

Statistiques avec



M2 Sciences du Langage

Remi.lafitte@univ-grenoble-alpes.fr

2023-2024

Statistiques descriptives et inférentielles

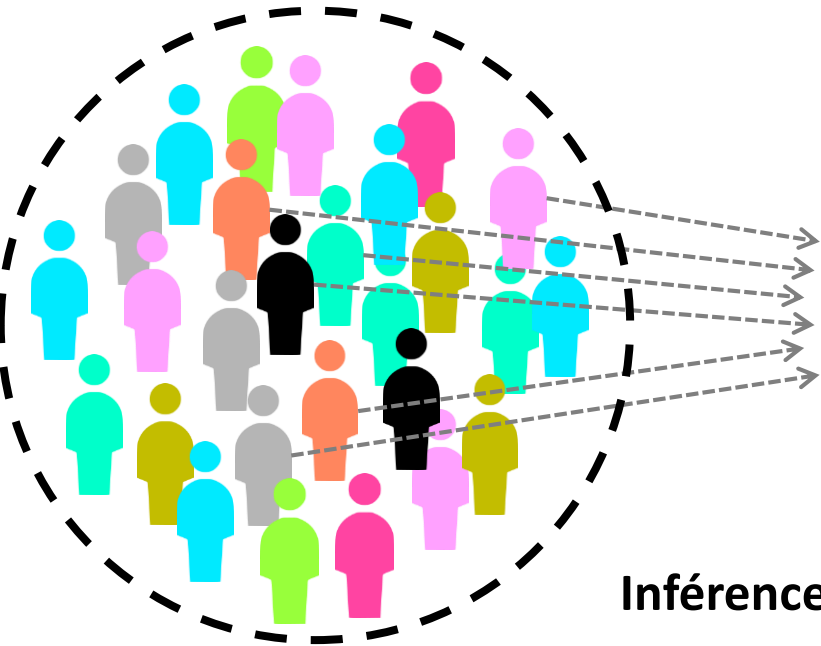


Quelques rappels en matière de statistiques

Les statistiques en bref



Population



Echantillon



Recueil de données

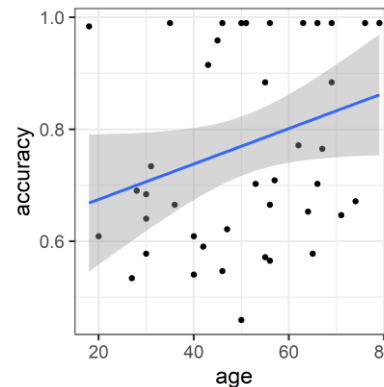
```
> DF
  sujet age accuracy
1     1  69  0.88386
2     2  67  0.76524
3     3  43  0.91508
4     4  18  0.98376
5     5  55  0.88386
6     6  57  0.70905
7     7  62  0.77148
8     8  50  0.99000
9     9  79  0.99000
```

Analyses

```
> cor.test(DF$age,DF$accuracy)

Pearson's product-moment correlation

data:  DF$age and DF$accuracy
t = 1.9074, df = 42, p-value = 0.06332
alternative hypothesis: true correlation is not
95 percent confidence interval:
 -0.01586572  0.53442782
sample estimates:
cor
0.2823446
```

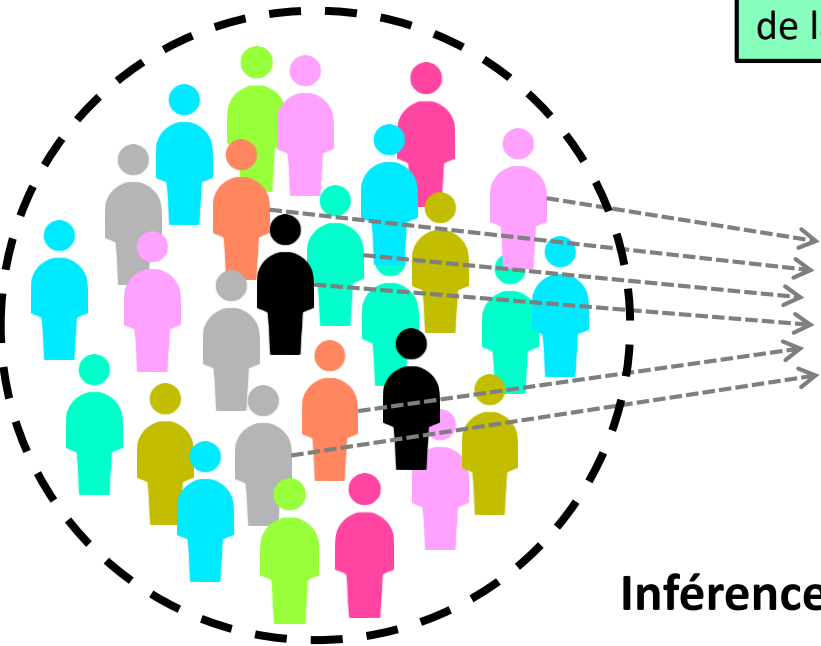


Inférence



Les statistiques en bref

Population



Doit être le plus représentatif de la population

Echantillon



Recueil de données

```
> DF
  sujet age accuracy
```

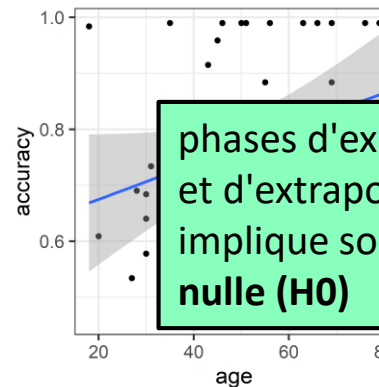
Via questionnaire, logiciel etc... l'outil de mesure doit être fiable

7	7	62	0.77148
8	8	50	0.99000
9	9	79	0.99000

Analyses

Inférence

On rejette H0 mais avec un risque d'erreur



```
> cor.test(DF$age, DF$accuracy)
Pearson's product-moment correlation
```

phases d'exploration (stat **descriptive**) et d'extrapolation (stat **inférentielle**) ; implique souvent de **tester l'hypothèse nulle (H0)**

Variable : nomenclature

Variable indépendante (VI)



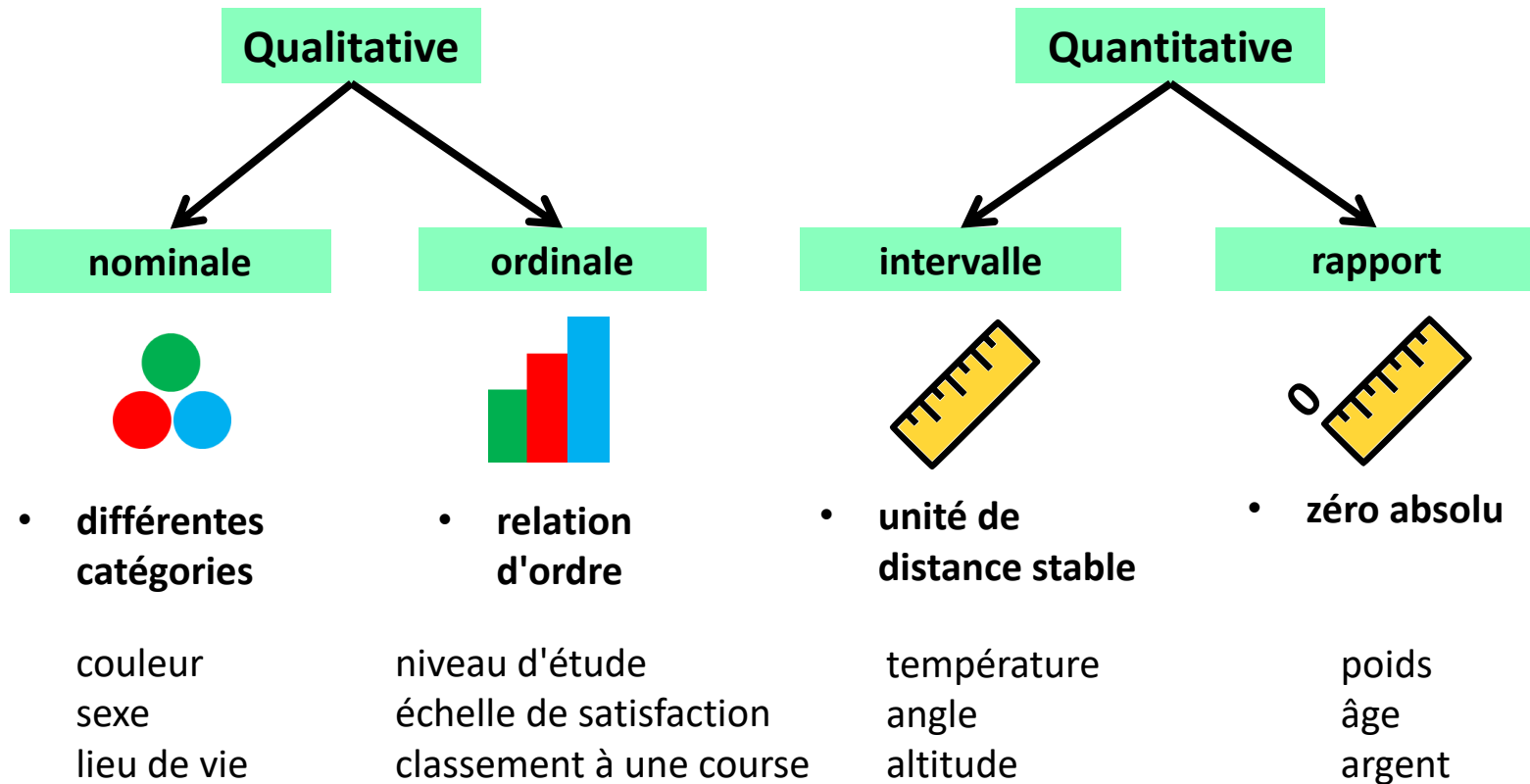
Variable dépendante (VD)

Si VI fluctue, VD fluctue aussi (au moins en partie)



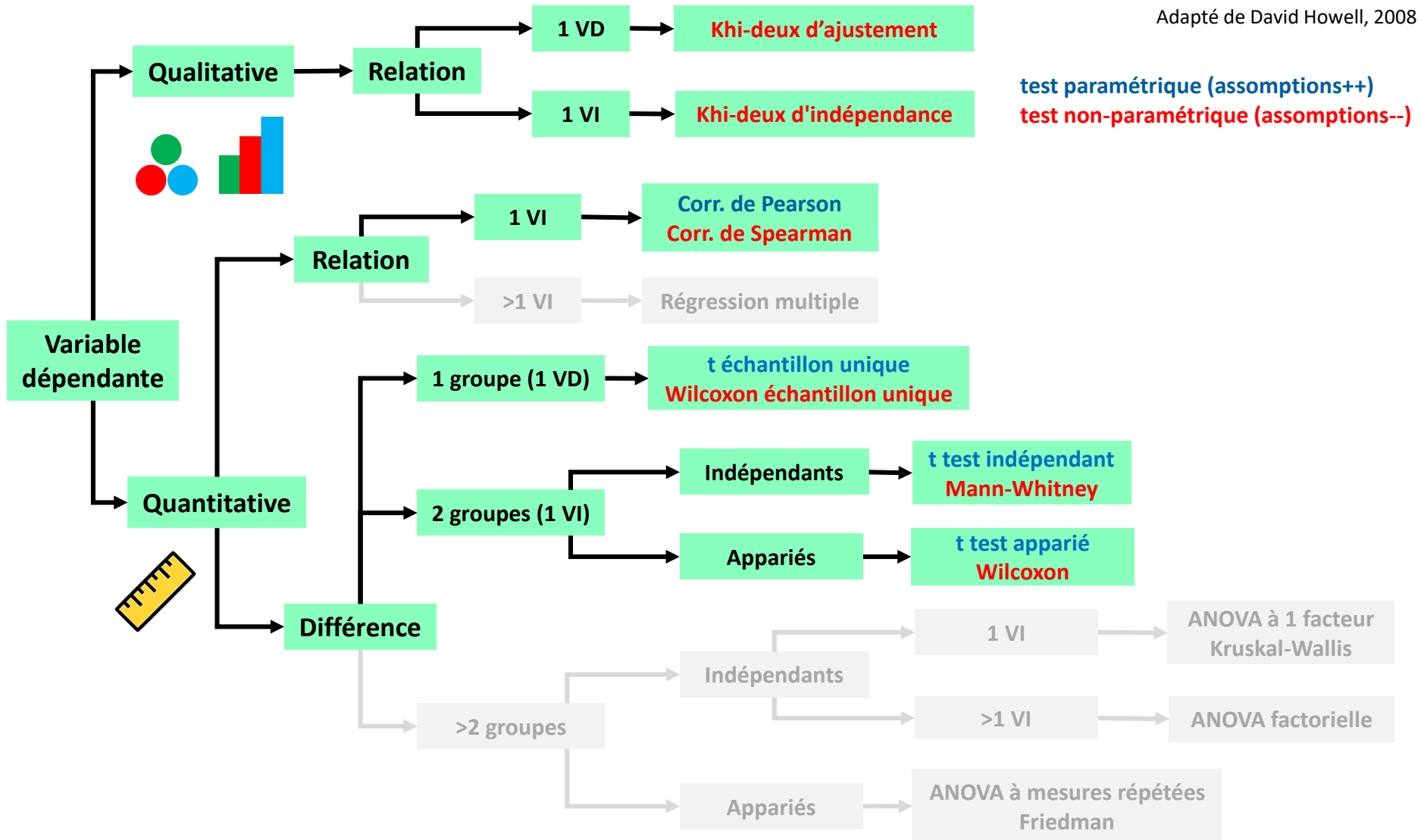
Variable : nomenclature

- Le choix du test statistique va dépendre de l'**échelle** des VD et VI



Arbre de décision statistique

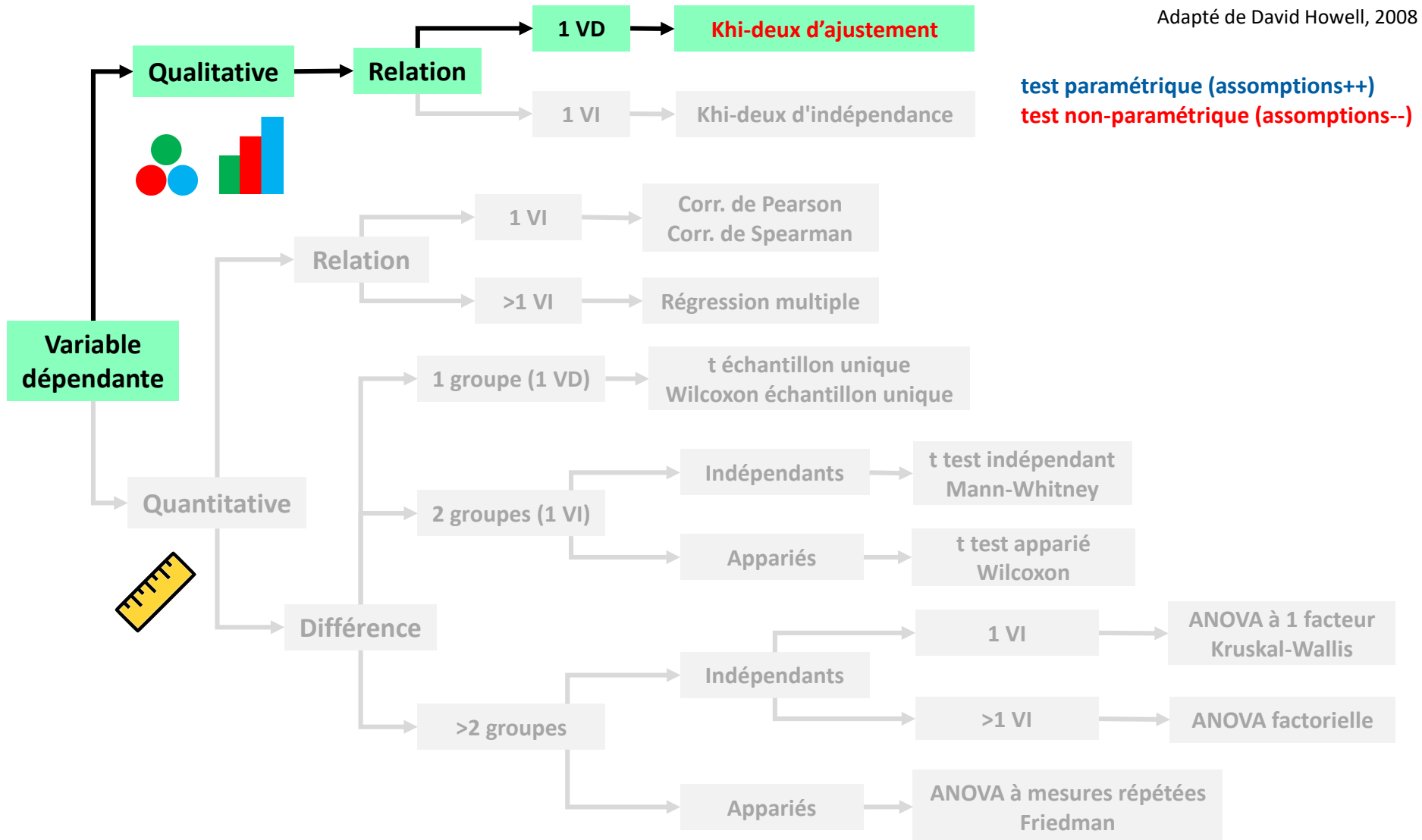
Adapté de David Howell, 2008



Khi-deux d'ajustement

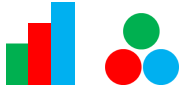


Adapté de David Howell, 2008



Khi-deux d'ajustement

H_0 = hyp nulle
 H_1 = hyp alternative

- Contexte
- VD : 1 variable qualitative (nominale, ordinale) 
- H_0/H_1 : la répartition des individus au sein des différentes catégories de la VD est **homogène** / **hétérogène** dans la **POPULATION**
- Questions de recherche abordées ici :
 - **Q1** : Est-ce que le nombre de personnes avec bac, licence, ou master diffère dans la POPULATION ?
 - **Q2** : Est-ce qu'il y a plus de femmes que d'hommes dans la POPULATION ?

Khi-deux d'ajustement

- **Q1** : Est-ce que le nombre de personnes avec bac, licence, ou master diffère **dans la POPULATION** ?

```
DF <- readxl::read_xlsx("Xhi-deux.xlsx")  
str(DF);head(DF)
```

##	sujet	etude	sexe
## 1	1	master	f
## 2	2	bac	f
## 3	3	bac	h
## 4	4	bac	f
## 5	5	licence	f
## 6	6	master	f

- **Analyse en deux étapes** :
 - stat **descriptives** (tables de fréquences, graphiques)
 - stat **inférentielles** (test d'hypothèse nulle)

$n = 50$ sujets

Khi-deux d'ajustement

- Statistiques descriptives

```
EFFECTIF <- table(DF$etude)
addmargins(EFFECTIF)                # effectifs bruts

EFFECTIF_PROP <- prop.table(EFFECTIF)
addmargins(EFFECTIF_PROP)          # effectifs en pourcentage

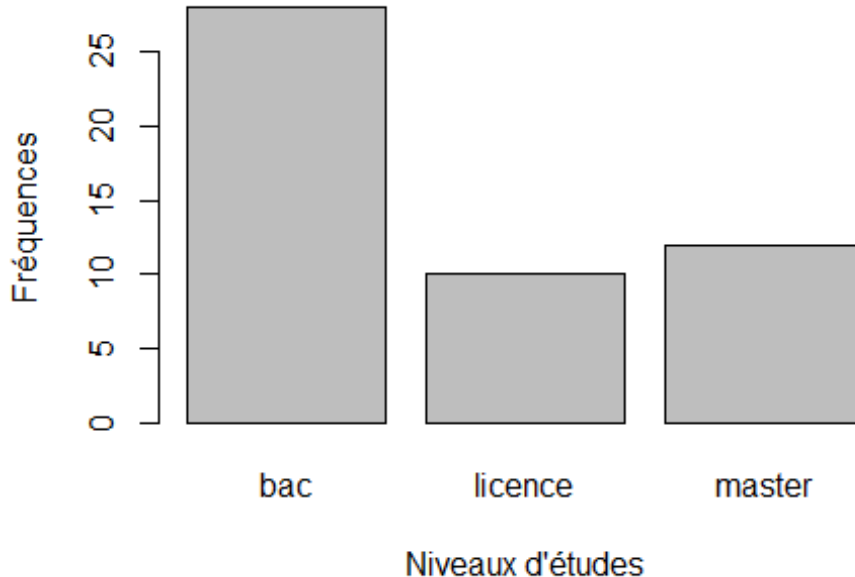
# il faut installer le paquet prettyR
prettyR::describe(DF)              # les deux à la fois
```

```
## etude      bac master licence
## Count      28      12      10
## Percent    56      24      20
## Mode bac
```

Khi-deux d'ajustement

graphique à barres

```
graphics::barplot(EFFECTIF,                                     # table d'effectifs
                  xlab = "Niveaux d'études",                   # nom axe x
                  ylab = "Fréquences")                         # nom axe y
```



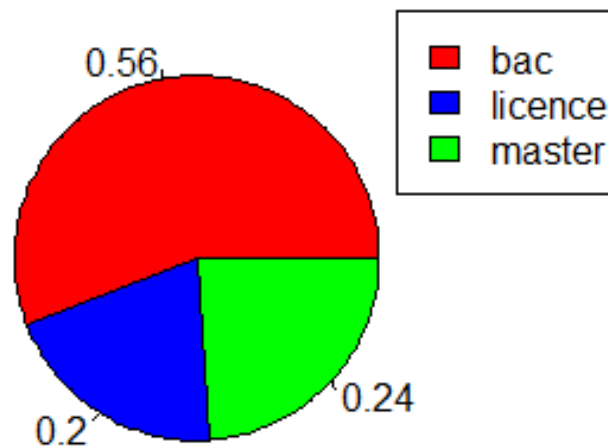
- Utilisez le paquet `graphics` pour exécutez des plots rapidement
- Utilisez `ggplot2` pour créer des graphiques de toute beauté (voir derniers TDs)

Khi-deux d'ajustement

```
# graphique camembert
graphics::pie(EFFECTIF_PROP,           # table d'effectifs
               labels = EFFECTIF_PROP, # table d'effectifs
               main   = "Niveaux d'études", # titre
               col    = c("red", "blue", "green")) # couleur des tranches

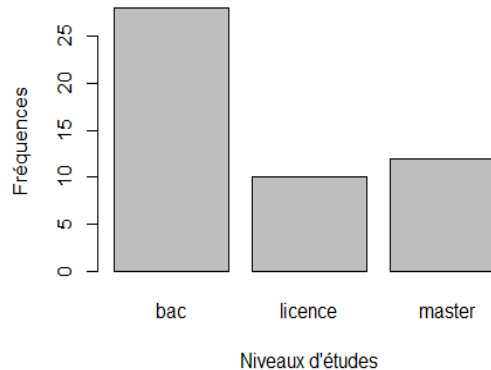
# rajoute une légende
legend("topright",                    # position
       legend=names(EFFECTIF),        # noms de légende
       fill   = c("red", "blue", "green")) # couleurs de légende
```

Niveaux d'études

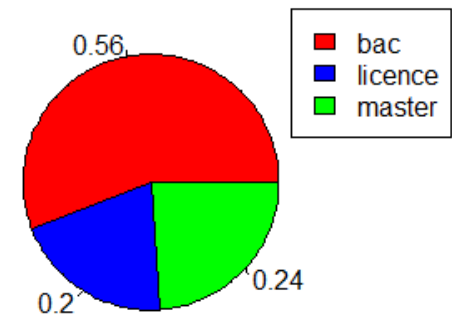


Khi-deux d'ajustement

##	etude	bac	master	licence
##	Count	28	12	10
##	Percent	56	24	20

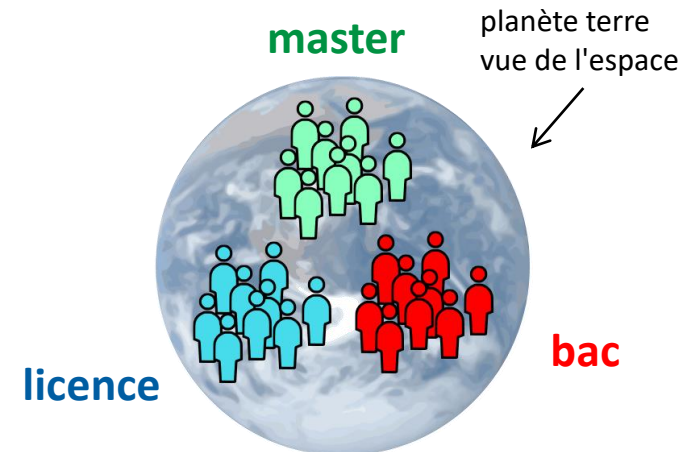


Niveaux d'études



Est ce que ces différences de proportions sont assez grandes pour pouvoir rejeter l'hypothèse nulle et être généralisées à la population ?

→ Seul le test du *khi-deux d'ajustement* peut nous donner la réponse



représentation imagée de H0

Khi-deux d'ajustement

- Statistiques inférentielles

la table des effectifs

```
KHI = stats::chisq.test(EFFECTIF)
KHI

##
## Chi-squared test for given probabilities
##
## data:  EFFECTIF
## X-squared = 11.68, df = 2, p-value = 0.002909
```

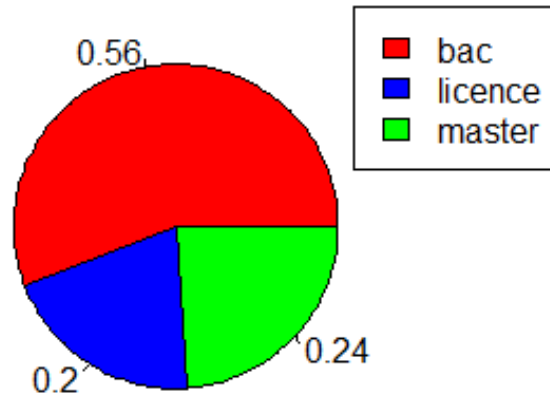
valeur du chi2
observé

degrés de libertés
(ici nombre de
modalités - 1)

probabilité d'obtenir un chi2
observé aussi extrême **si H0**
était vraie

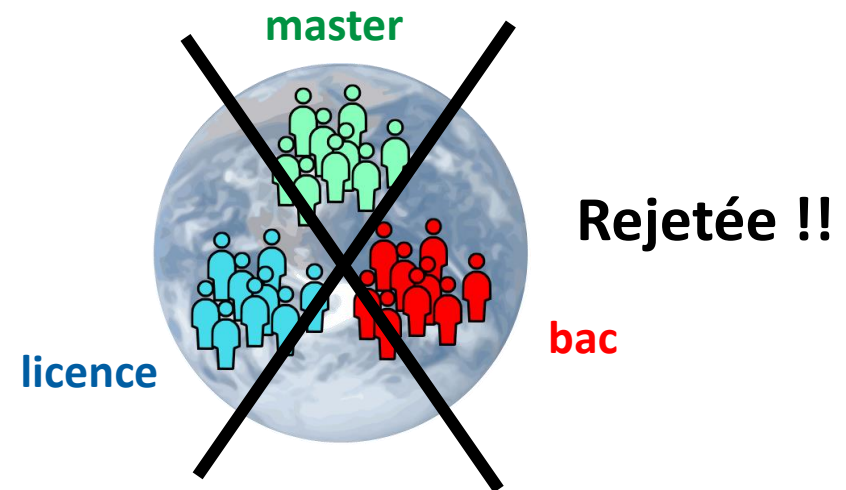
Khi-deux d'ajustement

Niveaux d'études



```
## Chi-squared test for given
## probabilities
##
## data:  EFFECTIF
## X-squared = 11.68, df = 2,
## p-value = 0.002909
```

Est ce que ces différences de proportions sont assez grandes pour rejeter l'hypothèse nulle et être généralisées à la population ? OUI !!



→ OUI les différences de proportions sont beaucoup trop IMPROBABLES si H_0 est vraie

→ On rejette H_0 et on GENERALISE nos résultats à la POPULATION

Khi-deux d'ajustement

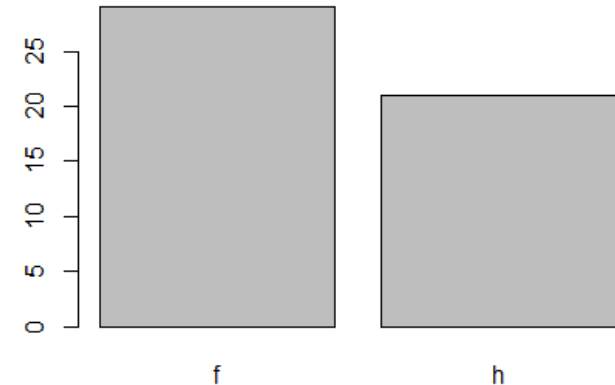
- Exemples de rédaction

Selon le test du Khi-deux d'ajustement, les 3 niveaux d'études (bac : [n = 28, 56%], licence : [n = 10, 20%], master : [n = 12, 24%]) était répartis de façon significativement hétérogène ($\chi^2[\text{df} = 2, N = 50] = 11.68, p < .01$).

According to the chi-square test of goodness-of-fit, the three education levels were not equally distributed in the population, $\chi^2[\text{df} = 2, N = 50] = 11.68, p < .01$.

Khi-deux d'ajustement (test binomial)

- **Q2** : Est-ce qu'il y a plus de femmes que d'hommes dans la **POPULATION** ?
- Quand la VD nominale n'a que **deux catégories**, le test binomial est plus adapté



```
EFFECTIF <- table(DF$sexe)
barplot(EFFECTIF)
binom.test(EFFECTIF)
```

```
## Exact binomial test
##
## data:  EFFECTIF
## number of successes = 29, number of trials = 50, p-value = 0.3222
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4320604 0.7181178
## sample estimates:
## probability of success
## 0.58
```

Annotations:

- nombre de femmes** (points to 29)
- nombre de sujets** (points to 50)
- valeur p** (points to 0.3222)
- % de femmes** (points to 0.58)

Khi-deux d'ajustement (test binomial)

- Exemple de rédaction

Selon le test binomial, le pourcentage de femmes ($n = 29$, 58%) n'était significativement pas supérieur au pourcentage des hommes ($n = 21$, 42%, $p = .32$).

The binomial test indicated that the number of women did not differ significantly from the number of men, $p = .32$

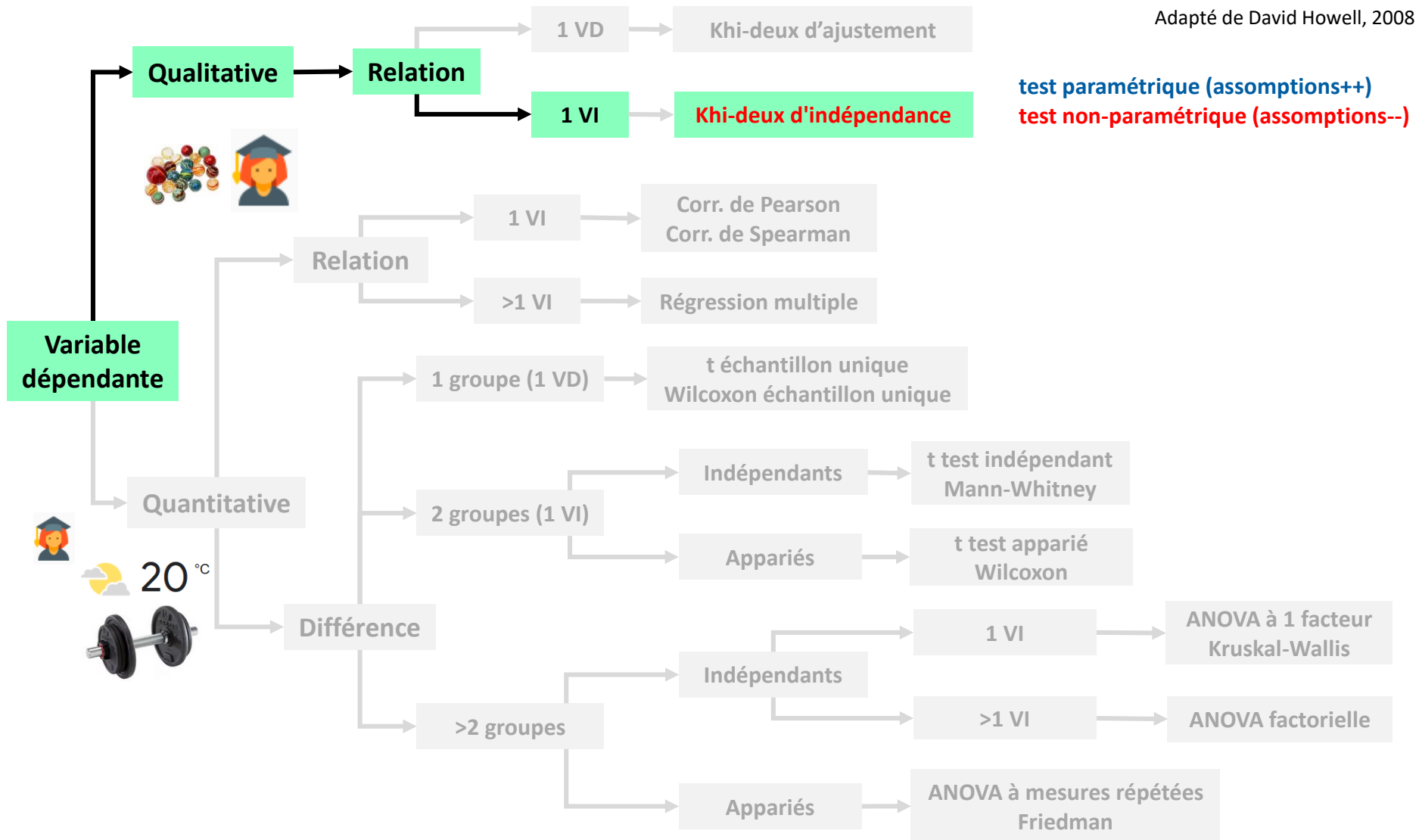
Khi-deux d'ajustement

- Mémo

VI-VD	stat descriptive	stat inférentielle
1 VD nominale	table() prop.table() prettyR::describe() barplot() pie()	chisq.test() binom.test()

Khi-deux d'indépendance


Adapté de David Howell, 2008



Khi-deux d'indépendance

H_0 = hyp nulle

H_1 = hyp alternative

- Contexte 
 - VD : 1 variable qualitative (nominale, ordinale)
 - VI : 1 variable qualitative (nominale, ordinale)
- ↔ permutables
- H_0/H_1 : la répartition des individus entre les catégories de la VD ne **diffère pas** / **diffère** en fonction des catégories de la VI. Il y a **indépendance** / **dépendance** entre les 2 variables dans la POPULATION
 - Exemple abordé ici :
 - Est ce que les femmes ont des niveaux d'études plus élevés que les hommes ?

##	sujet	etude	sexe
## 1	1	master	f
## 2	2	bac	f
## 3	3	bac	h
## 4	4	bac	f
## 5	5	licence	f

Khi-deux d'indépendance

- Statistiques descriptives

```
EFFECTIF <- table(DF$sexe, DF$etude)
addmargins(EFFECTIF)

EFFECTIF_PROP <- prop.table(EFFECTIF)
addmargins(EFFECTIF_PROP)
```

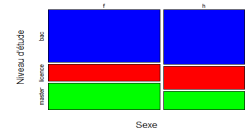
##		bac	licence	master	Sum
##	f	16	5	8	29
##	h	12	5	4	21
##	Sum	28	10	12	50

##		bac	licence	master	Sum
##	f	0.32	0.10	0.16	0.58
##	h	0.24	0.10	0.08	0.42
##	Sum	0.56	0.20	0.24	1.00

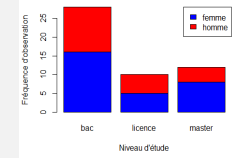
Khi-deux d'indépendance

- Statistiques descriptives

```
graphics::mosaicplot(EFFECTIF,
                      main = "",
                      xlab = "Sexe",
                      ylab = "Niveau d'étude",
                      col = c("blue", "red", "green"))
```

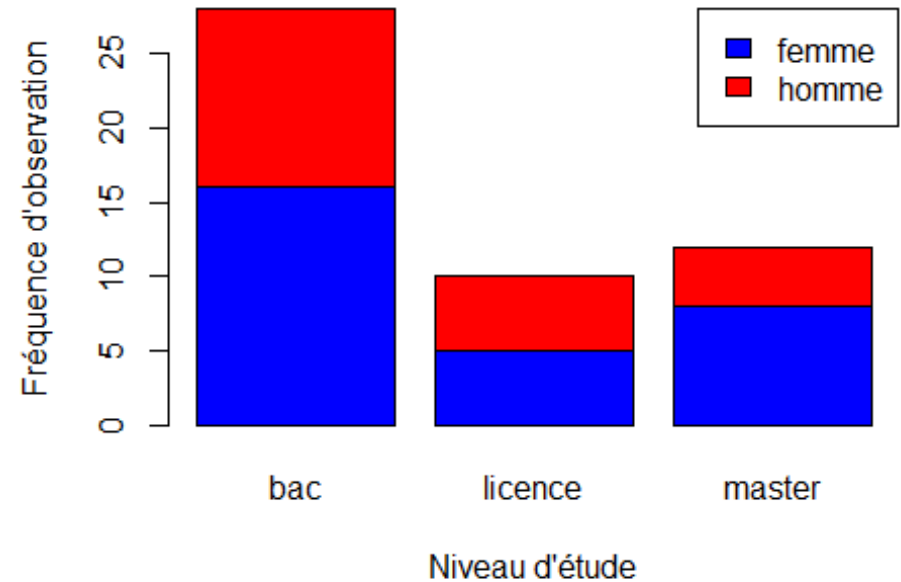
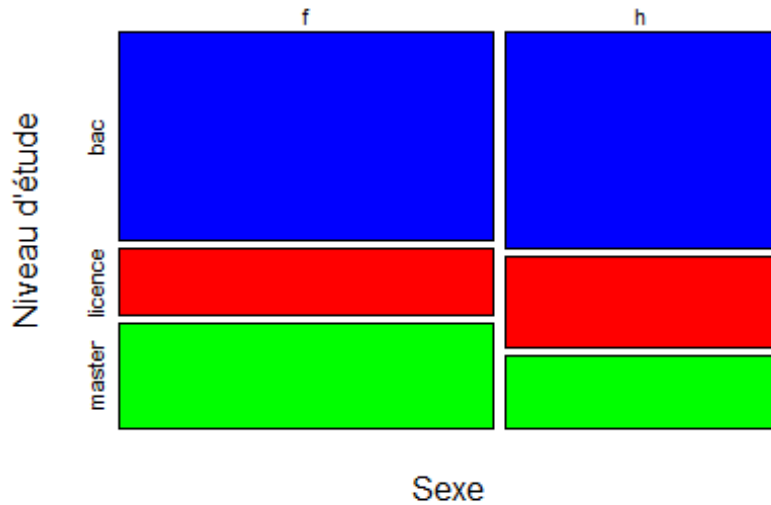


```
barplot(EFFECTIF,
        xlab = "Niveau d'étude",
        ylab = "Fréquence d'observation",
        col = c("blue", "red"))
legend("topright", legend = c("femme", "homme"), fill = c("blue", "red"))
```



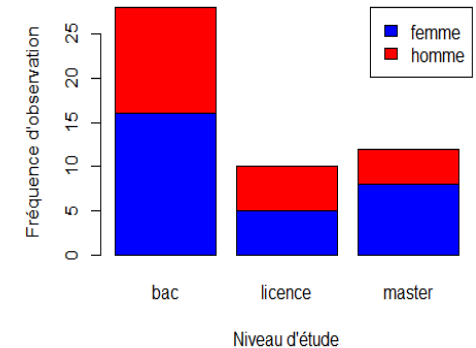
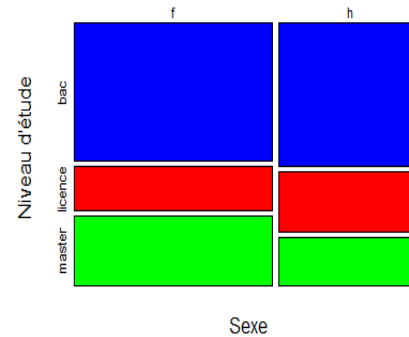
Khi-deux d'indépendance

- Statistiques descriptives



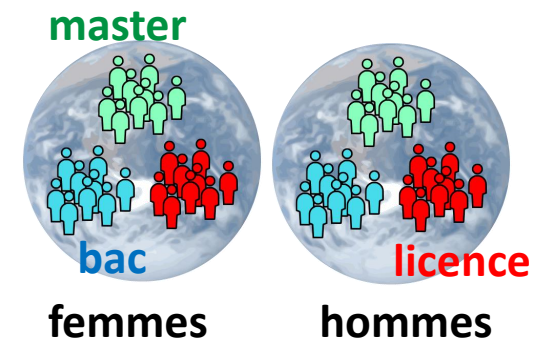
Khi-deux d'indépendance

##		bac	licence	master	Sum
##	f	16	5	8	29
##	h	12	5	4	21
##	Sum	28	10	12	50



Est ce que ces différences (certes minimales ici) de diplômes entre homme et femmes sont assez grandes pour pouvoir rejeter l'hypothèse nulle et être généralisées à la population ?

→ Seul le test du *khi-deux d'indépendance* peut nous donner la réponse



représentation imagée de H0 ; le diplôme est indépendant du sexe

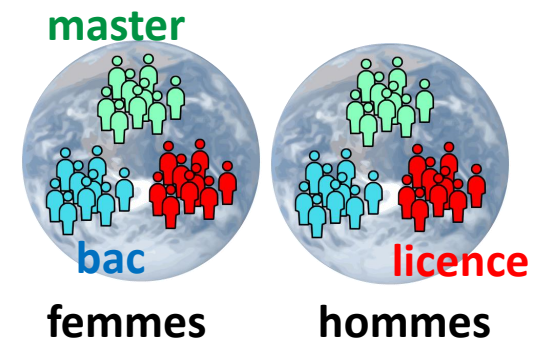
Khi-deux d'indépendance

- Statistiques inférentielles

```
KHI <- chisq.test(EFFECTIF)
KHI
## Pearson's Chi-squared test
## data:  EFFECTIF
## X-squared = 0.64118, df = 2, p-value = 0.7257
```

- Les différences de diplômes entre hommes et femmes observées dans notre échantillon sont donc **PROBABLES** si H_0 est vraie
- On **NE** rejette **PAS** H_0 et on **NE** GENERALISE **PAS** nos résultats à la POPULATION
- /!\ Cependant on **N'ACCEPTE PAS** H_0 car H_0 ne peut pas être prouvé /!\

Non rejetée



représentation imagée de H_0 ; le diplôme est indépendant du sexe

Khi-deux d'indépendance

- Exemple de rédaction

Nous n'avons pas observé d'association significative entre le niveau d'étude et le sexe ($\chi^2(df = 2, N = 50) = 0.64, p = .73$). Contrairement à notre hypothèse, les femmes n'avaient pas plus de plus hauts diplômes que les hommes.

The association between sex and education level was not significant ($\chi^2(df = 2, N = 50) = 0.64, p = .73$).



Quelques exemples en anglais ici :

<https://www.socscistatistics.com/tutorials/chisquare/default.aspx>

Khi-deux d'ajustement

- Mémo

VI-VD	stat descriptive	stat inférentielle
1 VD nominale 1 VI nominale	table() prop.table() barplot() mosaicplot()	chisq.test()

Exercice 5

- Importez et inspectez le DF "Xhi-deux_exo.xlsx"
- Réalisez un **test binomial** sur la variable "aime_coriandre"
- Réalisez un **chi2 d'ajustement** sur la variable "opinion_politique"
- Réalisez un **chi2 d'indépendance** entre ces deux variables
- **A chaque fois suivez les étapes :**
 - stat descriptives
 - graphique
 - stat inférentielles