

## Introduction

### Statistique Appliquée Statistiques

Luc Deneire – deneire@unice.fr

Xidian University, Polytech Nice Sophia

Avril 2020

#### Informations

- Cours sur : <http://jalon.unice.fr/public/pqg729/>
- Questions par e-mail : [deneire@unice.fr](mailto:deneire@unice.fr)

#### La statistique pour modéliser *le monde*

Permet de répondre à des questions du type

- Quelle est l'efficacité d'une usine ?
- Quelle est l'influence d'un changement dans un process ?

EN

- Recueillant des données
- En déduisant un modèle
- En appliquant les modifications à ce modèle

1

2

### Plan du cours - Partie Statistiques

- 1 Statistique : définitions
- 2 Statistique descriptive
- 3 Introduction à l'échantillonnage
- 4 Introduction à la statistique inférentielle
- 5 Estimation ponctuelle de paramètres
- 6 Estimation par intervalle de confiance
- 7 Test d'hypothèses liés aux moyennes, variance et covariance
- 8 Test d'adéquation à une loi
- 9 Régression linéaire et corrélation
- 10 Régression multi-linéaire
- 11 Petite conclusion

### Qu'est-ce qu'une statistique

#### Une statistique

est une quantité calculée à partir d'un certain nombre d'observations.

exemple : la moyenne; la médiane

#### Signification des termes dépendante du contexte

- En statistique : moyenne basée sur des observations
- En probabilité : moyenne basée sur un modèle probabiliste

**Un individu**

est l'unité statistique de base.

*exemple du sondage : un "sondé"*

**Une population**

est l'ensemble des individus que l'on souhaite étudier. Cette population peut être infinie ou finie.

*exemple de l'élection présidentielle : les "citoyens"*

**Un échantillon**

(d'une population) est un sous-ensemble de la population.

*exemple du sondage : 1034 "sondés"*

5

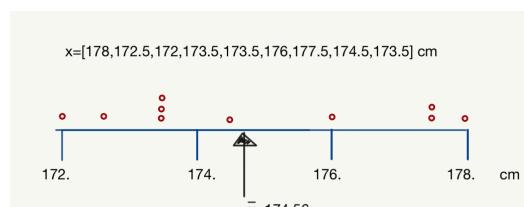
**Statistique Descriptive**

Moyenne et médiane

**Statistique descriptive : moyenne****DEFINITION :Moyenne de l'échantillon (sample mean)**

Soient  $n$  observations d'une population, notées  $x_1, x_2, \dots, x_n$ , la **moyenne de l'échantillon** est :

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$



Moyenne et médiane

6

**Statistique Descriptive**

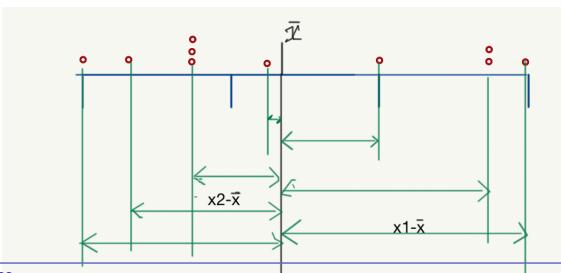
Moyenne et médiane

**Statistique descriptive : variance et écart-type****DEFINITION :Variance de l'échantillon (sample variance)**

Soient  $n$  observations d'une population, notées  $x_1, x_2, \dots, x_n$ , la **moyenne de l'échantillon** est :

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

l' **écart-type** vaut la racine carrée de la variance :  $s$ .



Moyenne et médiane

7

8

## Médiane et Mode

Soit un échantillon de  $N$  individus  $x_{(i)}$ , mis dans un ordre croissant, tels que

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$$

La **Médiane**; notée  $x_{\frac{N}{2}}$  est telle que la moitié des individus ont une valeur inférieure à  $x_{\frac{N}{2}}$  (et l'autre moitié une valeur supérieure). Si  $N$  est impair,

$$x_{\frac{N}{2}} = x_{(\frac{N+1}{2})}, \text{ si } N \text{ est pair}; x_{\frac{N}{2}} = 0.5 \times [x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}]$$

Le **Mode** est la valeur qui apparaît le plus souvent.

Dans notre exemple :  $x$  ordonnée vaut

$x = [172, 172.5, 173.5, 173.5, 173.5, 174.5, 176, 177.5, 178]$ ,  $N = 9$  est impair et  $x_{\frac{N}{2}} = x_{(5)} = 173.5$ .

On notera que la médiane est différente de la moyenne

Le mode vaut également 173.5

## Quartiles, quantiles, déciles

Soit un échantillon de  $N$  individus  $x_{(i)}$ , mis dans un ordre croissant, tels que

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$$

Le **Quantile**; notée  $x_p$ , aussi appelé  $p$ -quantile est tel que la fraction  $p$  des individus ont une valeur inférieure à  $x_p$

On a en particulier les premier, deuxième et troisième **Quartiles**, notés  $q_1, q_2$  et  $q_3$  tels que 25 % des valeurs de l'échantillon sont inférieures à  $q_1$ ,  $q_2 = x_{\frac{N}{2}}$  et  $q_3$  est tel que 75 % des valeurs de l'échantillon sont inférieures à  $q_3$ .

On note IQR l'**Intervalle Inter Quantile**  $IQR = q_3 - q_1$ .

Les **Déciles**; notés  $d_n$ ; sont tels que  $10^*n$  % des valeurs de l'échantillon sont inférieures à  $d_n$  (par exemple, 20 % des valeurs de l'échantillon sont inférieures à  $d_2$ ).

Les **Percentiles** sont les quantiles exprimés en pourcents.

## Quartiles, quantiles, déciles

```
6]: import numpy as np
#TRIAL=np.random.random(50)*5-3
#TRIAL=(TRIAL*100//1)/100
print(TRIAL)
print(np.sort(TRIAL))
print('La moyenne vaut : ',np.mean(TRIAL))
print('L\'écart-type vaut : ',np.std(TRIAL))

print('La médiane vaut : ',np.percentile(TRIAL,50))
print('Le premier quartile vaut : ',np.percentile(TRIAL,25))
print('Le troisième quartile vaut : ',np.percentile(TRIAL,75))
print('L\'espace interquartile vaut : ',np.percentile(TRIAL,75)-np.percentile(TRIAL,25))

[-2.67 -1.39 -0.18 -2.77  1.18  0.76  0.68  0.82 -2.28 -1.41 -1.47 -1.51
-2.8  1.98  1.8  -0.2 -1.52 -0.34 -1.91 -1.52  1.12  1.02 -2.96 -0.78
0.45 -1.72 -0.35  1.74 -0.97 -2.09 -2.29 -2.47 -0.8  -2.71  0.13 -2.77
-1.49 -1.58 -0.54 -2.7 -1.47 -2.32  0.34 -2.84  1.11  0.22  1.06 -2.68
-2.  -1.58]
[-2.96 -2.84 -2.8  -2.77 -2.77 -2.71 -2.7  -2.68 -2.67 -2.47 -2.32 -2.29
-2.28 -2.09 -2.  -1.91 -1.72 -1.58 -1.58 -1.52 -1.52 -1.51 -1.49 -1.47
-1.47 -1.41 -1.39 -0.97 -0.8  -0.78 -0.54 -0.35 -0.34 -0.2  -0.18  0.13
0.22  0.34  0.45  0.68  0.76  0.82  1.02  1.06  1.11  1.12  1.18  1.74
1.8  1.98]
La moyenne vaut : -0.9333999999999999
L'écart-type vaut : 1.454447812745442
La médiane vaut : -1.44
Le premier quartile vaut : -2.2325
Le troisième quartile vaut : 0.31
L'espace interquartile vaut : 2.5425
```

## Vers l'histogramme : Tiges et Feuilles

On peut classer assez simplement des données en CLASSES.

Une classe est un intervalle de valeurs (par exemple  $x_L \leq x < x_H$ ).

Supposons qu'on ait un intervalle  $3.0 \leq x < 4.0$ ; on considère alors que "3" est une tige, et que les données du type 3.9 ont une tige qui vaut 3 et une feuille qui vaut 0.9. L'affichage de ce tableau de "Tiges et Feuilles", avec, par feuille, le nombre de feuilles sur une tige, est un premier pas vers l'histogramme.

Par exemple, pour les données suivantes :

4.14	4.26	4.99	5.56	5.6	5.67	5.86	6.39	6.57	6.72	7.12	7.24
7.29	7.54	7.61	7.64	7.69	7.72	7.72	7.76	7.79	8.03	8.11	8.11
8.12	8.16	8.23	8.24	8.24	8.43	8.43	8.51	8.51	8.55	8.76	9.04
9.16	9.2	9.27	9.36	9.73	9.75	9.81	9.85	9.87	9.95	9.95	10.02
10.07	10.26	10.29	10.36	10.36	10.52	10.57	10.6	10.64	10.71	10.73	10.83
10.88	10.9	11.02	11.03	11.03	11.13	11.14	11.15	11.2	11.24	11.24	11.36
11.65	11.84	11.9	11.92	12.	12.03	12.47	12.53	12.62	12.63	12.66	12.67
12.8	12.82	12.84	12.88	12.93	12.97	12.98	13.34	13.55	13.58	13.61	13.76
13.96	14.02	14.54	15.23								

## Vers l'histogramme : Tiges et Feuilles

On obtient le diagramme de Tiges et Feuilles suivant :

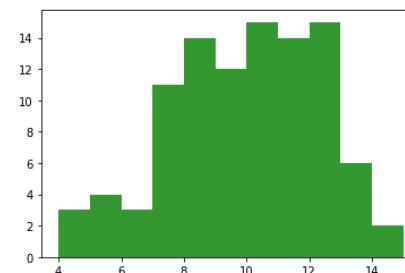
```
import matplotlib.pyplot as plt
import sys
def printf(format, *args):
    sys.stdout.write(format % args)
stems = TRIAL//1
stem=stems[0]
i=0
while( i < (np.size(TRIAL))):
    printf(" %2d :" ,stems[i])
    while(stems[i]==stem):
        printf(" %.2f",TRIAL[i]*stems[i])
        i=i+1
    stem=stems[i]
    printf("\n")
```

4 : 0.14 0.26 0.99  
 5 : 0.56 0.60 0.67 0.86  
 6 : 0.39 0.57 0.72  
 7 : 0.12 0.24 0.29 0.54 0.61 0.64 0.69 0.72 0.72 0.76 0.79  
 8 : 0.03 0.11 0.12 0.16 0.23 0.24 0.24 0.43 0.43 0.51 0.51 0.55 0.76  
 9 : 0.04 0.16 0.20 0.27 0.36 0.73 0.75 0.81 0.85 0.87 0.95 0.96  
 10 : 0.02 0.07 0.26 0.29 0.36 0.36 0.52 0.57 0.60 0.64 0.71 0.73 0.83 0.88 0.90  
 11 : 0.02 0.03 0.03 0.13 0.14 0.15 0.20 0.24 0.24 0.36 0.65 0.84 0.90 0.92  
 12 : 0.00 0.03 0.47 0.53 0.62 0.63 0.66 0.67 0.80 0.82 0.84 0.88 0.93 0.97 0.98  
 13 : 0.34 0.55 0.58 0.61 0.76 0.96  
 14 : 0.02 0.54  
 15 : 0.23

## Histogramme

Un **Histogramme** est un graphique permettant de représenter la répartition des données en la représentant avec des colonnes verticales. En abscisse, on divise les valeurs en classes de valeurs, et en ordonnée, on indique les fréquences absolues ou les fréquences relatives dans ces classes. En quelque sorte, c'est le diagramme "Tiges et Feuilles" tourné de 90 degrés.

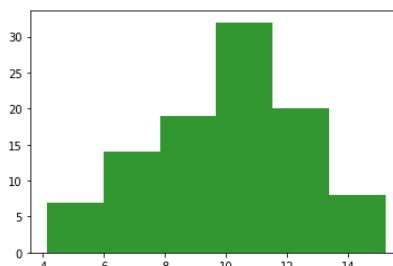
```
num_bins = 12
n, bins, patches = plt.hist(TRIAL,| bins=stems, facecolor='green', alpha=0.8)
plt.show()
```



## Histogramme

Un **Histogramme** est un graphique permettant de représenter la répartition des données en la représentant avec des colonnes verticales. En abscisse, on divise les valeurs en classes de valeurs, et en ordonnée, on indique les fréquences absolues ou les fréquences relatives dans ces classes.

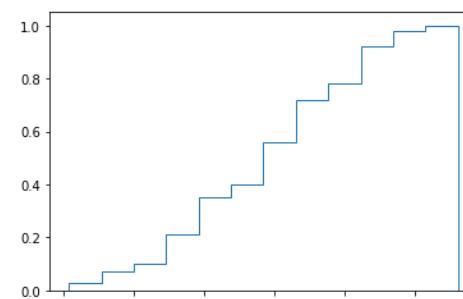
```
num_bins = 6
n, bins, patches = plt.hist(TRIAL, num_bins, facecolor='green', alpha=0.8)
plt.show()
```



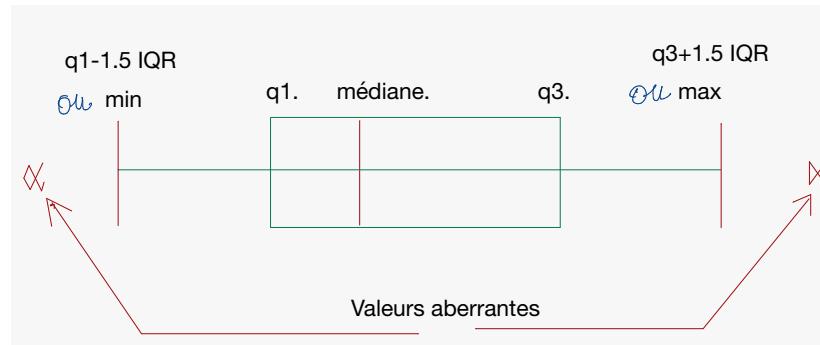
## Fonction de répartition (Cumulative Density Function-cdf)

La **cdf** est un graphique permettant de représenter la répartition des données en la représentant avec des colonnes verticales. En abscisse, on divise les valeurs en classes de valeurs, et en ordonnée, on indique le nombre ou la proportion de valeurs qui sont supérieure à la valeur la plus élevée de la classe.

```
# plot the cumulative histogram
n, bins, patches = plt.hist(TRIAL, bins, density=True, histtype='step',
                           cumulative=True, label='Empirical')
```



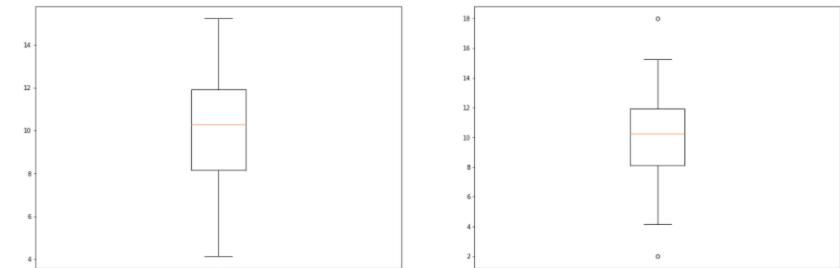
## Boîte à moustaches (whisker plot)



boîte à moustaches

## Boîte à moustaches (whisker plot)

```
fig,axs=plt.subplots(1,2,figsize=(24,8))
TRIALex=np.concatenate(([2],TRIAL,[18]))
axs[0].boxplot(TRIAL)
axs[1].boxplot(TRIALex)
plt.show()
```



17

boîte à moustaches

18

## Diagramme de Probabilités / Droite de Henry (Probability plot)

Les données précédentes pourraient être issues d'une loi Gaussienne. Le diagramme de probabilités permet de vérifier si les données d'un échantillon "semblent" issues d'une loi donnée. On dit alors qu'il y a adéquation entre les données et la loi. S'il s'agit d'une adéquation à la loi Gaussienne, ce diagramme de probabilités s'appelle la droite de Henry.

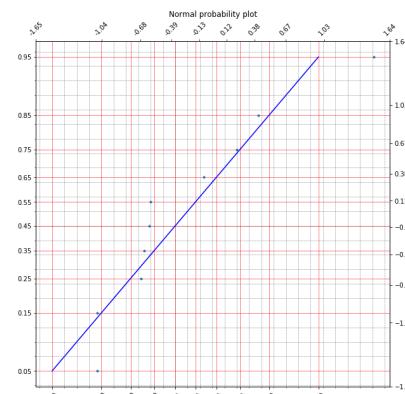
Soit un échantillon de  $n$  observations (données), on range ces observations dans l'ordre croissant telles que  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ . Ensuite, calculer  $n$  fréquences cumulées équiréparties sur  $[0, 1]$  ( $\frac{j-0.5}{n}$ ). Tracer les points de coordonnées  $(x_{(j)}, F_Z\left(\frac{j-0.5}{n}\right))$ ; où  $Z$  est la distribution normale centrée. Si ces points sont alignés, alors les données sont compatible avec un modèle Gaussien (il est possible que les données soient issues d'une population qui suit une loi Gaussienne).

boîte à moustaches

## Diagramme de Probabilités / Droite de Henry (Probability plot)

```
1 n=10 ; val=((np.arange(n)+0.5)/n
2 x = stats.norm.rvs(size=n)**10+10;
3 quantiles = stats.norm.ppf(val,0,1);
4 mean=np.mean(x)
5 median=np.median(x)
6 std=np.std(x)
7 labels = (quantiles*std*mean)**10;
8 ax.plot(quantiles,quantiles,'b-');
9 ax.set_xticks(quantiles);ax.set_yticks(quantiles);
10 ax.set_xlabel('Quantiles');
11 ax.set_ylabel('Normal probability plot');
12 print(np.sort(x));print((np.sort(x)-mean)/std,quantiles,'.')
13 ax.plot((np.sort(x)-mean)/std,quantiles,'.')

[-3.8080985 -3.7859277 3.93143899 4.50139982 5.42315314 5.64808115
-1.96339 -20.77161434 24.49913871 44.84736572]
[-1.08959692 -1.08855862 -0.54607356 -0.50600712 -0.44121315 -0.42540201
 0.22942474 0.63769547 0.89964872 2.33008244]
```

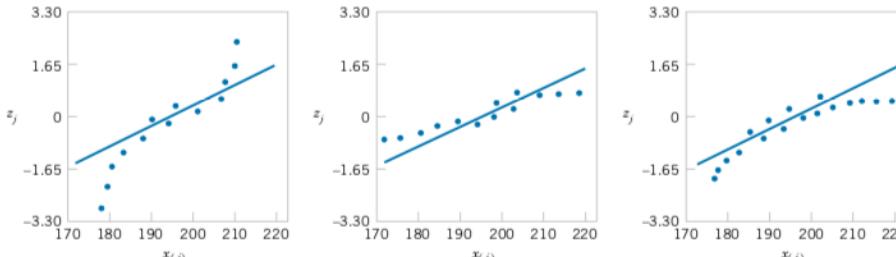


19

diagramme de probabilités

20

## Diagramme de Probabilités / Droite de Henry (Probability plot)



- (a) Distribution à queue légère (moins de valeurs extrêmes que la normale)
- (b) Distribution à queue lourde (plus de valeurs extrêmes que la normale)
- (c) Distribution asymétrique à droite (moins de valeurs faibles que la normale, plus de valeurs grandes que la normale)

## L'échantillonnage : une expérience aléatoire

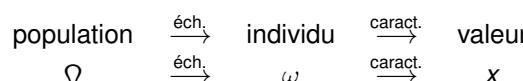
### DEFINITION : Échantillonnage

choisir au hasard  $n$  individus de la population  
Il y a deux types d'échantillonnage :

- 1 avec remplacement de l'individu choisi, ce qui mène à un traitement théorique plus simple - population infinie ;
- 2 sans remplacement: échantillonnage exhaustif, ce qui est une procédure naturelle ou obligatoire (contrôle destructif).

## L'échantillonnage : une expérience aléatoire

L'échantillonnage est une **expérience aléatoire**, : choisir au hasard un individu (ou un "petit nombre" d'individus) de la population.  
Chaque individu doit avoir la même probabilité d'être choisi.



Enfin, à partir de l'échantillon, on étudie la variable aléatoire  $X$  associée au caractère étudié et on peut, dans le meilleur des cas, déterminer la densité de probabilité  $f_X(x)$  (ou sa masse de probabilité  $p_X(x)$  s'il s'agit d'une variable aléatoire discrète).

Soit  $X$  une v.a. sur la population ( $f_X(x)$ )

l'échantillonnage correspond à la répétition de  $n$  expériences aléatoires identiques, :  $n$  v.a. indépendantes  $X_i$  ( $i = 1, \dots, n$ ) ayant la (même) densité de probabilité  $f_X(x)$ .

l'échantillonnage correspond à la répétition de  $n$  expériences aléatoires identiques

$X_i$  i.i.d. (indépendantes et identiquement distribuées).

$$f_{X_1}(x) = f_{X_2}(x) = \dots = f_{X_n}(x) = f_X(x) \text{ et} \\ f_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) = f_X(x_1) f_X(x_2) \dots f_X(x_n).$$

## Théorème de la limite centrale (Central Limit Theorem CLT)

Echantillonage avec remplacement et même probabilité de choisir chaque individu.

**DEFINITION :** Une statistique est

une fonction des variables aléatoires  $X_i$  ( $i = 1, \dots, n$ ) obtenue à partir d'un échantillon.

Une statistique est une variable aléatoire !

Soit  $n$  v.a. (i.i.d) :

- $X_1, X_2, \dots, X_n$ : série de v.a. indépendantes
- $f_{X_1}(x) = \dots = f_{X_n}(x) = f_X(x)$  (même distribution)
- $E[X_1] = \dots = E[X_n] = \mu_X$ ,  $\sigma_{X_1} = \dots = \sigma_{X_n} = \sigma_X$
- $S_n = X_1 + X_2 + \dots + X_n$ ,  $E[S_n] = n\mu_X$ ,  $\sigma_{S_n}^2 \stackrel{\text{ind}}{=} n\sigma_X^2$

$$Z_n = \frac{S_n - \mu_{S_n}}{\sigma_{S_n}} = \frac{X_1 + X_2 + \dots + X_n - n\mu_X}{\sqrt{n}\sigma_X}, \quad E[Z_n] = 0, \quad \sigma_{Z_n}^2 = 1$$

## Théorème Limite Central (2)

Théorème Limite Central

$$\lim_{n \rightarrow \infty} P(\{Z_n \leq z\}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z \exp\left(-\frac{1}{2}u^2\right) du$$

Théorème Limite Central

$$n \rightarrow \infty : Z_n \rightarrow N(0, 1), \quad S_n \rightarrow N(n\mu_X, n\sigma_X^2), \quad \frac{S_n}{n} \rightarrow N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

[http://onlinestatbook.com/stat\\_sim/sampling\\_dist/](http://onlinestatbook.com/stat_sim/sampling_dist/)

■ Échantillon aléatoire de taille  $n$ ; moyenne  $\bar{X}$

■ Population normale  $N(\mu, \sigma^2)$

- $\bar{X}$ : normale (combinaison linéaire de v.a. normales)
- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$  ( $\sigma$  connu)

■ Population non normale ( $\sigma$  connu)

- $n > 30$ :  $\bar{X} = N\left(\mu, \frac{\sigma^2}{n}\right)$  (tlc)
- $n < 30$ :  $\bar{X} = N\left(\mu, \frac{\sigma^2}{n}\right)$  si  $p_X(x)$  presque normale

■ Presque toujours:  $\bar{X} = N(\mu, \sigma/\sqrt{n})$

- $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$
- $P(Z > z_\alpha) = \alpha$  (définition de  $z_\alpha$  «valeur critique»)
- $P(Z < -z_\alpha) = \alpha$  (symétrie de la normale)

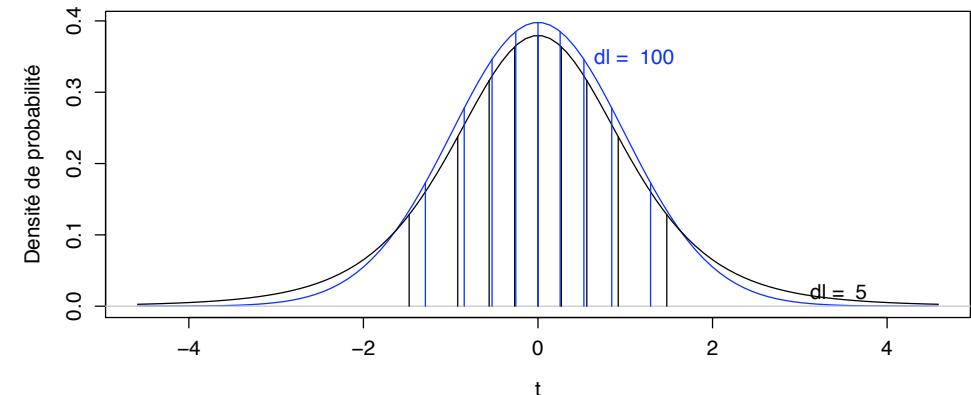
Distribution de la moyenne;  $\sigma_X$  inconnue

- $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \rightarrow N(0, 1)$
- $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{V/(n-1)}} = \frac{Z}{\sqrt{V/\nu}}$
- $V = \frac{(n-1)S^2}{\sigma^2}$ : loi du  $\chi^2$  à  $\nu = n - 1$  d.l.
- Condition: population normale
- $Z, V$  indépendantes
- $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ : loi de Student à  $\nu = n - 1$  d.l.
- $E[T] = 0$
- $\sigma_T^2 = \frac{\nu}{\nu-2} > 1$  (non définie pour  $\nu \leq 2$ )
- $P(T > t_\alpha) = \alpha$  (définition de  $t_\alpha$ , valeur critique)
- $P(T < -t_\alpha) = \alpha$  (symétrie de la loi  $t$ )
- $n \geq 30$ :  $s \rightarrow \sigma$  donc  $T \rightarrow Z$
- "Student": W.S. Gosset, 1908

## La distribution de Student

## La distribution de Student

## Distribution de Student



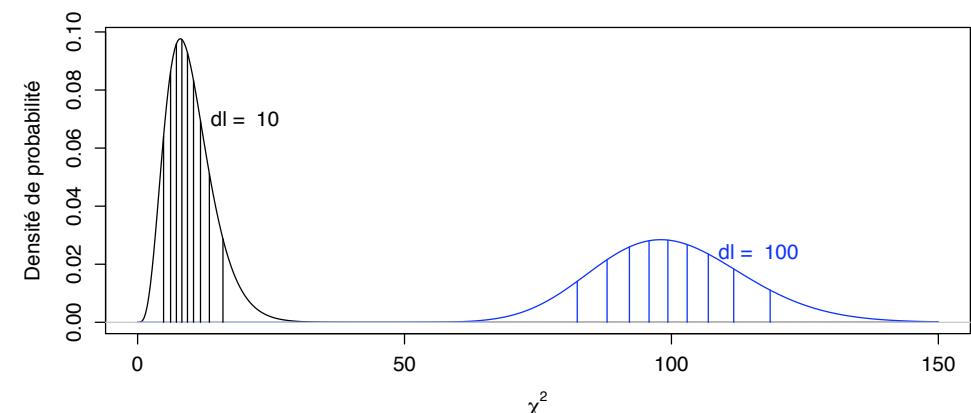
$$E[T] = 0 \quad , \quad \sigma_T^2 = \frac{\nu}{\nu-2} > 1 \text{ (non définie pour } \nu \leq 2\text{)}$$

## Distribution de la variance

- Échantillon aléatoire de taille  $n$ ; variance  $S^2$ 
  - Condition: population normale  $N(\mu, \sigma^2)$
  - $X^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$
  - $X^2$ : v.a. loi du  $\chi^2$  à  $\nu = n - 1$  degrés de liberté (d.l.)
  - $X^2 > 0$
  - $E[X^2] = n - 1 \rightarrow E[S^2] = \sigma^2$
  - $\sigma_{X^2}^2 = 2(n-1) \rightarrow \sigma_{S^2}^2 = 2\sigma^4/(n-1)$
  - $P(X^2 > \chi_\alpha^2(\nu)) = \alpha$  (définition de  $\chi_\alpha^2(\nu)$ , valeur critique)

Distribution du  $\chi^2$ 

## Distribution du Khi-deux



$$E[X^2] = n - 1 \quad , \quad \sigma_{X^2}^2 = 2(n - 1)$$

## Distribution de la proportion

- Proportion :  $\pi$ : proportion d'individus possédant un caractère qualitatif ( $\pi \neq 3.1415!$ )
- Échantillon aléatoire de taille  $n$ 
  - n.v.a.  $X_i ; x_i \in \{0, 1\}$  : Bernoulli indépendantes, de paramètre  $\pi$
  - $\sum_{i=1}^n X_i$ : nombre d'individus possédant le caractère (fréquence)
  - $\hat{P} = \frac{1}{n} \sum_{i=1}^n X_i$ : proportion d'individus (fréquence relative)
- Conditions:
  - $n > 30$  (grand échantillon: théorème limite central)
  - $n\hat{p} \geq 5$  (fréquence de présence du caractère)
  - $n(1 - \hat{p}) = n - n\hat{p} \geq 5$  (fréquence d'absence du caractère)
  - ni  $\hat{p} \approx 0$ , ni  $\hat{p} \approx 1$
- Distribution:
  - $\mu_{\hat{P}} = (n\mu_X)/n = \mu_X = \pi$  ,  $\sigma_{\hat{P}}^2 \stackrel{\text{ind}}{=} (n\sigma_X^2)/n^2 = \pi(1 - \pi)/n$
  - $\hat{P}$ : normale  $N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \rightarrow Z$ : normale  $N(0, 1)$

## Théorie d'échantillonage – deux échantillons

- Conditions:  $\sigma_1, \sigma_2$  connus et
  - populations normales  $N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)$  ou
  - $n_1 > 30$  et  $n_2 > 30$ , ou
  - populations «presque» normales
- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$ ; moyennes  $\bar{X}_1, \bar{X}_2$ 
  - $\bar{X}_1 - \bar{X}_2$ : normale
  - $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma_{\bar{X}_1 - \bar{X}_2}^2 \stackrel{\text{ind}}{=} \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- D'autres cas à examiner ultérieurement...

## Distribution du rapport des variances

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Provenant de populations normales de variances  $\sigma_1^2, \sigma_2^2$
- Variances des échantillons:  $S_1^2, S_2^2$
- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\nu_1/\nu_1}{\nu_2/\nu_2}$
- $V_i = \frac{(n_i - 1)S_i^2}{\sigma_i^2}$ : v.a. indépendantes, loi du  $\chi^2$  à  $\nu_i = n_i - 1$  d.l.
- $F$ : loi de Fisher (1924) - Snedecor (1934) avec  $\nu_1$  et  $\nu_2$  d.l.
- $F \geq 0$
- $E[F] = \frac{\nu_2}{\nu_2 - 2}$  ( $\nu_2 > 2$ )
- $\sigma_F^2 = \frac{\nu_2^2(2\nu_1 + 2\nu_2 - 4)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}$  ( $\nu_2 > 4$ )
- $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$  (définition de  $f_\alpha(\nu_1, \nu_2)$ , v.c.)
- $f_\alpha(\nu_1, \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2, \nu_1)}$  (propriété de la loi  $F$ )

Démonstration de  $f_\alpha(\nu_1, \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2, \nu_1)}$

On considère:  $\mathbb{P}(F > c) = \alpha$  où  $c$ , par définition, vaut  $f_\alpha(\nu_1, \nu_2)$ . Par définition de  $F$ :

$$\mathbb{P}\left(\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} > c\right) = \alpha$$

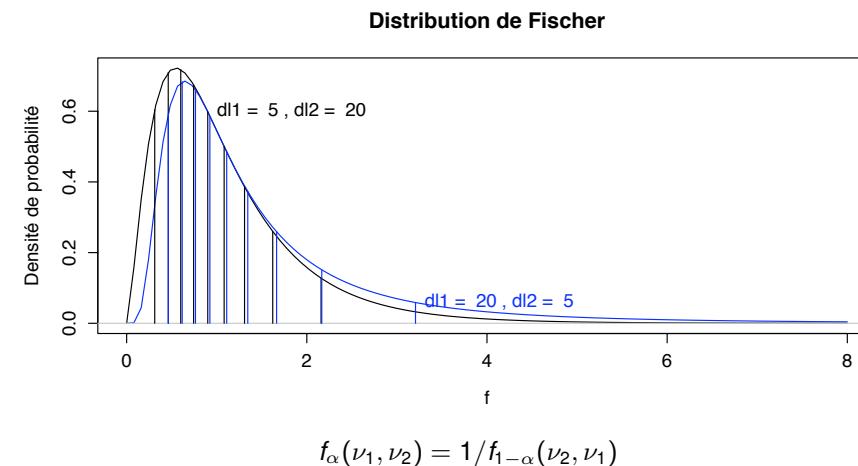
En inversant l'inégalité on obtient :

$$\mathbb{P}\left(\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} < \frac{1}{c}\right) = \alpha \text{ et } 1 - \mathbb{P}\left(\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} < \frac{1}{c}\right) = 1 - \alpha$$

Il s'en suit :

$$\mathbb{P}\left(\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} > \frac{1}{c}\right) = 1 - \alpha$$

Et on reconnaît que  $\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$  suit une loi de Fisher de paramètres  $\nu_2, \nu_1$  et donc  $\frac{1}{c} = f_{1-\alpha}(\nu_2, \nu_1)$ . CQFD.

**Distribution de Fisher****DEFINITION : Objectif principal :**

obtenir, à partir de mesures sur une *partie* de la population (échantillon), des informations (de caractère *probabiliste*) sur la *totalité* de celle-ci.

**Inférence :**

On prélève un **échantillon** de la population, on en *déduit* (ou encore on *infère*) des caractéristiques de la **population**.

**Objectif de la Statistique Inférentielle**

Théorie d'échantillonnage: Population → Échantillon  
Statistique inférentielle: Échantillon → Population

Échantillon		Population $p_X(x)$
v.a.	valeur	paramètre
une population		
$\bar{X}$	$m = \bar{x}$	$\mu_X = E[X]$
$S^2$	$s^2$	$\sigma_X^2 = \text{var}[X]$
$\hat{P}$	$\hat{p}$	$\pi$
deux populations		
$\bar{X}_2 - \bar{X}_1$	$m_2 - m_1 = \bar{x}_2 - \bar{x}_1$	$\mu_2 - \mu_1$
$S_2^2 / S_1^2$	$(s_2 / s_1)^2$	$(\sigma_2 / \sigma_1)^2$
$\hat{P}_2 - \hat{P}_1$	$\hat{p}_2 - \hat{p}_1$	$\pi_2 - \pi_1$

- Estimer les paramètres de la population
- Calculer des intervalles de confiance
- Formuler des hypothèses et les tester

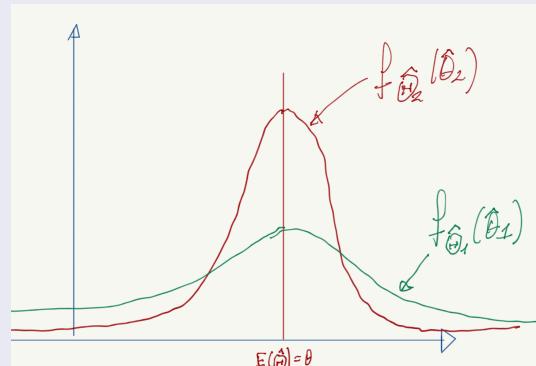
**DEFINITION : Un estimateur ponctuel**

est une statistique qui donne une valeur (unique) estimée de la grandeur recherchée.

- Paramètre à estimer:  $\theta$
- Estimateur: v.a.  $\hat{\Theta}$
- Estimateur non biaisé:  $E[\hat{\Theta}] = \theta$
- Biais =  $E[\hat{\Theta}] - \theta$
- Un bon Estimateur: sans biais; de faible variance
- Estimateur efficace: minimise l'erreur quadratique moyenne  
 $E[(\hat{\Theta} - \theta)^2] = \sigma_{\hat{\Theta}}^2 + (\text{biais})^2$
- Estimateur convergent:  $n \rightarrow \infty$  :  $E[\hat{\Theta}] = \theta$  et  $\text{var}[\hat{\Theta}] = 0$

## DEFINITION : MVUE Estimateur Non Biaisé à Variance Minimale

Soient tous les estimateurs non biaisés de  $\theta$   
L'estimateur MVUE (Minimum Variance Unbiased Estimator) est celui qui a la variance la plus petite.



Ici :  $\hat{\theta}_2$  a une variance plus faible que  $\hat{\theta}_1$ . On préférera donc  $\hat{\theta}_2$ .

## Estimateur au Maximum de Vraisemblance

Développé par Sir R.A. Fisher, statisticien Britannique, autour de 1920, l'estimateur au Maximum de Vraisemblance (MLE : Maximum Likelihood Estimator) est un des meilleurs estimateurs ponctuels.

### DEFINITION : Vraisemblance

Soit  $X$  une v.a. de densité de probabilité  $f_X(x; \theta)$ , où  $\theta$  est un paramètre inconnu. Soient  $x_1, x_2, \dots, x_n$  les valeurs observées d'un échantillon aléatoire de taille  $n$ , la **Vraisemblance** de l'échantillon vaut :

$$L(\theta) = f_X(x_1; \theta) \times f_X(x_2; \theta) \times \dots \times f_X(x_n; \theta)$$

Notez que  $L(\theta)$  est une fonction du seul paramètre  $\theta$ .

Note : on a besoin de la densité de probabilité de  $X$ , c'est le **MODELE**.

Exemple :  $X \sim \mathcal{N}(\mu, \sigma^2)$  avec  $\mu, \sigma^2$  les paramètres inconnus.

On a alors que les  $X_i \sim \mathcal{N}(\mu, \sigma^2)$  et sont indépendantes.

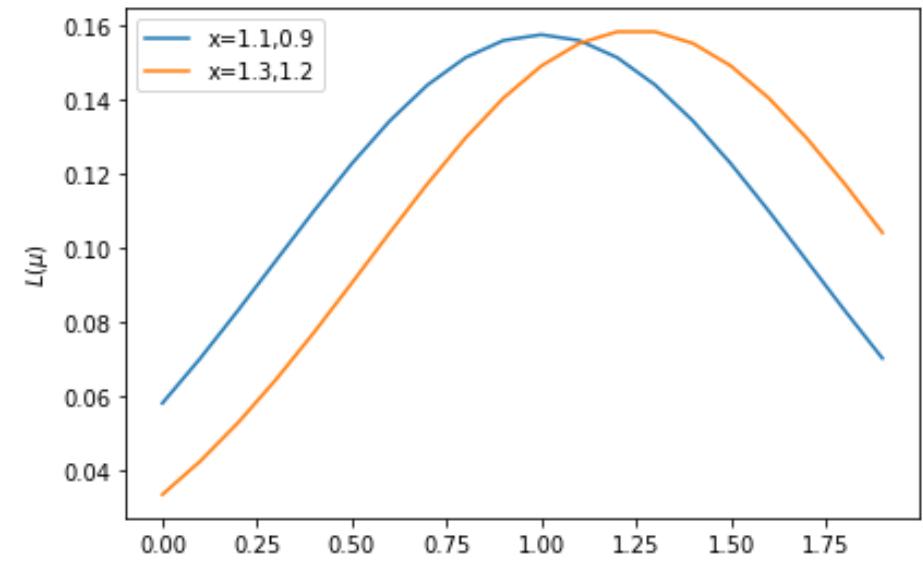
$$\text{On a donc } L(\mu, \sigma^2) = f_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

La moyenne de l'échantillon est un estimateur MVUE de  $\mu$

- Ce cours ne donne pas de méthode pour trouver un MVUE (mais ça existe).
- Soient  $X_1, X_2, \dots, X_n$  un échantillon aléatoire tiré d'une population normale de moyenne  $\mu$  et de variance  $\sigma^2$ ; alors la moyenne de l'échantillon  $\bar{X}$  est un MVUE de  $\mu$ .
- Même si on ne connaît pas de MVUE, on peut utiliser ce principe. Par exemple  $X$  et  $Y$  sont deux estimateurs non biaisés de  $\mu$ , mais  $\sigma_X^2 = \frac{\sigma^2}{n} < \sigma_Y^2$ , et donc on préférera  $\sigma_X^2$ .

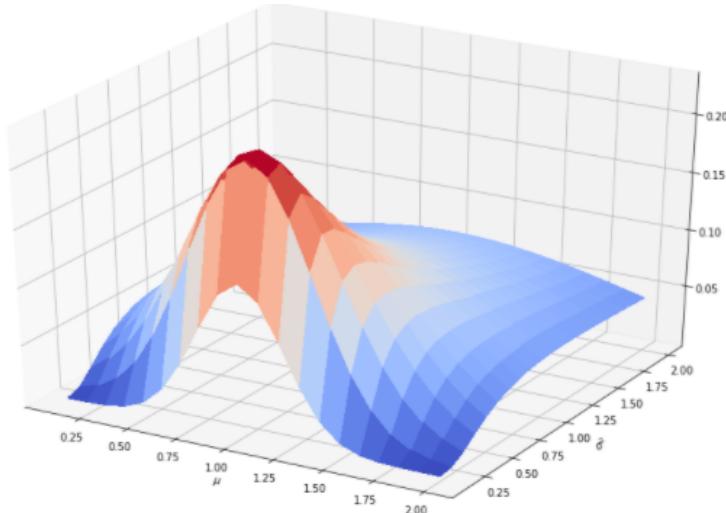
## Estimateur au Maximum de Vraisemblance

Exemple pour  $L(\mu)$  avec  $x = [1.1, 0.9]$  et  $x = [1.3, 1.2]$



## Estimateur au Maximum de Vraisemblance

Exemple pour  $L(\mu, \sigma^2)$  avec  $x = [1.5, 0.5]$



## Estimateur Maximum de Vraisemblance pour un modèle normal

Pour un modèle normal, on a

$$L(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2}$$

Soit, en prenant le logarithme :

$$\begin{aligned} \log L(\mu, \sigma^2) &= -n/2 \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \\ \frac{\partial \log L(\mu, \sigma^2)}{\partial \mu} &= \frac{\sum_{i=1}^n (x_i - \mu)}{\sigma^2} \end{aligned}$$

Et la valeur de  $\mu$  qui annule cette dérivée partielle vaut

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$ , qui est l'estimée ML de  $\mu$ .

Dans ce cas-ci, l'estimateur ML de  $\mu$  vaut  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$ , et est donc également une variable aléatoire !

## Estimateur au Maximum de Vraisemblance

## DEFINITION : Estimée au maximum de Vraisemblance

Soit  $X$  une v.a. de densité de probabilité  $f_X(x; \theta)$ , où  $\theta$  est un paramètre inconnu. Soient  $x_1, x_2, \dots, x_n$  les valeurs observées d'un échantillon aléatoire de taille  $n$ , l'estimée au maximum de vraisemblance de l'échantillon vaut :

$$\hat{\theta} = \arg \max_{\theta} (L(\theta))$$

## DEFINITION : L'estimateur au maximum de Vraisemblance

Soit  $X$  une v.a. de densité de probabilité  $f_X(x; \theta)$ , où  $\theta$  est un paramètre inconnu. Soit  $X = [X_1, X_2, \dots, X_n]$  un échantillon aléatoire de taille  $n$ , l'Estimateur au maximum de vraisemblance de l'échantillon vaut :

$$\hat{\Theta} = \arg \max_{\theta} (L(X; \theta))$$

## Estimateur Maximum de Vraisemblance pour un modèle normal

## L'Estimateur au Maximum de Vraisemblance pour un modèle normal

$$\frac{\partial \log L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4}$$

Et donc, on a les deux estimateurs ML qui valent :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

## Propriétés de l'Estimateur au Maximum de Vraisemblance

## Propriétés de l'Estimateur au Maximum de Vraisemblance

### DEFINITION : Régularité d'une pdf

Une pdf  $f_X(x; \theta)$  est régulière si et seulement si :

- 1 Le domaine de variation de  $X$ ,  $R_X = x : f_X(x; \theta) > 0$  ne dépend pas de  $\theta$ .
- 2  $f_X(x; \theta)$  est au moins trois fois continument dérivable.
- 3 La vraie valeur de  $\theta$  est dans un ensemble compact  $\Theta$ .

Sous les conditions de régularité :

- 1 L'estimateur au maximum de vraisemblance est **consistant**
- 2 L'estimateur au maximum de vraisemblance est **asymptotiquement normal**
- 3 L'estimateur au maximum de vraisemblance est **Efficace**
- 4 L'estimateur au maximum de vraisemblance est **Invariant**

## l'Estimateur au Maximum de Vraisemblance est consistant

Sous les conditions de régularité,

$$\hat{\Theta} \xrightarrow[n \rightarrow \infty]{P} \theta_o$$

où  $\theta_o$  est la vraie valeur du paramètre.

Note :  $X_n \xrightarrow[n \rightarrow \infty]{P} X$  signifie que la suite  $X_n$  converge vers  $X$  en probabilité :

$$\forall \epsilon > 0, \lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

## L'Estimateur au Maximum de Vraisemblance est asymptotiquement normal

Sous les conditions de régularités,

L'estimateur ML est asymptotiquement Normal

pour  $n$  grand,

$$\hat{\Theta} \xrightarrow{asy} \mathcal{N}(\theta_o, I_N^{-1}(\theta_o))$$

où  $\theta_o$  est la vraie valeur du paramètre et  $I_N^{-1}(\theta_o)$  est la matrice d'information de Fisher associée à l'échantillon (qui sort du cadre de ce cours).

## L'Estimateur ML non biaisé est asymptotiquement efficace

- 1 Si  $\hat{\Theta}_{UB}$  est un estimateur (linéaire ou non linéaire) non biaisé, le meilleur estimateur est celui dont la variance est la plus petite. L'indice *UB* signifie UnBiased (non biaisé).
- 2 On peut trouver une borne inférieure de la variance de tout estimateur  $\hat{\Theta}_{UB}$ , appelé borne de Cramer-Rao.

### Efficacité de l'estimateur ML

Sous les conditions de régularité, l'estimateur ML est asymptotiquement efficace et atteint la borne de Cramer-Rao :

$$\text{var}[\hat{\Theta}] = I_N^{-1}(\theta)$$

### Estimateur par intervalle de confiance

### DEFINITION :Estimateur par intervalle de confiance

une statistique qui donne un intervalle dans lequel la grandeur recherchée se trouve, avec un indice de confiance. Cet indice de confiance donne le niveau de confiance avec lequel on peut "croire" que la grandeur recherchée se trouve à l'intérieur de cet intervalle.

- v.a.  $\hat{\Theta}_L, \hat{\Theta}_H$ : estimateurs ponctuels
- $P(\hat{\Theta}_L < \theta < \hat{\Theta}_H) = 1 - \alpha$
- $\hat{\theta}_L < \theta < \hat{\theta}_H$ : intervalle de confiance
- $1 - \alpha$ : niveau de confiance (en anglais : confidence coefficient).

## L'Estimateur ML non biaisé est invariant

### Invariance de l'estimateur au maximum de vraisemblance

Soit  $\hat{\Theta}_1, \dots, \hat{\Theta}_k$  les estimateurs ML de  $\theta_1, \dots, \theta_k$ . On considère une fonction  $h(\theta_1, \dots, \theta_k)$  de ces paramètres. Alors :

$$\hat{h}(\theta_1, \dots, \theta_k) = h(\hat{\Theta}_1, \dots, \hat{\Theta}_k)$$

L'estimateur ML de la fonction des paramètres est la fonction des estimateurs ML des paramètres.

Par exemple, l'estimateur ML de l'écart-type d'un modèle Gaussien vaut :

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^{1/2}$$

### Estimateur par intervalle de confiance

### Propriétés de l'estimateur ML

#### Estimation de la moyenne d'une v.a. normale

### Propriétés et intervalle de confiance : Moyenne

- $X$ , normale de moyenne  $\mu$  inconnue et de variance  $\sigma^2$  connue
- $\bar{X}$ : normale  $N(\mu, \sigma^2/n)$
- $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ : normale  $N(0, 1)$
- $\bar{X}$  estimateur non biaisé et convergent de  $\mu$
- $P(Z > z_{\alpha/2}) = \alpha/2$  (définition de  $z_{\alpha/2}$ )
- $P(Z < -z_{\alpha/2}) = \alpha/2$  (symétrie de la normale)
- $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$
- $P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$
- $P(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $P(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $\hat{\Theta}_L = \bar{X} - z_{\alpha/2} \sigma_{\bar{X}}, \hat{\Theta}_H = \bar{X} + z_{\alpha/2} \sigma_{\bar{X}}$
- $1 - \alpha = 0.95, z_{\alpha/2} = 1.96$
- $1 - \alpha = 0.99, z_{\alpha/2} = 2.56$

## Taille de l'échantillon

- $X$ , normale de moyenne  $\mu$  inconnue et de variance  $\sigma^2$  connue
- $P(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $P(|\bar{X} - \mu| < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$
- $e = |\bar{X} - \mu|$ : erreur
- $e_{\max} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ : marge d'erreur à  $1 - \alpha$
- $n_{\min} = \left( \frac{z_{\alpha/2}\sigma}{e_{\max}} \right)^2$ : taille d'échantillon minimale
- $\bar{X} - e_{\max} < \mu < \bar{X} + e_{\max}$  à  $1 - \alpha$
- Cas particulier: échantillonnage d'une population finie, sans remplacement
  - Population de taille  $N$
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \stackrel{N \gg 1}{\approx} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N}} = \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
  - $n_{\min} = \frac{Nz_{\alpha/2}^2 \sigma^2}{Ne_{\max}^2 + z_{\alpha/2}^2 \sigma^2}$ : taille d'échantillon minimale

## Note sur l'Estimation de la moyenne pour des petits échantillons

- Si la population est normale,  $n < 30$  et  $\sigma$  connu :  $\bar{X}$  est normale.
- Si la population est normale,  $n < 30$  et  $\sigma$  inconnu :  $(\bar{X} - \mu)/(S/\sqrt{n})$  suit une loi du Student !
- Si la population n'est PAS NORMALE et  $n < 30$ , ON NE PEUT RIEN DIRE (dans ce cours).
- Si la population n'est PAS NORMALE et  $n < 30$  ... il y a beaucoup de publications scientifiques qui précisent ....

## Estimation de la moyenne, variance inconnue

- $X$ , normale de moyenne  $\mu$  inconnue et de variance  $\sigma^2$  inconnue (et donc estimée par  $S^2$ ).
- $T = (\bar{X} - \mu)/(S/\sqrt{n})$ : Student à  $n - 1$  d.l.
- $P(T > t_{\alpha/2}) = \alpha/2$  (définition de  $t_{\alpha/2}$ )
- $P(T < -t_{\alpha/2}) = \alpha/2$  (symétrie de la loi t)
- $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha$
- $P(-t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}) = 1 - \alpha$
- $P(-t_{\alpha/2} \frac{S}{\sqrt{n}} < \bar{X} - \mu < t_{\alpha/2} \frac{S}{\sqrt{n}}) = 1 - \alpha$
- $P(\bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}) = 1 - \alpha$
- $\hat{\Theta}_L = \bar{X} - t_{\alpha/2} \frac{S}{\sqrt{n}}$ ,  $\hat{\Theta}_H = \bar{X} + t_{\alpha/2} \frac{S}{\sqrt{n}}$
- $1 - \alpha = 0.95$ ,  $t_{\alpha/2} = 2.05$
- $1 - \alpha = 0.99$ ,  $t_{\alpha/2} = 2.76$
- Rappel:  $n \geq 30$ ,  $T \rightarrow Z$
- $T$ : petits échantillons!

## Estimation de la variance (un échantillon)

- Condition: population normale  $N(\mu, \sigma^2)$
- $X^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$
- $X^2$ : v.a. loi du  $\chi^2$  à  $\nu = n - 1$  degrés de liberté (d.l.)
- $P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha$
- $P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1 - \alpha$
- $P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha$
- $P\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}\right) = 1 - \alpha$
- Intervalle de confiance:  

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}}$$
 à un niveau de confiance de  $(1 - \alpha)100\%$

## Estimation de la proportion (= moyenne)

## ■ Caractère qualitatif

- Proportion:  $\hat{P} = \frac{1}{n} \sum_{i=1}^n X_i$
- $n > 30, n\hat{p} \geq 5, n(1-\hat{p}) \geq 5$ , ni  $\hat{p} \approx 0$ , ni  $\hat{p} \approx 1$
- $\hat{P} = N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

## ■ Les proportions (fréquences relatives) sont des moyennes!

■  $\bar{X} \rightarrow \hat{P}$ : remplacer

- $\mu \rightarrow \pi$
- $\sigma \rightarrow \sqrt{\pi(1-\pi)}$

## ■ Caractère qualitatif

- $P\left(\hat{P} - z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}} < \pi < \hat{P} + z_{\alpha/2} \sqrt{\frac{\pi(1-\pi)}{n}}\right) = 1 - \alpha$
- Intervalle de confiance à un niveau de confiance de  $(1 - \alpha)100\%$ :  

$$\hat{P} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}} < \pi < \hat{P} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}$$
- $n_{\min} = \left(\frac{z_{\alpha/2}}{\theta_{\max}}\right)^2 \hat{p}(1 - \hat{p})$ : taille d'échantillon minimale  
estimer  $\hat{p}$  (1er échantillonage,  $n \geq 30$ ) ou prendre  $\hat{p} = 0.5$  (pire scénario)

## Estimation du rapport des variances (deux échantillons)

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Provenant de populations normales de variances  $\sigma_1^2, \sigma_2^2$
- Variances des échantillons:  $S_1^2, S_2^2$
- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{V_1/\nu_1}{V_2/\nu_2}$
- $V_i = \frac{(n_i-1)S_i^2}{\sigma_i^2}$ : v.a. indépendantes, loi du  $\chi^2$  à  $\nu_i = n_i - 1$  d.l.
- $F$ : loi de Fisher - Snedecor avec  $\nu_1$  et  $\nu_2$  d.l.
- $P(f_{1-\alpha/2}(\nu_1, \nu_2) < F < f_{\alpha/2}(\nu_1, \nu_2)) = 1 - \alpha$
- $P\left(f_{1-\alpha/2}(\nu_1, \nu_2) < \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(\nu_1, \nu_2)\right) = 1 - \alpha$
- $P\left(\frac{S_1^2}{S_2^2} \frac{1}{f_{1-\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_1^2}{S_2^2} \frac{1}{f_{\alpha/2}(\nu_1, \nu_2)}\right) = 1 - \alpha$
- $P\left(\frac{S_1^2}{S_2^2} \frac{1}{f_{\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_1^2}{S_2^2} f_{\alpha/2}(\nu_2, \nu_1)\right) = 1 - \alpha$

## Tests d'hypothèses

- Hypothèse: énoncé concernant les caractéristiques d'une population
- Hypothèse nulle: fixer un paramètre  $\theta$  à une valeur particulière  $\theta_0$ 
  - $H_0: \theta = \theta_0$
- Hypothèse alternative (trois choix possibles)
  - $H_1: \theta \neq \theta_0$  (test bilatéral)
  - $H_1: \theta < \theta_0$  (test unilatéral)
  - $H_1: \theta > \theta_0$  (test unilatéral)
- Test: procédure suivie afin d'accepter/rejeter  $H_0$
- Rejet > Acceptation (non-rejet)
- En pratique: formuler  $H_0$  comme l'opposé de ce qu'on veut démontrer!

## Types et probabilités d'erreur

Types d'erreur		
décision \ état du monde	$H_0$ vraie	$H_1$ vraie
non-rejet de $H_0$	OK	Type II
rejet de $H_0$	Type I	OK

■  $P(\text{Type I}) = P(\text{rejet de } H_0 | H_0 \text{ vraie}) = \alpha$

■  $P(\text{Type II}) = P(\text{non-rejet de } H_0 | H_1 \text{ vraie}) = \beta$

Probabilités d'erreur		
décision \ état du monde	$H_0$ vraie	$H_1$ vraie
non-rejet de $H_0$	$1 - \alpha$	$\beta$
rejet de $H_0$	$\alpha$	$1 - \beta$

■  $\alpha$ : seuil de signification (calculé dans l'univers de  $H_0$ , ok)

■  $1 - \beta$ : puissance du test (calculée dans l'univers de  $H_1$ , ???)

■ Préciser  $H_1$ , ensuite calculer une valeur de  $\beta$  liée à cette  $H_1$

## Tests : la procédure à suivre

- 1 Formuler les hypothèses  $H_0$  et  $H_1$
- 2 Choisir le seuil de signification  $\alpha$  (typiquement 1% ou 5%)
- 3 Déterminer la statistique utilisée ainsi que sa distribution
- 4 Définir la région critique (région de rejet de  $H_0$ )
- 5 Adopter une règle de décision (à partir des valeurs critiques)
- 6 Prélever un échantillon et faire les calculs
- 7 Décider

## Test sur une moyenne

1  $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$  (test bilatéral)

2  $\alpha$  à définir

3 Statistique à utiliser:  $\bar{X}$ ; distribution:

$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$  si on connaît  $\sigma$  ou  $n$  grand (cas présenté dans la suite)

$T = (\bar{X} - \mu) / (S / \sqrt{n})$  si on ne connaît pas  $\sigma$  et  $n$  petit (population normale)

4  $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$

$P(\text{non-rejet de } H_0 | \mu = \mu_0) = 1 - \alpha$

$P(z_{1-\alpha/2} < Z < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$

$P(-z_{\alpha/2} < Z < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$

$P(-z_{\alpha/2} < (\bar{X} - \mu) / (\sigma / \sqrt{n}) < z_{\alpha/2} | \mu = \mu_0) = 1 - \alpha$

$P(-z_{\alpha/2} < (\bar{X} - \mu_0) / (\sigma / \sqrt{n}) < z_{\alpha/2}) = 1 - \alpha$

région critique :  $Z = (\bar{X} - \mu_0) / (\sigma / \sqrt{n}) < -z_{\alpha/2}$  et

$Z = (\bar{X} - \mu_0) / (\sigma / \sqrt{n}) > z_{\alpha/2}$

5 Règle de décision:

rejeter  $H_0$  si  $\bar{X} < \bar{X}_{c1} = \mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$  ou  $\bar{X} > \bar{X}_{c2} = \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

## Test Unilatéral (moyenne)

1  $H_0: \mu = \mu_0, H_1: \mu > \mu_0$  (test unilatéral)

2  $\alpha$  à définir

3 Statistique à utiliser:  $\bar{X}$ ; distribution:

$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$  si on connaît  $\sigma$  ou  $n$  grand (cas présenté dans la suite)

$T = (\bar{X} - \mu) / (S / \sqrt{n})$  si on ne connaît pas  $\sigma$  et  $n$  petit (population normale)

4  $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$

$P(\text{non-rejet de } H_0 | \mu = \mu_0) = 1 - \alpha$

$P(Z < z_\alpha | \mu = \mu_0) = 1 - \alpha$

$P((\bar{X} - \mu) / (\sigma / \sqrt{n}) < z_\alpha | \mu = \mu_0) = 1 - \alpha$

$P((\bar{X} - \mu_0) / (\sigma / \sqrt{n}) < z_\alpha) = 1 - \alpha$

région critique :  $Z = (\bar{X} - \mu_0) / (\sigma / \sqrt{n}) > z_\alpha$

5 Règle de décision:

rejeter  $H_0$  si  $\bar{X} > \bar{X}_c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

## Taille de l'échantillon

- $H_0: \mu = \mu_0, H_1: \mu > \mu_0$  (test unilatéral)
- $\alpha = P(\text{rejet de } H_0 | H_0 \text{ vraie}) = P(\text{rejet de } H_0 | \mu = \mu_0) = P(Z > z_\alpha | \mu = \mu_0) = P((\bar{X} - \mu)/(\sigma/\sqrt{n}) > z_\alpha | \mu = \mu_0) = P((\bar{X} - \mu_0)/(\sigma/\sqrt{n}) > z_\alpha)$
- Règle de décision: rejeter  $H_0$  si  $\bar{X} > \bar{x}_c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$
- $\beta = P(\text{rejet de } H_1 | H_1 \text{ vraie}) = P(\text{non-rejet de } H_0 | H_1 \text{ vraie}) = P(\bar{X} < \bar{x}_c | H_1 \text{ vraie})$
- Préciser  $H_1: \mu = \mu_0 + \delta$
- $\beta = P(\bar{X} < \bar{x}_c | \mu = \mu_0 + \delta) = P(Z < (\bar{x}_c - \mu)/(\sigma/\sqrt{n}) | \mu = \mu_0 + \delta) = P(Z < \frac{\bar{x}_c - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{\sigma/\sqrt{n}}) = P(Z < z_\alpha - \frac{\delta}{\sigma/\sqrt{n}})$
- $-z_\beta = z_\alpha - \frac{\delta}{\sigma/\sqrt{n}}$
- $n = (z_\alpha + z_\beta)^2 \frac{\sigma^2}{\delta^2}$

## Test bilatéral (Variance)

- $H_0: \sigma = \sigma_0, H_1: \sigma \neq \sigma_0$  (test bilatéral)
- $\alpha$  à définir
- Statistique à utiliser:  $S$ ; distribution:  $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$ , v.a. loi du  $\chi^2$  à  $\nu = n - 1$  degrés de liberté (population normale)
- $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$   
 $P(\text{non-rejet de } H_0 | \sigma = \sigma_0) = 1 - \alpha$   
 $P(\chi_{1-\alpha/2}^2 < \chi^2 < \chi_{\alpha/2}^2 | \sigma = \sigma_0) = 1 - \alpha$   
 $P\left(\chi_{1-\alpha/2}^2 < \frac{(n-1)S^2}{\sigma_0^2} < \chi_{\alpha/2}^2\right) = 1 - \alpha$   
 $P\left(\frac{\chi_{1-\alpha/2}^2 \sigma_0^2}{(n-1)} < S^2 < \frac{\chi_{\alpha/2}^2 \sigma_0^2}{(n-1)}\right) = 1 - \alpha$   
 région critique :  $\chi^2 < \chi_{1-\alpha/2}^2$  et  $\chi^2 > \chi_{\alpha/2}^2$
- Règle de décision:  
 rejeter  $H_0$  si  $s^2 < s_{c1}^2 = \chi_{1-\alpha/2}^2 \sigma_0^2 / (n-1)$  ou  $s^2 > s_{c2}^2 = \chi_{\alpha/2}^2 \sigma_0^2 / (n-1)$

## Test unilatéral (variance)

- $H_0: \sigma = \sigma_0, H_1: \sigma < \sigma_0$  (test unilatéral)
- $\alpha$  à définir
- Statistique à utiliser:  $S$ ; distribution:  $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$ , v.a. loi du  $\chi^2$  à  $\nu = n - 1$  degrés de liberté (population normale)
- $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$   
 $P(\text{non-rejet de } H_0 | \sigma = \sigma_0) = 1 - \alpha$   
 $P(\chi_{1-\alpha}^2 < \chi^2 | \sigma = \sigma_0) = 1 - \alpha$   
 $P\left(\chi_{1-\alpha}^2 < \frac{(n-1)S^2}{\sigma_0^2}\right) = 1 - \alpha$   
 $P\left(\frac{\chi_{1-\alpha}^2 \sigma_0^2}{(n-1)} < S^2\right) = 1 - \alpha$   
 région critique :  $\chi^2 < \chi_{1-\alpha}^2$
- Règle de décision:  
 rejeter  $H_0$  si  $s^2 < s_c^2 = \chi_{1-\alpha}^2 \sigma_0^2 / (n-1)$

- $H_0: \pi = \pi_0, H_1: \pi \neq \pi_0$  (test bilatéral)
- $\alpha$  à définir
- Statistique à utiliser:  $\hat{P}$ ; distribution:  $Z = (\hat{P} - \pi) / (\sqrt{\pi(1 - \pi)} / \sqrt{n})$
- $P(\text{non-rejet de } H_0 | H_0 \text{ vraie}) = 1 - \alpha$   
 $P(\text{non-rejet de } H_0 | \pi = \pi_0) = 1 - \alpha$   
 $P(-z_{\alpha/2} < (\hat{P} - \pi_0) / (\sqrt{\pi_0(1 - \pi_0)} / \sqrt{n}) < z_{\alpha/2}) = 1 - \alpha$   
 région critique :  $Z < -z_{\alpha/2}$  et  $Z > z_{\alpha/2}$
- Règle de décision:  
 rejeter  $H_0$  si  $\hat{p} < \hat{p}_{c1} = \pi_0 - z_{\alpha/2} \frac{\sqrt{\pi_0(1 - \pi_0)}}{\sqrt{n}}$  ou  
 $\hat{p} > \hat{p}_{c1} = \pi_0 + z_{\alpha/2} \frac{\sqrt{\pi_0(1 - \pi_0)}}{\sqrt{n}}$
- $H_0: \pi = \pi_0, H_1: \pi > \pi_0$  (test unilatéral)
- ...
- Règle de décision: rejeter  $H_0$  si  $z > z_\alpha$   
 c.à.d.  $\hat{p} > \hat{p}_c = \pi_0 + z_\alpha \frac{\sqrt{\pi_0(1 - \pi_0)}}{\sqrt{n}}$

Paramètre $\theta$	$\mu$		
Population	— connu	— connu	— connu
Écart-type $\sigma$	connu	inconnu	inconnu
Échantillon	—	$n > 30$	$n > 30$
Statistique $\hat{\Theta}$	$\bar{X}$		
St. normalisée	$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$	$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$
Distribution	$N(0, 1)$		Student ( $\nu$ )
D.L.	—		$n - 1$
Mesure $\hat{\theta}$	$\bar{X}$		

Paramètre $\theta$	$\pi$	$\sigma^2$
Population	—	—
Écart-type $\sigma$	—	—
Échantillon	$n > 30$ $n\hat{p} \geq 5$ , $n(1 - \hat{p}) \geq 5$ , ni $\hat{p} \approx 0$ , ni $\hat{p} \approx 1$ .	—
Statistique $\hat{\Theta}$	$\hat{P}$	$S^2$
St. normalisée	$Z = \frac{\hat{P} - \pi}{\sqrt{\pi(1-\pi)/n}}$	$X^2 = \frac{(n-1)S^2}{\sigma^2}$
Distribution	$N(0, 1)$	khi-deux ( $\nu$ )
D.L.	—	$n - 1$
Mesure $\hat{\theta}$	$\hat{p}$	$s^2$

Stat. norm.	Intervalle de confiance	Test d'hypothèse $H_0 : \theta = \theta_0$		
		$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$
$Z$	$-Z_{\alpha/2} < Z < Z_{\alpha/2}$	$Z < -Z_{\alpha/2}$ ou $Z > Z_{\alpha/2}$	$Z < -Z_\alpha$	$Z > Z_\alpha$
$T$	$-t_{\alpha/2} < t < t_{\alpha/2}$	$t < -t_{\alpha/2}$ ou $t > t_{\alpha/2}$	$t < -t_\alpha$	$t > t_\alpha$
$\chi^2$	$\chi^2_{1-\frac{\alpha}{2}} < \chi^2 < \chi^2_{\frac{\alpha}{2}}$	$\chi^2 < \chi^2_{1-\frac{\alpha}{2}}$ ou $\chi^2 > \chi^2_{\frac{\alpha}{2}}$	$\chi^2 < \chi^2_{1-\alpha}$	$\chi^2 > \chi^2_\alpha$
	mettre sous la forme: $\theta_L < \theta < \theta_H$	«entrer dans le monde de $H_0$ »: $\theta = \theta_0$ , calculer $z, t, \chi^2$ à partir des mesures; décisions de <i>rejet</i> de $H_0$		

- Intervalle de confiance: niveau de confiance  $1 - \alpha$
- Tests d'hypothèse: seuil de signification  $\alpha$
- Voir tableaux unifiés en annexe.

## Différence de moyennes : variances connues, populations “normales”

- Conditions:  $\sigma_1, \sigma_2$  connus et
  - populations normales  $N(\mu_1, \sigma_1), N(\mu_2, \sigma_2)$  ou
  - $n_1 > 30$  et  $n_2 > 30$ , ou
  - populations «presque» normales
- Échantillons aléatoires et ind. de tailles  $n_1, n_2$ ; moyennes  $\bar{X}_1, \bar{X}_2$ 
  - $\bar{X}_1 - \bar{X}_2$ : normale :  $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_{\bar{X}_1} - \mu_{\bar{X}_2} = \mu_1 - \mu_2$
  - $\sigma_{\bar{X}_1 - \bar{X}_2}^2 = \sigma_{\bar{X}_1}^2 + \sigma_{\bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
- $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightarrow N(0, 1)$
- Intervalle de confiance :
 
$$(\bar{X}_1 - \bar{X}_2) - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
- Test d'hypothèse :
  1.  $H_0: \mu_1 - \mu_2 = d_0, H_1: \mu_1 - \mu_2 \neq d_0$  (test bilatéral)
  5. Règle de décision: rejeter  $H_0$  si  $z < -Z_{\alpha/2}$  ou  $z > Z_{\alpha/2}$
$$(\bar{X}_1 - \bar{X}_2) < (\bar{X}_1 - \bar{X}_2)_{c1} = d_0 - Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \text{ ou}$$

$$(\bar{X}_1 - \bar{X}_2) > (\bar{X}_1 - \bar{X}_2)_{c2} = d_0 + Z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

## variances inconnues, populations normales et grands échantillons

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Populations normales **et** grands échantillons ( $n_1 > 30, n_2 > 30$ )
- $\sigma_1, \sigma_2$ : inconnus
- $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow \approx N(0, 1)$
- Équivalent de  $T \rightarrow Z$  pour grands échantillons
- Intervalle de confiance :
 
$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
- Test d'hypothèse :
  1.  $H_0: \mu_1 - \mu_2 = d_0, H_1: \mu_1 - \mu_2 > d_0$  (test unilatéral)
  5. Règle de décision: rejeter  $H_0$  si  $z > z_\alpha$
$$(\bar{x}_1 - \bar{x}_2) > (\bar{x}_1 - \bar{x}_2)_c = d_0 + z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

## Variances inconnues mais égales, petits échantillons

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Populations normales **et** petits échantillons ( $n_1 < 30$  ou  $n_2 < 30$ )
- $\sigma_1, \sigma_2$ : inconnus mais  $\sigma_1 = \sigma_2$  (à tester)
- $T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_c^2}{n_1} + \frac{s_c^2}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightarrow$  Student
- Variance commune :  $S_c^2 = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2}{(n_1-1)+(n_2-1)} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$
- $T$ : Student à  $(n_1 + n_2 - 2)$  d.l.
- Intervalle de confiance :
 
$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$
- Test d'hypothèse: ...
- À propos des conditions:
  - $\sigma_1 \approx \sigma_2$  ou populations  $\approx$  normales: OK
  - $\sigma_1 \neq \sigma_2$  et normales: OK si  $n_1 = n_2$

## Variances inconnues et différentes - petits échantillons

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Populations normales **et** petits échantillons ( $n_1 < 30$  ou  $n_2 < 30$ )
- $\sigma_1, \sigma_2$ : inconnus et  $\sigma_1 \neq \sigma_2$  (à tester)
- $T = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \rightarrow$  Student à  $\nu$  d.l. ;  $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$
- Arrondir  $\nu$  au nombre entier *inférieur*.
- Intervalle de confiance :
 
$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$
- Test d'hypothèse :
  1.  $H_0: \mu_1 - \mu_2 = d_0, H_1: \mu_1 - \mu_2 < d_0$  (test unilatéral)
  5. Règle de décision: rejeter  $H_0$  si  $t < t_\alpha$
$$(\bar{x}_1 - \bar{x}_2) < (\bar{x}_1 - \bar{x}_2)_c = d_0 - t_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

## Echantillons appariés

- Échantillons aléatoires et **appariés** de tailles  $n_1 = n_2 = n$
- Appariés: «avant / après»
- Population: nouvelle v.a.  $D = X_1 - X_2$  ( $\mu_D, \sigma_D$ )
- Échantillon: calculer  $d_i = x_{1i} - x_{2i}$ ; oublier  $X_1, X_2$ !
- Population normale ou grands échantillons ( $n > 30$ ),  $\sigma_D$  connu :
 
$$Z = \frac{\bar{D} - \mu_D}{\sigma_D/\sqrt{n}} \rightarrow N(0, 1)$$
- Population normale et petits échantillons ( $n < 30$ ),  $\sigma_D$  inconnu :
 
$$T = \frac{\bar{D} - \mu_D}{s_D/\sqrt{n}} \text{ à } (n-1) \text{ d.l.}$$
- Intervalle de confiance :  $\bar{D} - t_{\alpha/2} \frac{s_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{\alpha/2} \frac{s_D}{\sqrt{n}}$
- Test d'hypothèse : ...
- Échantillons appariés : un seul nouvel échantillon!

## Distribution de la différence des proportions

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Grands échantillons ( $n_1 > 30, n_2 > 30$ )
- Proportions:  $\hat{P}_i = N(\pi_i, \sqrt{\pi_i(1-\pi_i)}/\sqrt{n_i})$
- $Z = \frac{(\hat{P}_1 - \hat{P}_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \rightarrow N(0, 1)$
- Intervalle de confiance :  $(\hat{P}_1 - \hat{P}_2) - Z_{\alpha/2} \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} < \pi_1 - \pi_2 < (\hat{P}_1 - \hat{P}_2) + Z_{\alpha/2} \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$ ; remplacer  $\pi_i(1-\pi_i) \rightarrow \hat{P}_i(1-\hat{P}_i)$
- Test d'hypothèse :
  - 1.  $H_0: \pi_1 - \pi_2 = d_0$  ( $\pi_1 = \pi_2 + d_0$ ),  $H_1: \pi_1 - \pi_2 > d_0$  (test unilatéral)
  - 5. Règle de décision: rejeter  $H_0$  si  $z > z_\alpha$
$$(\hat{P}_1 - \hat{P}_2) > (\hat{P}_1 - \hat{P}_2)_c = d_0 + z_\alpha \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$$

Si  $d_0 = 0$ ,  $\pi_1 = \pi_2$ : remplacer  $\pi_j \rightarrow \hat{P} = \frac{\sum_{i=1}^{n_1} x_{1i} + \sum_{i=1}^{n_2} x_{2i}}{n_1 + n_2} = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2}$

Si  $d_0 \neq 0$ : remplacer  $\pi_j \rightarrow \hat{P}_j$

- Échantillons aléatoires et indépendants de tailles  $n_1, n_2$
- Provenant de populations normales de variances  $\sigma_1^2, \sigma_2^2$
- Variances des échantillons:  $S_1^2, S_2^2$
- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{V_1/\nu_1}{V_2/\nu_2}; V_i = \frac{(n_i-1)S_i^2}{\sigma_i^2}$ : v.a. ind., loi du  $\chi^2$  à  $\nu_i = n_i - 1$  d.l.
- $F$ : loi de Fisher (1924) - Snedecor (1934) avec  $\nu_1$  et  $\nu_2$  d.l.
- $F \geq 0$
- $P(F > f_\alpha(\nu_1, \nu_2)) = \alpha$  (définition de  $f_\alpha(\nu_1, \nu_2)$ )
- $f_\alpha(\nu_1, \nu_2) = \frac{1}{f_{1-\alpha}(\nu_2, \nu_1)}$  (propriété de la loi  $F$ )
- $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2}$
- Intervalle de confiance (niveau de confiance  $1 - \alpha$ ) :
  - $f_{1-\alpha/2}(\nu_1, \nu_2) < f < f_{\alpha/2}(\nu_1, \nu_2)$
  - $\frac{s_1^2}{s_2^2} \frac{1}{f_{\alpha/2}(\nu_1, \nu_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \frac{1}{f_{1-\alpha/2}(\nu_1, \nu_2)}$
- Test d'hypothèse  $H_0: \sigma_1 = \sigma_2$
- Règle de décision: rejeter  $H_0$  si
  - $H_1: \sigma_1 \neq \sigma_2$   
 $f < f_{1-\alpha/2}$  ou  $f > f_{\alpha/2}$  c-à-d  $s_1^2/s_2^2 < f_{1-\alpha/2}$  ou  $s_1^2/s_2^2 > f_{\alpha/2}$
  - $H_1: \sigma_1 > \sigma_2$   
 $f > f_\alpha$  c-à-d  $s_1^2/s_2^2 > f_\alpha$
  - $H_1: \sigma_1 < \sigma_2$   
 $f < f_{1-\alpha/2}$  c-à-d  $s_1^2/s_2^2 < f_{1-\alpha/2}$

Paramètre $\theta$	$\mu_2 - \mu_1$		
Populations Écart-types $\sigma_1, \sigma_2$	≈ normales connus	— connus	≈ normales
			inconnus
Échantillons	—	$n_1 > 30$ et $n_2 > 30$	$n_1 > 30$ et $n_2 > 30$
Statistique $\hat{\Theta}$	$\bar{X}_2 - \bar{X}_1$		
St. normalisée	$Z = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$Z = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	
Distribution	$N(0, 1)$		
Degrés de liberté	—		
Mesure $\hat{\theta}$	$\bar{X}_2 - \bar{X}_1$		

Paramètre $\theta$	$\mu_2 - \mu_1$	
Populations	≈ normales	
Écart-types $\sigma_1, \sigma_2$	inc., $\sigma_1 = \sigma_2$ ou $n_1 = n_2$	inc., $\sigma_1 \neq \sigma_2$ et $n_1 \neq n_2$
Échantillons	$n_1 < 30$ ou $n_2 < 30$	
Statistique $\hat{\Theta}$	$\bar{X}_2 - \bar{X}_1$	
St. normalisée	$T = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{S_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$T = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Distribution	Student ( $\nu$ )	
Degrés de liberté	$n_1 + n_2 - 2$	$\nu^*$
Mesure $\hat{\theta}$	$\bar{X}_2 - \bar{X}_1$	
Rappels	$S_c$ : diapo #78	$\nu^*$ : diapo #79

Paramètre $\theta$	$\pi_2 - \pi_1$	$\sigma_1^2/\sigma_2^2$
Populations	—	$\approx$ normales
Écart-types $\sigma_1, \sigma_2$	—	—
Échantillons	$n_1 > 30$ et $n_2 > 30$ (+++)	—
Statistique $\hat{\Theta}$	$\hat{P}_2 - \hat{P}_1$	$F$
St. normalisée	$Z = \frac{(\hat{P}_2 - \hat{P}_1) - (\pi_2 - \pi_1)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}}$	$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$
Distribution	$N(0, 1)$	Fischer ( $\nu_1, \nu_2$ )
Degrés de liberté	—	$n_1 - 1, n_2 - 1$
Mesure $\hat{\theta}$	$\hat{p}_2 - \hat{p}_1$	$s_1^2/s_2^2$

Stat. norm.	Intervalle de confiance	Test d'hypothèse $H_0 : \theta = \theta_0$		
		$H_1 : \theta \neq \theta_0$	$H_1 : \theta < \theta_0$	$H_1 : \theta > \theta_0$
$Z$	$-Z_{\frac{\alpha}{2}} < Z < Z_{\frac{\alpha}{2}}$	$Z < -Z_{\frac{\alpha}{2}}$ ou $Z > Z_{\frac{\alpha}{2}}$	$Z < -Z_\alpha$	$Z > Z_\alpha$
$T$	$-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}$	$t < -t_{\frac{\alpha}{2}}$ ou $t > t_{\frac{\alpha}{2}}$	$t < -t_\alpha$	$t > t_\alpha$
$F$	$f_{1-\frac{\alpha}{2}} < f < f_{\frac{\alpha}{2}}$	$f < f_{1-\frac{\alpha}{2}}$ ou $f > f_{\frac{\alpha}{2}}$	$f < f_{1-\alpha}$	$f > f_\alpha$
	mettre sous la forme: $\theta_L < \theta < \theta_H$	«entrer dans le monde de $H_0$ »: $\theta = \theta_0$ , calculer $z, t, \chi^2$ à partir des mesures; décisions de <i>rejet</i> de $H_0$		

- Intervalle de confiance: niveau de confiance  $1 - \alpha$
- Tests d'hypothèse: seuil de signification  $\alpha$
- Voir tableaux unifiés en annexe

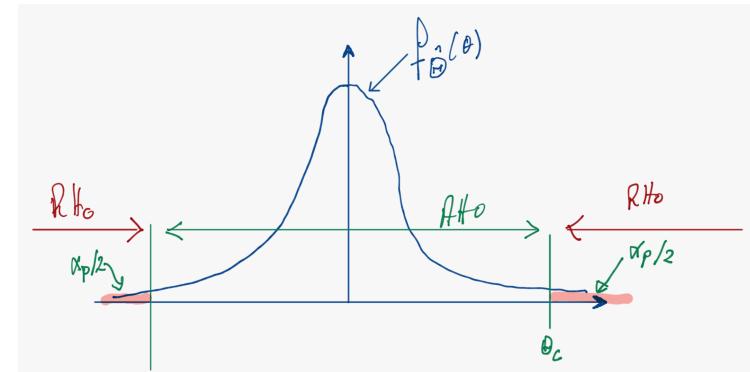
## Comment ne (pas) choisir $\alpha$

- Test statistique: «2. Choisir le seuil de signification  $\alpha$ »
- «Typiquement 1% ou 5%»
- Comment choisir?
- Comment décider?
- Pourquoi choisir  $\alpha$ ?
- Tests classiques:
  - Mesurer  $\hat{\theta}$ ; comparer  $\hat{\theta}$  aux valeurs critiques  $\hat{\theta}_c$
  - Valeurs critiques dépendent de  $\alpha$
- Alternative
  - Calculer  $\alpha_p$  (p-value) telle que  $\hat{\theta} = \hat{\theta}_c$
  - $\alpha_p$ : rejeter  $H_0$  de façon marginale
- P-valeur (seuil descriptif): la plus petite valeur de  $\alpha = P(\text{rejeter } H_0 | H_0 \text{ vraie})$  qui conduirait au rejet de  $H_0$
- La probabilité de se retrouver «au moins aussi loin» de la  $H_0$  – dans le sens de la  $H_1$  – que l'échantillon examiné, si  $H_0$  est vraie.

## Définition formelle de la P-valeur.

### DEFINITION : P-valeur

La **P-valeur** est le plus petit seuil de signification qui amènerait à rejeter l'hypothèse nulle  $H_0$ , au vu des données.



## Exemple de test classique

- Test sur la moyenne, petit échantillon, population normale,  $\sigma$  inconnu

1  $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$  (test bilatéral)

2  $\alpha$  à définir

3 Statistique à utiliser:  $\bar{X}$ ; distribution:  $T = (\bar{X} - \mu) / (S/\sqrt{n})$

4 Région critique:  $T < -t_{\alpha/2}$  et  $T > t_{\alpha/2}$

5 Règle de décision: rejeter  $H_0$  si  $t < -t_{\alpha/2}$  ou  $t > t_{\alpha/2}$

6 Prélever un échantillon et faire les calculs

Population  $N(0.5, 1), n = 5$

-> x=0.5+np.random.normal(0,1,n)

-> x= [ 0.191 -0.456 0.95 -0.057 2.005] mean: [0.527] stdev : [0.972]

$\mu_0 = 0$ , calculer  $t$ :

->t=(xm-0.0) / (xs/np.sqrt(n)); t: [1.211]

$\alpha = 0.05$ , calculer  $t_c = t_{\alpha/2}$ :

->talpha2=stats.t.ppf(1-0.025,df=4); talpha2 : [2.776]

7 Décider :  $-t_{\alpha/2} < t < t_{\alpha/2}$ , on ne peut pas rejeter  $H_0 : \mu = \mu_0 = 0$

Tests: au delà du seuil de signification

## Exemple de p-valeur

- Test sur la moyenne, petit échantillon, population normale,  $\sigma$  inconnu

1  $H_0: \mu = \mu_0, H_1: \mu \neq \mu_0$  (test bilatéral)

2 Statistique à utiliser:  $\bar{X}$ ; distribution:

$T = (\bar{X} - \mu) / (S/\sqrt{n})$

3 Prélever un échantillon et faire les calculs

Population  $N(0.5, 1), n = 5$

-> x=0.5+np.random.normal(0,1,n)

-> x= [ .191 0.456 0.95 -.057 2.005] mean: [.527] stdev : [.972]

$\mu_0 = 0$ , calculer  $t$ :

->t=(xm-0.0) / (xs/np.sqrt(n)); t: [1.211]

Quelle est la valeur de  $\alpha$  qui donne  $t = t_c = t_{\alpha/2}$ ?

-> pvalue=(1-stats.t.cdf(t, df=4))\*2 : [0.292]

7. Décider : échantillon probable si  $H_0$  est vraie : on accepte  $H_0$ .

Tests: au delà du seuil de signification

90

## Exemple de p-valeur

- Autre exemple

-> x=[10. 10. 10. 10. 5.] mean: [9.] stdev : [2.23607]

->t=(xm-0.0) / (xs/np.sqrt(n)); t: [9]

-> pvalue=(1-stats.t.cdf(t, df=4))\*2 : [0.00084]

7. Décider : échantillon probable si  $H_0$  rejeté : on rejette  $H_0$ .

Tests d'hypothèses

Tests: au delà du seuil de signification

90

## Comment utiliser la p-valeur - cas de grands échantillons

A priori une très petite  $p$  – valeur conduit à rejeter  $H_0$  avec force (c'est pour cela que  $H_0$  est souvent l'inverse de ce qu'on veut prouver).

Cependant, quand les échantillons sont grands, un petit écart sur  $\hat{x}$  peut mener à une très faible  $p$  – valeur, et fausser les conclusions.

### p-valeur pour un test de vitesse

On mesure la vitesse sur autoroute, en considérant qu'une vitesse supérieure à 120 km/h est dangereuse. Considérant que l'écart-type de la vitesse est de 10 km/h, on obtient une valeur  $\bar{x} = 125$  km/h.

Les valeurs de la p-valeur et de la puissance du test à  $\alpha = 0.05$  pour une hypothèse alternative  $\mu_1 = 125$  sont alors :

n	p-valeur	puissance du test
10	0.215	0.196
25	0.106	0.346
50	0.039	0.549
100	0.006	0.804
400	$2.8 \cdot 10^{-7}$	0.9996
1000	$1.3 \cdot 10^{-15}$	1

## Test d'adéquation à une loi

Comparer, à l'issue d'une expérience aléatoire, des fréquences expérimentales aux fréquences prévues par la théorie (Pearson, 1900).

- $k$ : nombre de fréquences à comparer (nombre de classes)
- $o_i$ : fréquences Observées (obtenues expérimentalement)
- $e_i$ : fréquences «Espérées» (théoriques, à calculer)
- 

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

- Loi du  $\chi^2$  à  $\nu$  degrés de liberté; si  $o_i = e_i$ ,  $\chi^2 = 0$ , sinon  $\chi^2 > 0$
- Calculer  $\chi^2$  à partir de  $o_i$ ,  $e_i$ ; obtenir  $\alpha = P(X^2 > \chi^2)$ , la p-value
- $\nu = k - 1 - (\text{nombre de paramètres estimés utilisés dans le calcul de } e_i)$
- Condition:  $e_i \geq 5$  au moins pour 80% des classes;  $e_i > 0$  pour les autres
- Applications: test d'adéquation, d'indépendance, d'homogénéité, de proportions

## Test d'adéquation (ou d'ajustement)

$H_0$ : les données expérimentales ont été obtenues à partir d'une population suivant la loi  $p_X(x)$  (p.ex., normale, uniforme, etc.).

- Exemple: données sur plusieurs lancers d'un dé (données simulées...)

Face	1	2	3	4	5	6	Total N
Fréquence ( $o_i$ )	1037	937	1055	1034	929	1008	6000
O = [ 1037 937 1055 1034 929 1008 ]							

- $H_0$ : le dé est bien équilibré;  $p_i = 1/6$ ,  $e_i = p_i N = 1000$   
 $e = np.ones(6) * N / 6$
- Conditions: OK (sinon grouper des classes voisines)
- Calculer  $\chi^2 = 14.624$  ( $\text{chi2}=np.sum((O-e)**2) / (N/6)$ )
- $\nu = 6 - 1 - 0 = 5$
- p-value:  $P(X^2 > 14.624) =$   
 $Q=1-\text{stats.chi2.cdf(chi2, df=5)}$   
 $Q= 0.0120957$
- On peut rejeter  $H_0$  au seuil de signification 5%

## Test d'indépendance / tableau de contingence

On mesure, sur chaque individu d'un échantillon aléatoire de taille  $n$ , deux caractères  $X$  et  $Y$ , à  $I$  et  $c$  modalités, respectivement.  
 $H_0$ : les deux caractères  $X$  et  $Y$  sont indépendants.

- Exemple: le tabac et les jeunes, INPES, baromètre santé 2000.
- 

Sexe \ Fumeur	Oui	Non	Total
Homme	340 (310)	314 (344)	654
Femme	289 (319)	384 (354)	673
Total	629	698	1327

- $H_0$ :  $X$  et  $Y$  sont indépendants ;  $\pi_{ij} = \pi_i \pi_j$  ( $i = 1, \dots, I$ ;  $j = 1, \dots, c$ )
- On estime  $\pi_i$  et  $\pi_j$  à partir des fréquences marginales de l'échantillon
- $\pi_{ij} = \pi_i \pi_j \rightarrow \frac{e_{ij}}{n} = \frac{\sum_{j=1}^c o_{ij}}{n} \frac{\sum_{i=1}^I o_{ij}}{n} \rightarrow e_{ij} = \frac{1}{n} \sum_{j=1}^c o_{ij} \sum_{i=1}^I o_{ij}$
- Degrés de liberté  $\nu = (Ic - 1) - [(I - 1) + (c - 1)] = (I - 1)(c - 1)$
- Conditions: OK (sinon? augmenter la taille de l'échantillon!)

## Test d'indépendance: correction de Yates

- Si  $\nu = 1$  (tableau  $2 \times 2$ ) utiliser:

$$\chi^2 = \sum_{i,k} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}$$

- Calculer  $\chi^2 = 10.52$
- $\nu = (2 - 1)(2 - 1) = 1$
- p-value:  $P(X^2 > 3.657) = 0.0011$
- On peut rejeter  $H_0$  au seuil de signification 1%

## Test d'homogénéité

À partir de  $c$  populations, on obtient  $c$  échantillons aléatoires et indépendants, de taille  $n_j$  ( $j = 1, \dots, c$ ). On mesure sur chaque individu le même caractère  $X$ , à  $I$  modalités.

$H_0$ : la proportion d'individus appartenant à la  $i$ -ème modalité ( $i = 1, \dots, I$ ), reste la même pour toutes les populations (les populations sont homogènes par rapport au caractère étudié).

- Example: notes (fictives) échantillonées dans trois parcours

Note \ Parcours	I	II	III	Total
$0 \leq x < 6$	32	15	8	55
$6 \leq x < 12$	123	60	43	226
$12 \leq x \leq 20$	145	125	149	419
Total ( $n_j$ )	300	200	200	700

- $H_0$ : proportion de chaque modalité constante ;  
 $\pi_{i1} = \pi_{i2} = \dots = \pi_{ic} = \pi_i$  ( $i = 1, \dots, I$ )
- On estime  $\pi_i$  à partir des fréquences marginales de l'échantillon

## Test d'homogénéité

Note \ Parcours	I	II	III	Total
$0 \leq x < 6$	32 (23.57)	15 (15.71)	8 (15.71)	55
$6 \leq x < 12$	123 (96.86)	60 (64.57)	43 (64.57)	226
$12 \leq x \leq 20$	145 (179.57)	125 (119.71)	149 (119.71)	419
Total ( $n_j$ )	300	200	200	700

- $H_0$ : proportion de chaque modalité constante;  
 $\pi_{i1} = \pi_{i2} = \dots = \pi_{ic} = \pi_i$  ( $i = 1, \dots, I$ )
- On estime  $\pi_i$  à partir des fréquences marginales de l'échantillon

$$\pi_{ij} = \pi_i \rightarrow \frac{e_{ij}}{n_j} = \frac{\sum_{j=1}^c o_{ij}}{n} \rightarrow e_{ij} = \frac{1}{n} \sum_{j=1}^c o_{ij} \underbrace{\sum_{i=1}^I o_{ij}}_{n_j}$$

- Degrés de liberté  $\nu = (Ic - 1) - [(I - 1) + (c - 1)] = (I - 1)(c - 1)$
- Conditions: OK (sinon? augmenter la taille de l'échantillon!)

## Test d'homogénéité

- Même formule que le test d'indépendance!

$$\chi^2 = \sum_{i,k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Calculer  $\chi^2 = 35.4729$
- $\nu = (3 - 1)(3 - 1) = 4$
- p-value:  $P(\chi^2 > 35.4729) =$   
 $[P \ Q] = \text{cdfchi}("PQ", 35.4729, 4)$   
 $Q = 3.714026 \ 10e7 \ P = 0.9999996$
- On peut rejeter  $H_0$  pratiquement à n'importe quel seuil de signification!

À partir de  $c$  populations, on obtient  $c$  échantillons aléatoires et indépendants, de taille  $n_j$  ( $j = 1, \dots, c$ ). On mesure sur chaque individu le même caractère  $X$ , à 2 modalités («oui» / «non»).

$H_0$ : la proportion de «oui» reste la même pour toutes les populations (cas spécial du test d'homogénéité,  $I = 2$ ).

- Exemple: nombre de pièces défectueuses et moment de production

Pièces \ Créneau	Matin	Après-midi	Nuit	Total
Défectueuses («O»)	45 (56.97)	55 (56.67)	70 (56.37)	170
Normales («N»)	905 (893.03)	890 (888.33)	870 (883.63)	2665
Total ( $n_j$ )	950	945	940	2835

- $H_0$ :  $\pi_1 = \pi_2 = \dots = \pi_c = \pi$
- On estime  $\pi$  à partir des fréquences marginales de l'échantillon
- «Oui»:  $\pi_j = \pi \rightarrow \frac{e_{1j}}{n_j} = \frac{\sum_{j=1}^c o_{1j}}{n}$
- «Non»:  $1 - \pi_j = 1 - \pi \rightarrow \frac{e_{2j}}{n_j} = \frac{\sum_{j=1}^c o_{2j}}{n}$

$$\blacksquare \quad e_{ij} = \frac{n_j}{n} \sum_{j=1}^c o_{ij} \rightarrow \boxed{e_{ij} = \frac{1}{n} \sum_{j=1}^c o_{ij} \sum_{i=1}^l o_{ij}}$$

- Même formule que le test d'indépendance / d'homogénéité!
- Degrés de liberté  $\nu = (2 - 1)(c - 1) = c - 1$
- Conditions: OK (sinon? augmenter les tailles des échantillons!)
- Calculer  $\chi^2 = 6.2339$
- $\nu = (3 - 1) = 2$
- p-value:  $P(X^2 > 6.2339) =$   
 $[P \ Q] = \text{cdfchi}(\text{"PQ"}, 6.2339, 2)$   
 $Q = 0.04429$
- On peut rejeter  $H_0$  au seuil de signification 5%

Même contexte qu'avant:  $c$  populations,  $c$  échantillons, caractère  $X$  à deux modalités.

$H_0$ : les proportions de «oui»,  $\pi_1, \dots, \pi_c$ , sont égales à  $p_1, \dots, p_c$  (pas d'estimation de paramètres).

- «Oui»:  $\pi_j = p_j \rightarrow \frac{e_{1j}}{n_j} = p_j$
- «Non»:  $1 - \pi_j = 1 - p_j \rightarrow \frac{e_{2j}}{n_j} = 1 - p_j$
- $\nu = c$ : on ne perd aucun degré de liberté
- Exemple précédent avec:  
 $p_1 = 0.05, p_2 = 0.06, p_3 = 0.08 (\neq 170/2835 \approx 0.06)$
- Calculer  $\chi^2 = 0.5836$
- $\nu = 3$
- p-value:  $P(X^2 > 0.5836) = 0.9002$
- On ne peut pas rejeter  $H_0$

$H_0$ : les données expérimentales (échantillon de taille  $n$ ) ont été obtenues à partir d'une population normale.

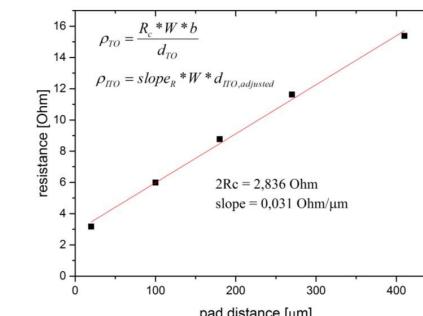
- Procédure «classique»: test du  $\chi^2$ 
  - 1 Répartir les données en classes (histogramme)
  - 2 Estimer  $\mu$  et  $\sigma$  avec `cdfnor`
  - 3a. Calculer les probabilités théoriques  $p_j$  des classes  
 Calculer les fréquences théoriques  $e_j = p_j n$   
 Vérifier les conditions sinon regrouper les classes
  - 3b. Ou répartir en  $(M + 1)$  classes équiprobabiles:  $e_j = n/(M + 1)$
  4. Calculer  $\chi^2$  (on perd deux d.l. avec l'estimation de  $\mu$  et  $\sigma$ !)
- Une grande p-value permet de ne pas rejeter l'hypothèse de normalité

## La Régression : faire le lien entre variables aléatoires

La régression est un ensemble de techniques pour faire le lien entre deux grandeurs, à partir de mesures faites sur ces grandeurs.

C'est un des outils simples, mais fondamentaux de l'Intelligence Artificielle.  
 La régression linéaire établit un lien linéaire entre deux grandeurs.

Simple mais très puissant !



Krause, Stephan ; Kaufmann, Kai ; Lancaster, Kevin ; Naumann, Volker. (2016). *Fs-laser micro machining for mu-TLM resistivity test structures in photovoltaic TCO multilayers*.

## Régression : Expliquer $Y$ à partir de $X$

### L'objectif de la régression (multi-) Linéaire

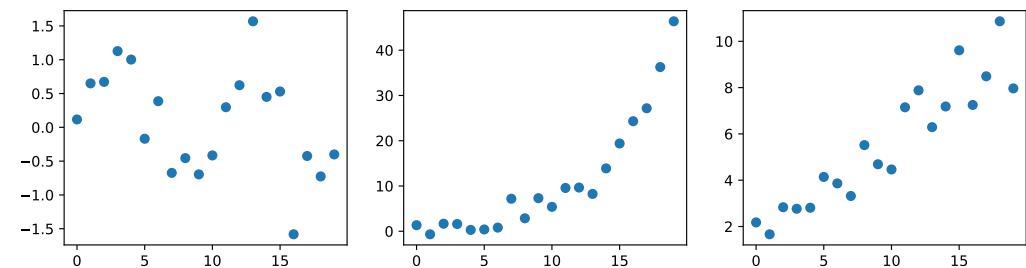
- Expliquer une variable  $Y$   
**(La variable dépendante)**
- A partir de  $X$  (cas multi-linéaire :  $X_i, (i = 1, \dots, q)$ )  
**(La/les variable(s) indépendantes ou explicatives)**

Avec une relation de type Linéaire :

$$Y = \beta_0 + \beta_1 X \quad \text{Multi-linéaire : } Y = \beta_0 + \sum_{i=1}^q \beta_i X_i$$

## Première étape : tracer les “nuages de points”

Sur l'exemple ci-dessous : seul le troisième nuage semble indiquer une relation linéaire :



## La régression linéaire simple : Modélisation

$$Y = \beta_0 + \beta_1 X$$

- $\beta_0$  = ordonnée du point d'intersection de la droite avec l'axe vertical ( $X = 0$ )
- $\beta_1$  = pente de la droite

### Modélisation à partir des données

Soient les données issues d'un échantillon :  $(x_i, y_i), i = 1, \dots, n$ .  $X$  est ici connu parfaitement (non aléatoire).  $Y$  est aléatoire et modélisé par :

$$Y = \beta_0 + \beta_1 X + E$$

où  $E \sim \mathcal{N}(0, \sigma^2)$ , ce qui, exprimé pour les données donne :

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

On notera qu'on a trois paramètres déterministes inconnus :  $\beta_0, \beta_1, \sigma^2$

## La régression linéaire simple : Solution par les moindres carrés

Trouver  $\beta_0, \beta_1$ , grâce à l'hypothèse de bruit Gaussien :

$$(\beta_0, \beta_1) = \arg \min_{\beta_0, \beta_1} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

## La régression linéaire simple : Solution par les moindres carrés

En dérivant par rapport aux paramètres et en égalant les dérivées à 0, on obtient :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}$$

$$\hat{\beta}_1 = \boxed{r \frac{s_y}{s_x}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

où  $\bar{x}, \bar{y}, s_x, s_y$  sont les moyennes et écarts-types des échantillons et  $r$  le coefficient de corrélation :

$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

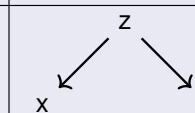
## Corrélation ne signifie pas causalité

### 4 cas de corrélation

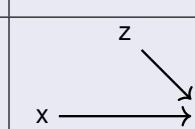
**Liens Causaux** : Causalité possible dans les deux sens



**Cause cachée** :  $Z$  est une cause commune à  $X$  et  $Y$



**Facteur Commun** :  $Z$  et  $X$  sont ensemble cause de  $Y$



**Coïncidence** : Corrélation "par accident"

## coefficient de corrélation, de détermination et erreur résiduelle

Note :  $r^2$  est appelé **Coefficient de détermination**.

C'est la proportion de la variabilité expliquée par le modèle linéaire.

Note: Certains écrivent également

$$\hat{\beta}_1 = \boxed{\frac{S_{xy}}{S_{xx}}}$$

ou

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^n y_i (x_i - \bar{x})$$

On qualifie l'erreur résiduelle par l'erreur quadratique :

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

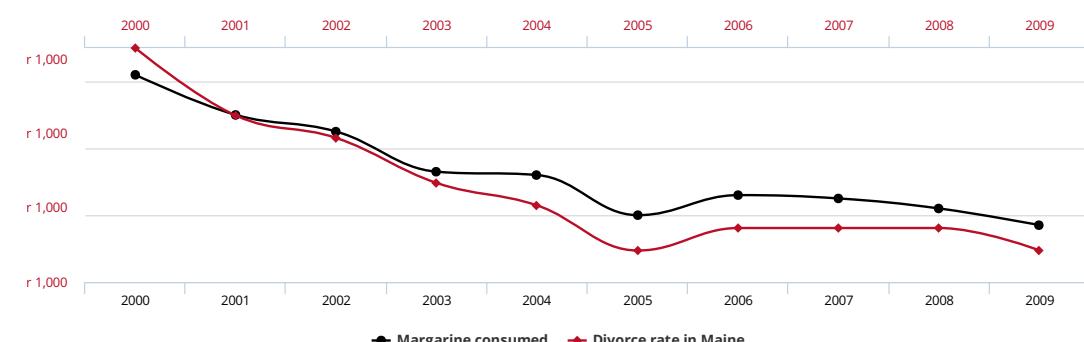
Et on peut montrer que  $E[SS_E] = (n-2)\sigma^2$ .

## Exemple de coïncidence étrange !

### Divorce rate in Maine

correlates with

### Per capita consumption of margarine



From Tyler Vigen, *Spurious Correlations*

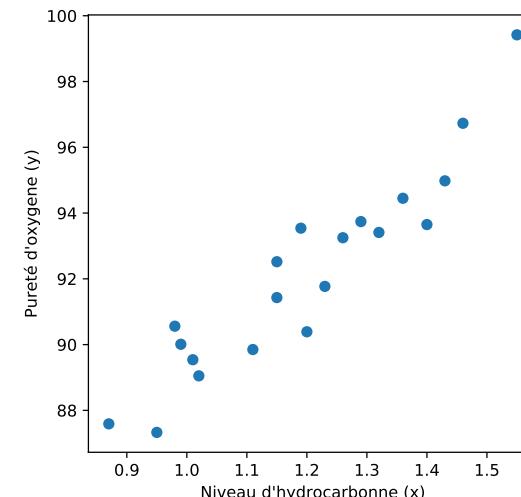
## Exemple de travail.

## Exemple de travail.

On observe un processus de distillation chimique, avec comme variable  $X$  le pourcentage d'hydrocarbone présente dans le distillateur et comme variable  $Y$  la niveau de pureté de l'oxygène en pourcent. On observe un échantillon de taille  $n = 20$ .

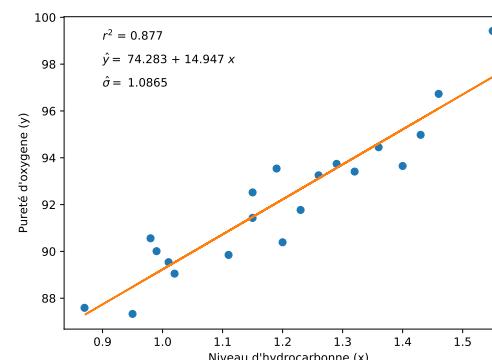
$x$	0.99	1.02	1.15	1.29	1.46	1.36	0.87	1.23	1.55	1.40
$y$	90.01	89.05	91.43	93.74	96.73	94.45	87.59	91.77	99.42	93.65
$x$	1.19	1.15	0.98	1.01	1.11	1.20	1.26	1.32	1.43	0.95
$y$	93.54	92.52	90.56	89.54	89.85	90.39	93.25	93.41	94.98	87.32

From Montgomery, *Applied Statistics*, table 11-1



## Exemple de travail.

## Estimation de la pente



L'estimateur de la pente est non biaisé

En remplaçant  $y_i$  par  $\beta_0 + \beta_1 x_i + \epsilon_i$ , avec  $E[\epsilon_i] = 0$  dans l'expression de  $\hat{\beta}_1$ , on trouve que

$$E[\hat{\beta}_1] = \beta_1$$

La variance de  $\hat{\beta}_1$  étant inconnue, la statistique normalisée de  $\hat{\beta}_1$  est :

$$t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{S_{\beta_1}}$$

Qui suit une loi de Student à  $n - 2$  degrés de liberté.

## Estimation de la pente

En effet, on peut montrer que

$$s_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Les  $x$  sont proches les uns des autres

Et on a vu précédemment que

Les erreurs sont grandes

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 2}$$

Remarquez qu'en effet,  $s_{\hat{\beta}_1}$  est d'autant plus grand que les  $x_i$  sont proches les uns des autres (il est difficile de faire passer une ligne) et d'autant plus petit que l'erreur sur la prédiction des  $y_i$  est petite.

De la même manière, la statistique du point d'interception est :

$$t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}}$$

qui suit également une loi de Student à  $n - 2$  degrés de liberté. Son écart-type vaut :

$$s_{\hat{\beta}_0} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}$$

## Estimation du coefficient de corrélation

La statistique pour la corrélation est la corrélation standardisée :

$$t_r = r \sqrt{\frac{n - 2}{1 - r^2}}$$

qui suit également une distribution de Student à  $n - 2$  degrés de liberté.

## Estimation du prédicteur

Soit une valeur  $x^o$ , (qui peut ne pas appartenir aux données prélevées). Pour cette valeur, on peut prédire la valeur de  $y$  :

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^o$

Cela correspond au modèle probabiliste :

$$Y|x^o = \underbrace{\beta_0 + \beta_1 x^o}_{\text{défini comme } \mu(x^o)} + \epsilon$$

$\epsilon \sim \mathcal{N}(0, \sigma^2)$

## Estimation du prédicteur

$$\begin{array}{l} n \uparrow \rightarrow \hat{\sigma}_{\mu} \downarrow \\ x^o \text{ 与 } \bar{x} \text{ 高近} \rightarrow \hat{\sigma}_{\mu} \end{array}$$

L'écart type de  $\hat{\mu}(x^o)$  (c'est à dire la moyenne de la droite en  $x^o$ ) vaut :

$$s_{\hat{\mu}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}$$

D'autre part, l'écart-type de  $\hat{y}$  est presque identique, à part un "+1" qui vient du bruit dans l'équation  $Y|x^o = \mu(x^o) + \epsilon$

$$s_{\hat{y}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} + 1}$$

## Test d'hypothèse en régression linéaire

Test sur la pente

$y$  与  $x$  有关?

$$\begin{array}{ll} \beta_1 = 0 & \checkmark \\ \beta_1 \neq 0 & \times \end{array}$$

avec  $t_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$  qui suit une loi de Student à  $n - 2$  degrés de liberté.

De même, on peut faire le test :

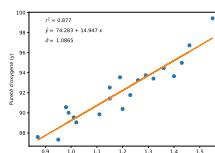
faire

$$\begin{array}{ll} H_0 : \beta_0 = \beta_{0,o} & \\ H_1 : \beta_0 \neq \beta_{0,o} & \end{array}$$

avec  $t_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,o}}{s_{\hat{\beta}_0}}$  qui suit une loi de Student à  $n - 2$  degrés de liberté.

## Exemple de travail

Pour notre exemple de travail, nous allons vérifier qu'il y a bien un lien entre le niveau d'hydrocarbone et la pureté de l'oxygène, c'est à dire que :  $\beta_1 \neq 0$ .



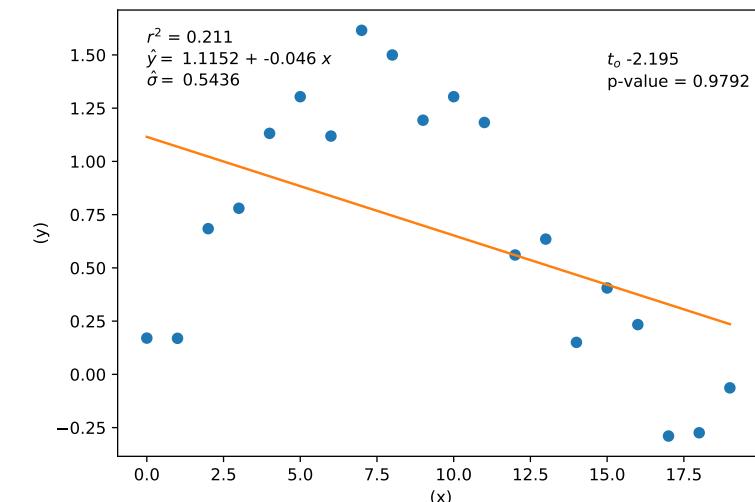
$$\begin{array}{ll} H_0 : \beta_1 = 0 & \\ H_1 : \beta_1 \neq 0 & \end{array}$$

Nous avons

obtenu  $\hat{\beta}_1 = 14.97$ ,  $\hat{\sigma} = 1.086$ , ce qui conduit à  $s_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = 1.317$ .

On obtient  $t_{ech} = 11.35$ , pour une loi de Student à 18 degrés de libertés, qui mène à une p-valeur de 0 selon les tables ( $610^{-10}$  à l'ordinateur), et on rejette clairement  $H_0$ .

## Autre Exemple de travail



## Intervalle de confiance sur la pente et le point d'interception

En fonction des statistiques données précédemment :

$$\text{Intervalle de confiance au niveau } (1 - \alpha) \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

L'intervalle de confiance sur la pente est donné par :

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \hat{\sigma} / \sqrt{S_{xx}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \hat{\sigma} / \sqrt{S_{xx}}$$

L'intervalle de confiance sur le point d'interception est donné par :

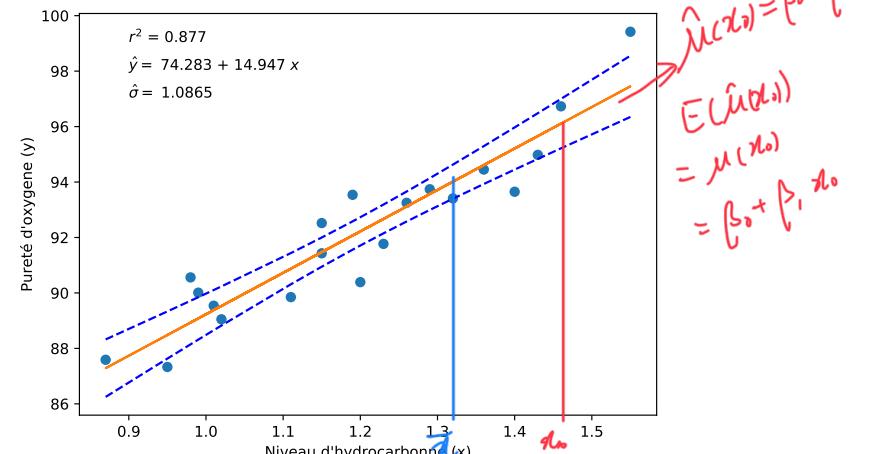
$$\hat{\beta}_0 - t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

Exemple de travail : on a  $t_{0.025, 18} = 2.101$ , ce qui donne avec les nombres donnés précédemment

$$12.19 \leq \beta_1 \leq 17.7$$

Exemple de travail : intervalles de confiance à 95 % de la moyenne de la réponse

$$y = \beta_0 + \beta_1 x$$



## Intervalle de confiance sur la valeur moyenne de la réponse

Soit une valeur quelconque  $x^o$ , on a que la valeur moyenne de  $y$  en  $x^o$  vaut :

$$\hat{\mu}_{Y|x^o} = \hat{\beta}_0 + \hat{\beta}_1 x^o$$

avec

$$s_{\hat{\mu}} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}}}$$

Intervalle de confiance au niveau  $(1 - \alpha)$ 

L'intervalle de confiance sur le prédicteur est donné par :

$$\hat{\mu}_{Y|x^o} - t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}}} \leq \mu_{Y|x^o} \leq \hat{\mu}_{Y|x^o} + t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}}}$$

## Intervalle de confiance sur le prédicteur

$$E \sim N(\mu, \sigma^2)$$

Soit une valeur quelconque  $x^o$ , on a que la valeur prédictée de  $y$  en  $x^o$  vaut :

$$\hat{y}^o = \hat{\mu}_{Y|x^o} + \epsilon = \hat{\beta}_0 + \hat{\beta}_1 x^o + \epsilon$$

avec

$$s_{\hat{y}^o} = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}} + 1}$$

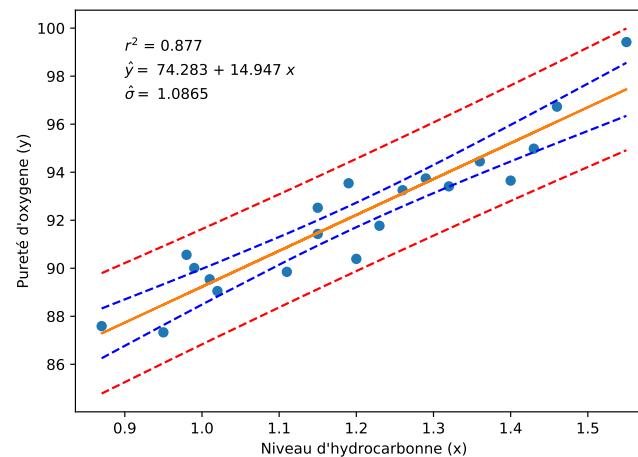
Intervalle de confiance au niveau  $(1 - \alpha)$ 

L'intervalle de confiance sur le prédicteur  $\hat{y}^o$  est donné par :

$$\hat{y}^o - t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}} + 1} \leq y^o \leq \hat{y}^o + t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^o - \bar{x})^2}{S_{xx}} + 1}$$

## Exemple de travail : intervalles de confiance à 95 % de la prédition

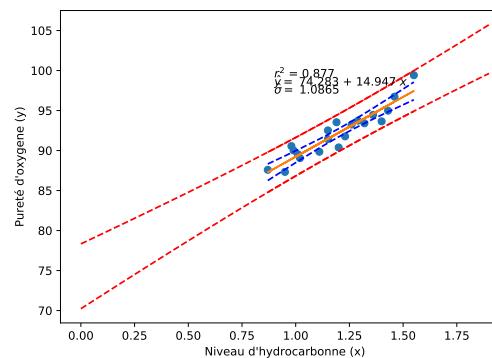
## Interpolation et Extrapolation



## Exemple de travail : intervalles de confiance à 95 % de la prédition

L'interpolation se fait dans l'intervalle hors de la partie où les points sont présents.

La validité du modèle dans cet intervalle n'est pas garantie !



Dans le cadre de la régression on parle :

- **d'interpolation** : quand on estime les valeurs de  $\hat{y}$  pour les valeurs de  $x$  comprises DANS l'intervalle des données d'origine.
- **d'extrapolation** : quand on estime les valeurs de  $\hat{y}$  pour les valeurs de  $x$  comprises HORS l'intervalle des données d'origine.

L'extrapolation est toujours nettement plus hasardeuse que l'interpolation.

## Régression non linéaire “simple”

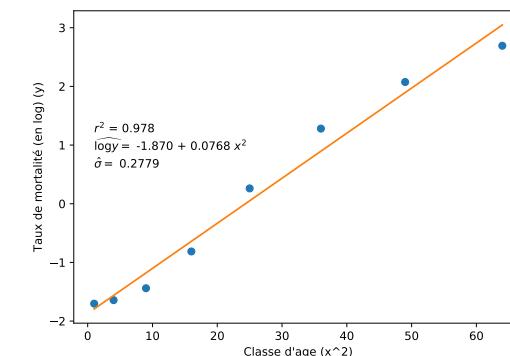
Prenons un exemple où on suppose que :

$$Y = ae^{bx^2}$$

On peut transformer cette équation en un problème de régression linéaire :

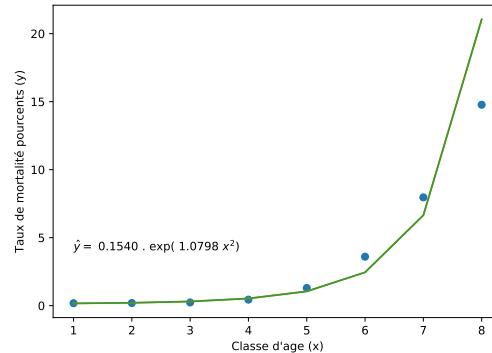
$$Y_i = \log(a) + bX_i, \text{ avec } Y_i = \log(Y) \text{ et } X_i = X^2.$$

Exemple : les données du TD 8.4.2



## Régression non linéaire "simple"

Qui donne, une fois revenu dans le domaine d'origine :



Remarquez qu'il faut bien faire un test d'adéquation ensuite pour vérifier !  
(voir exercice 8.4.2)

## Un introduction à la régression multilinéaire et polynomiale

Supposons que  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon$ ,  $\epsilon$  étant un bruit gaussien.

On notera qu'il est facile, si on peut résoudre le problème précédent, de faire de la régression polynomiale en posant  $X_k = X^k$ .

On fait  $n$  mesures de  $Y$  :  $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i$ ,  
Soit, sous forme matricielle :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

## Un introduction à la régression multilinéaire et polynomiale

avec

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Ici, la solution est également la solution d'un problème de moindres carrés qui vaut :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

où  $\mathbf{X}^T$  est la transposée de la matrice  $\mathbf{X}$ .

Nous n'irons pas plus loin, mais c'est le début de "l'intelligence artificielle".

Notons également que la régression polynomiale, par exemple, pose beaucoup d'autres problèmes, comme l'overfitting ... voir la vidéo

<https://youtube.com/watch?v=DQWI1kvmwRg>