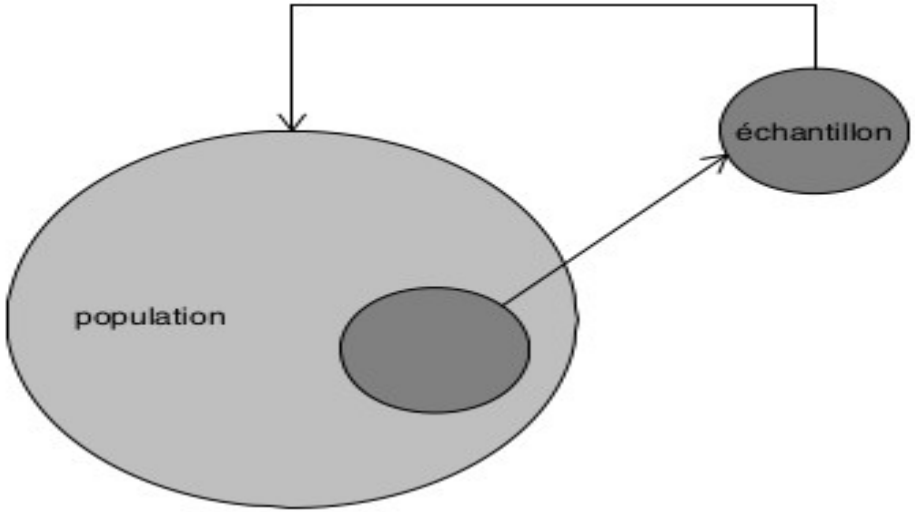
- Terminologie
- Principaux types de variables
- Distribution d’une variable
- Mesures de tendance centrale
- Boxplot et quantiles
- Mesures de variabilité

Plan

- Variable standardisée (z-scores)
- Distribution normale
- Mesures de non-normalité
- Estimation ponctuelle
- Estimation par intervalle de confiance
- Principe d’un test statistique
- Conclusion

Population , variable , échantillon

- **la population : est l'ensemble des individus d'intérêt d'une étude, que ce soient des patients, des plantes, des insectes ou différents lancers d'une pièce de monnaie**
- **une variable : est une caractéristique d'intérêt mesurable sur les individus de la population**
- **un échantillon : est un ensemble de quelques individus représentatifs de la population pour lesquels une variable est effectivement mesurée**


échantillon
population

Principaux types de variables

- Variable qualitative
  - Nominale
  - ordinale
- Variable quantitative
  - Continue
  - Discret

Distribution d’une variable

- **Distribution d’une variable qualitative**

O	O	O	A	A	A	A	A	O	O	O	O	O	A	A	O
O	O	O	O	O	A	O	O	O	B	O	O	O	O	O	A
A	A	A	O	O	B	O	O	O	O	O	O	A	O	A	O
A	B	O	A	B	O	A	A	A	A	AB	O	A	O	O	O
A	AB	O	AB	O	O	B	A	A	A	O	A	O	O	B	O
O	O	O	A	O	A	A	A	O	O	A	O	A	A	A	O
O	AB	A	A	O	A	O	A	A	A	O	O	A	O	A	O
O	A	O	A	A	O	O	A	A	O	A	O	A	B	A	O
O	A	O	B	O	O	O	A	O	O	A	O	A	A	O	O
O	A	O	A	A	B	A	O	O	O	B	A	A	A		

Barplot

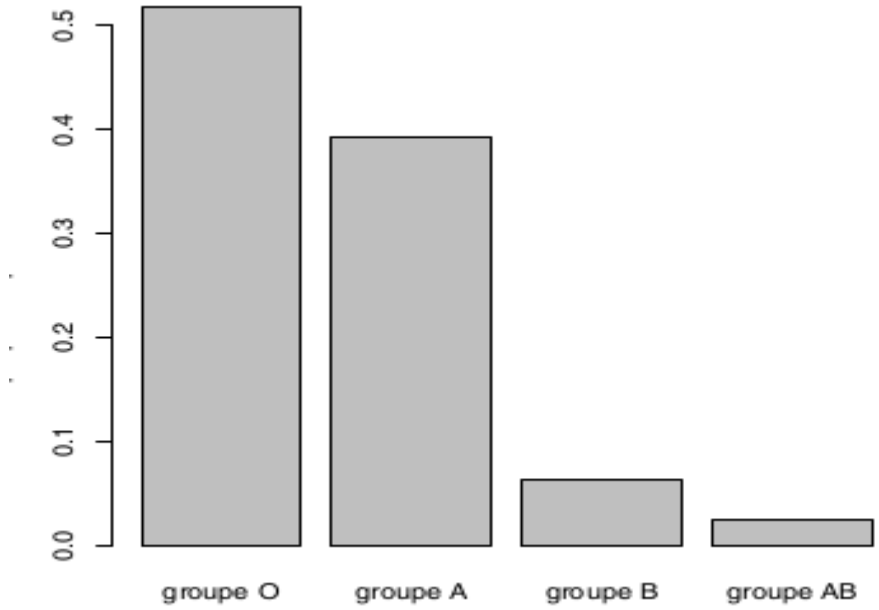


Table de contigence

groupe O	groupe A	groupe B	groupe AB	total
82	62	10	4	158
52 %	39 %	6 %	3 %	100 %

Distribution d’une variable continue

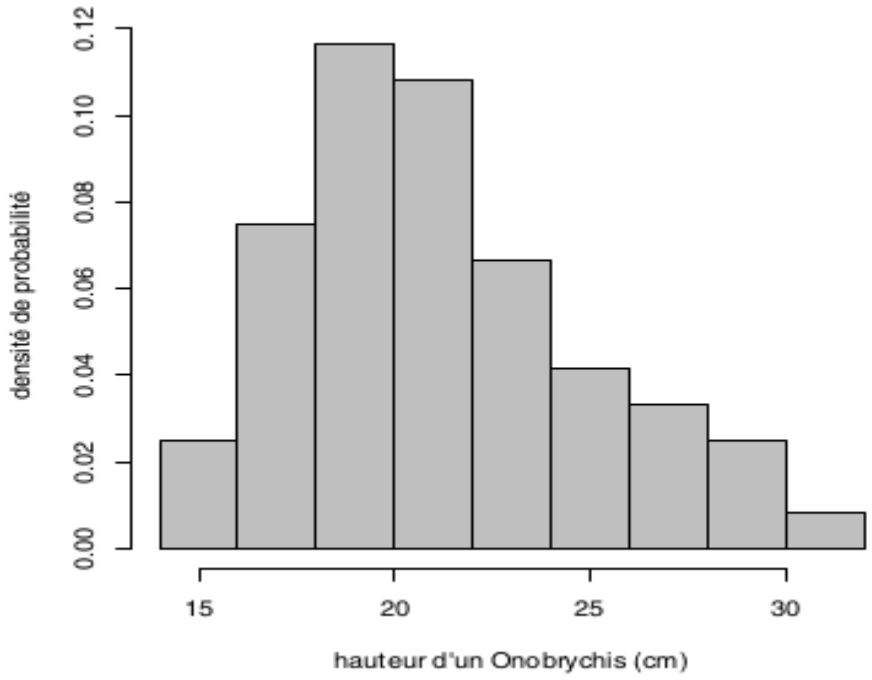
- **Afin de résumer graphiquement la distribution d’une variable continue, on peut calculer un histogramme. Il s’agit de compartimenter l’ensemble des valeurs possibles en un certain nombre d’intervalles de même longueur et de compter le nombre d’observations dans chaque intervalle**

Exemple

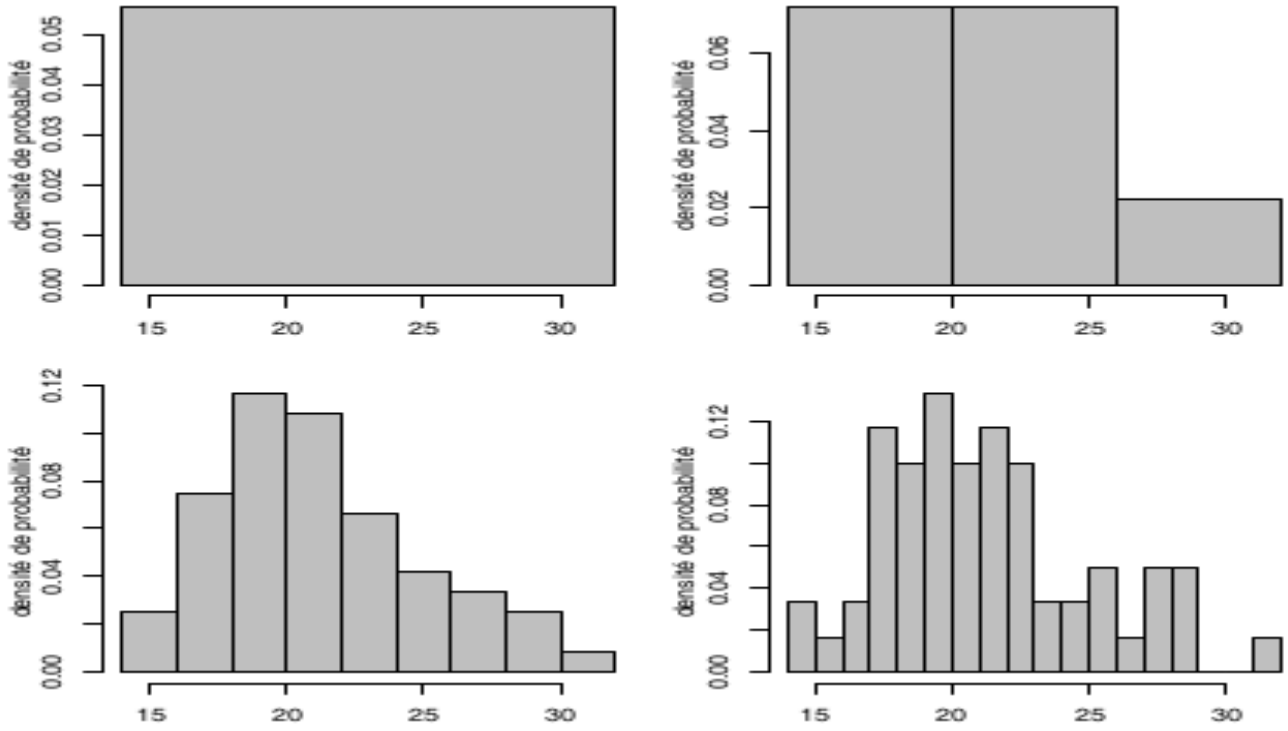
- on a mesuré la hauteur (en cm) de n = 60 spécimens d'Onobrychis viciifolia (une plante herbacée vivace)

21	21	23	22	23	29	24	21	18	23	19	18	20	24	20
20	19	19	22	21	18	20	23	17	20	25	23	21	14	18
29	28	28	14	28	26	22	22	22	29	19	26	16	17	23
18	25	22	20	22	18	32	26	21	20	27	20	19	19	18

- on considère par exemple des intervalles de longueur
- 2cm: 14–16 cm, 16–18 cm, etc., jusqu'à 30–32 cm


densité de probabilité
0.00
0.02
0.04
0.06
0.08
0.10
0.12
15
20
25
30
hauteur d'un Onobrychis (cm)

Inconvénient de cette méthode



Mesures de tendance centrale

- Une mesure de tendance centrale est une caractéristique sur la position du centre (milieu) des données
- La moyenne :
- Moyenne(Y) = $\frac{1}{n}\sum_i y_i$
- la moyenne est la somme des observations divisée par le nombre des observations

- **La médiane :**

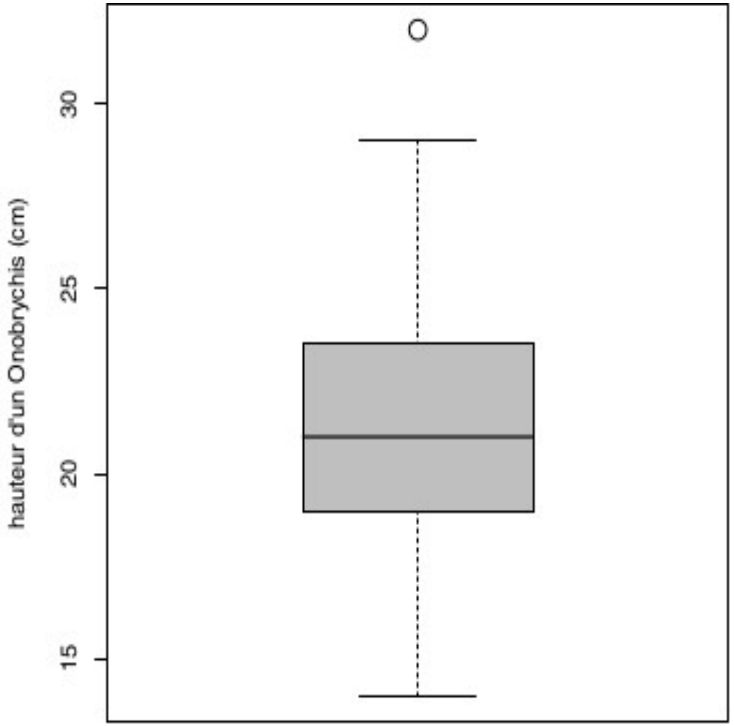
- **median(Y)=** $\begin{cases} y_{((n+1)/2)} & \text{si } n \text{ impair} \\ \frac{1}{2}(y_{(n/2)} + y_{(n/2+1)}) & \text{si } n \text{ pair} \end{cases}$

- **pour calculer la médiane, il s'agit au préalable d'ordonner les observations (y(i) dénotant la i-ième observation ordonnée)**

- **la médiane se trouve au milieu des données, dans le sens où 50 % des observations sont plus grandes et 50 % plus petites que la médiane**

Boxplot et quantiles

- **Le boxplot est parfois appelé résumé à cinq valeurs. En effet, les cinq caractéristiques numériques suivantes sont représentées dans un boxplot**
  - **Le minimum**
  - **Le quantile 25 %**
  - **Le quantile 50 %**
  - **Le quantile 75 %**
  - **Le maximum**


hauteur d'un Onobrychis (cm)
15
20
25
30

Mesures de variabilité

- La moyenne ou la médiane ne résument qu'un aspect de la distribution d'une variable quantitative, nous informant sur la position du centre des données. Un autre aspect important, on y revient, est la variabilité

La variance

- **La variance** $variance(Y) = mean((Y - mean(Y))^2) = \frac{1}{n}\sum_i (y_i - \bar{y})^2$

- **L'ecart type :** $stdev(Y) = \sqrt{variance(Y)} = \sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2}$

Variable standardisée

- **Lorsque l’on standardise une variable, en lui soustrayant d’abord sa moyenne (c’est-à-dire en la centrant) puis en la divisant par son écart type, les unités que l’on obtient sont des nombres d’écarts types par rapport à la moyenne**

Exemple

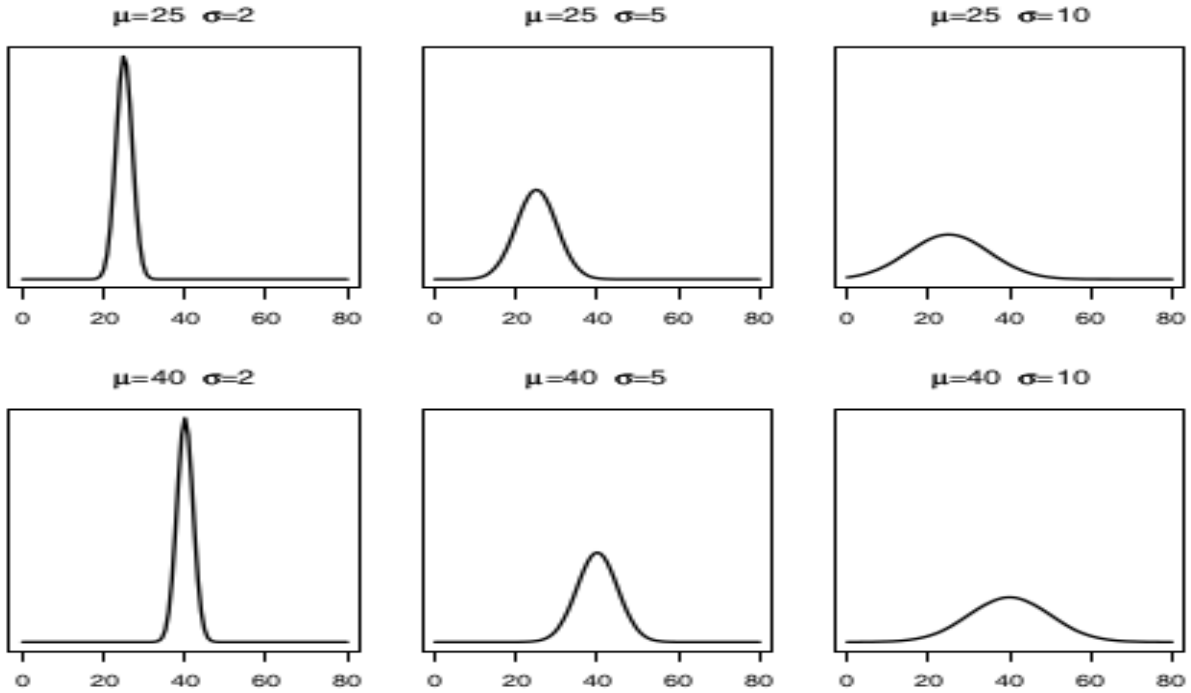
21	21	23	22	23	29	24	21	18	23	19	18	20	24	20
20	19	19	22	21	18	20	23	17	20	25	23	21	14	18
29	28	28	14	28	26	22	22	22	29	19	26	16	17	23
18	25	22	20	22	18	32	26	21	20	27	20	19	19	18

−2.03	−2.03	−1.50	−1.24	−1.24	−0.97	−0.97	−0.97	−0.97	−0.97
−0.97	−0.97	−0.71	−0.71	−0.71	−0.71	−0.71	−0.71	−0.45	−0.45
−0.45	−0.45	−0.45	−0.45	−0.45	−0.45	−0.18	−0.18	−0.18	−0.18
−0.18	−0.18	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.34
0.34	0.34	0.34	0.34	0.34	0.61	0.61	0.87	0.87	1.13
1.13	1.13	1.39	1.66	1.66	1.66	1.92	1.92	1.92	2.71

Distribution normale

- On parle d’une distribution normale si les données sont symétrique et si les données ne contiennent pas des valeurs extrêmes
- Il s’agit d’une famille de distributions indexée par deux paramètres : la moyenne, que l’on notera ici μ, et l’écart type, que l’on notera σ

Exemple



38.3 %	des observations dans l'intervalle	$\mu \pm 0.5 \cdot \sigma$
50.0 %	des observations dans l'intervalle	$\mu \pm 0.675 \cdot \sigma$
68.3 %	des observations dans l'intervalle	$\mu \pm 1 \cdot \sigma$
86.6 %	des observations dans l'intervalle	$\mu \pm 1.5 \cdot \sigma$
90.0 %	des observations dans l'intervalle	$\mu \pm 1.645 \cdot \sigma$
95.0 %	des observations dans l'intervalle	$\mu \pm 1.96 \cdot \sigma$
95.4 %	des observations dans l'intervalle	$\mu \pm 2 \cdot \sigma$
99.0 %	des observations dans l'intervalle	$\mu \pm 2.58 \cdot \sigma$
99.7 %	des observations dans l'intervalle	$\mu \pm 3 \cdot \sigma$
99.9 %	des observations dans l'intervalle	$\mu \pm 3.29 \cdot \sigma$
99.99 %	des observations dans l'intervalle	$\mu \pm 4 \cdot \sigma$

Mesures de non-normalité

- On rappelle que deux caractéristiques importantes de la normalité sont la symétrie de la distribution et le fait qu'il y a peu de valeurs extrêmes.
- Pour mesurer la symétrie de la distribution on utilise :
  - le coefficient d'asymétrie de fisher (Skewness)
- Pour mesurer l'aplatissement
  - le coefficent d'aplatissement (kurtosis)

$$skewness(Y) = \frac{mean((Y - mean(Y))^3)}{stdev^3(Y)}$$

$$kurtosis(Y) = \frac{mean((Y - mean(Y))^4)}{stdev^4(Y)} - 3$$

- **Si Y est normale , on aura skewness(Y)=0 et kurtosis(y)=0**

Estimation

- En pratique, on calcule les caractéristiques de la population sur les données de notre échantillon. En théorie,on pourrait aussi les calculer sur les données de la population, si seulement elles étaient disponibles. Ainsi, bien que l'on calcule en pratique la moyenne de l'échantillon, on peut concevoir l'existence de la moyenne de la population

Estimation ponctuelle

- **Estimateur exacte**
- **Estimateur biaisé**

Estimation par intervalle de confiance

- Intervalle de confiance pour la moyenne
  - Intervalle de wald
  - intervalle de student
- Intervalle de confiance pour la variance
-

Principe d’un test statistique

- L’hypothèse nulle H0
- L’hypothèse alternative H1
- On démontre h1 en rejetant H0

Erreur de première et du seconde espèce

	probabilité de rejeter H_0	probabilité de ne pas rejeter H_0
H_0 vraie	α	$1-\alpha$
H_0 fausse	$1-\beta$	β

- **α = seuil du test = probabilité de rejeter H0 alors que H0 est vraie = 5 %.**

Concept de valeur p

- la valeur p peut être définie comme « la probabilité que le hasard de l'échantillonnage puisse produire des données aussi éloignées (ou encore plus éloignées) de l'hypothèse nulle que le sont les données de notre échantillon, si l'hypothèse nulle était vraie ».
- valeur p = seuil minimal au-delà duquel on rejette H0
- On rejette H0 au seuil α si $p \leq \alpha$

Statistique de test

- définir une statistique de test T_stat calculable sur un échantillon
- établir mathématiquement la distribution théorique de T_stat sous H0
- calculer la réalisation t_stat de T_stat sur notre échantillon
- comparer t_stat avec la distribution théorique de T_stat sous H0

- **Les deux premières étapes sont des étapes théoriques, fondées sur les mathématiques. Les données entrent en jeu à partir de la troisième étape. La quatrième étape est le calcul de la valeur p, qui mesure à quel point les données sont incompatibles avec l'hypothèse nulle, et qui nous permet de décider si on rejette ou non l'hypothèse nulle**
- **un test statistique sera dit exact si on connaît mathématiquement (et si on utilise effectivement) la distribution de la statistique de test sous H0 , alors qu'il sera dit valide si à défaut de la connaître exactement, on dispose d'une bonne approximation de cette distribution**

Tests du khi-deux pour tables de contingence

couleur yeux	couleur cheveux blond	roux	châtain	brun	noir	total
bleu	326	38	241	110	3	718
	(45 %)	(5 %)	(34 %)	(15 %)	(0 %)	(100 %)
vert	688	116	584	188	4	1580
	(44 %)	(7 %)	(37 %)	(12 %)	(0 %)	(100 %)
brun	343	84	909	412	26	1774
	(19 %)	(5 %)	(51 %)	(23 %)	(1 %)	(100 %)
noir	98	48	403	681	85	1315
	(7 %)	(4 %)	(31 %)	(52 %)	(6 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

Distribution théorique

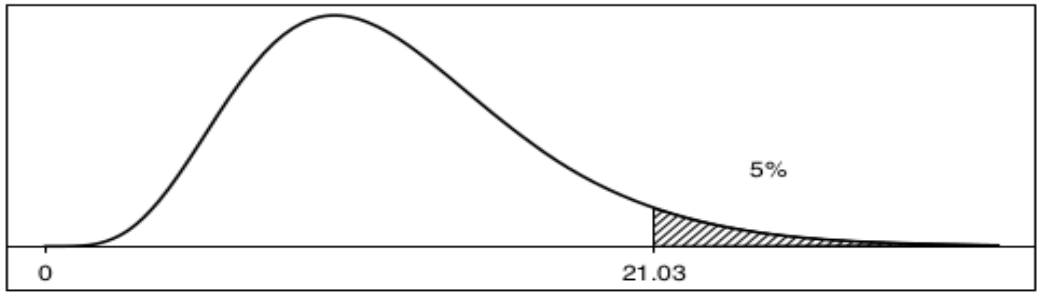
couleur yeux	couleur cheveux					
	blond	roux	châtain	brun	noir	total
bleu	193.9	38.1	284.8	185.4	15.7	718
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
vert	426.7	83.9)	626.8	408.0	34.6	1580
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
brun	479.1	94.2	703.7	458.1	38.9	1774
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
noir	355.2	69.8	521.7	339.6	28.8	1315
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)
total	1455	286	2137	1391	118	5387
	(27 %)	(5 %)	(40 %)	(26 %)	(2 %)	(100 %)

$$
\begin{aligned}
t_{stat} &= \frac{(326-193.9)^2}{193.9} + \frac{(38-38.1)^2}{38.1} + \frac{(241-284.8)^2}{284.8} + \frac{(110-185.4)^2}{185.4} + \frac{(3-15.7)^2}{15.7} \\
&\quad + \frac{(688-426.7)^2}{426.7} + \frac{(116-83.9)^2}{83.9} + \frac{(584-626.8)^2}{626.8} + \frac{(188-408-0)^2}{408.0} + \frac{(4-34.6)^2}{34.6} \\
&\quad + \frac{(343-479.1)^2}{479.1} + \frac{(84-94.2)^2}{94.2} + \frac{(909-703.7)^2}{703.7} + \frac{(412-458.1)^2}{458.1} + \frac{(26-38.9)^2}{38.9} \\
&\quad + \frac{(98-355.2)^2}{355.2} + \frac{(48-69.8)^2}{69.8} + \frac{(403-521.7)^2}{521.7} + \frac{(681-339.6)^2}{339.6} + \frac{(85-28.8)^2}{28.8} \\
&= 1240.0
\end{aligned}
$$

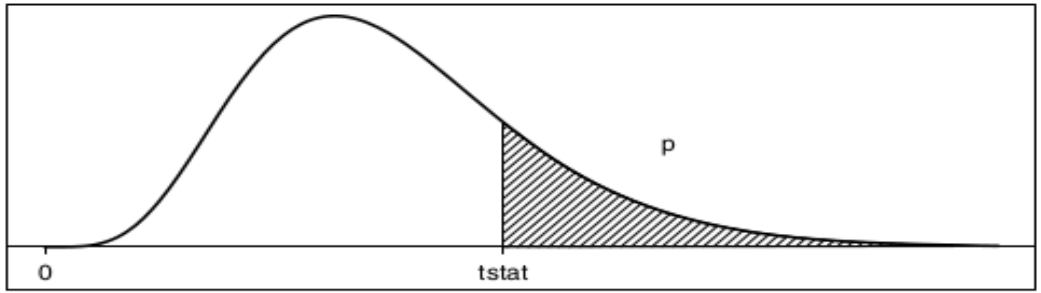
Les étapes

- la première étape consiste à définir une statistique de test T_stat comme on vient de le faire ci-dessus
- la deuxième étape consiste à établir mathématiquement la distribution de cette statistique de test sous l'hypothèse nulle ; il se trouve que la distribution de T_stat sous H0 est ici (approximativement) une distribution du khi-deux avec $(I-1)(J-1)$ dl, l'approximation étant bonne si la majorité (par exemple 80 %) des fréquences attendues sont supérieures à 5, auquel cas le test du khi-deux est dit valide (dans notre exemple, il s'agit donc de $(4-1)(5-1) = 12$ dl)
- la troisième étape consiste à calculer la réalisation t_stat de la variable aléatoire T_stat dans notre échantillon (dans notre exemple t_stat = 1240.0)
- la quatrième étape consiste à comparer la valeur observée (notre exemple t_stat = 1240.0) avec la distribution théorique de T_stat sous H0 (dans notre exemple, une distribution du khi-deux avec 12 dl), l'idée étant de rejeter hypothèse nulle si la statistique de test observée t_stat est incompatible (trop grande) par rapport à la distribution théorique

région de rejet à 5% pour test du khi–deux avec 12 dl



calcul valeur p pour test du khi–deux avec 12 dl



Exemple pratique