Séance 10

Pierre-Marc Juneau

27 octobre 2022

Plan

1. Distributions discrètes et continues définies
2. Distributions pour l'inférence statistique
3. Tests d'hypothèses (1ère partie)
4. Évaluation formative

1- Distributions discrètes et continues définies

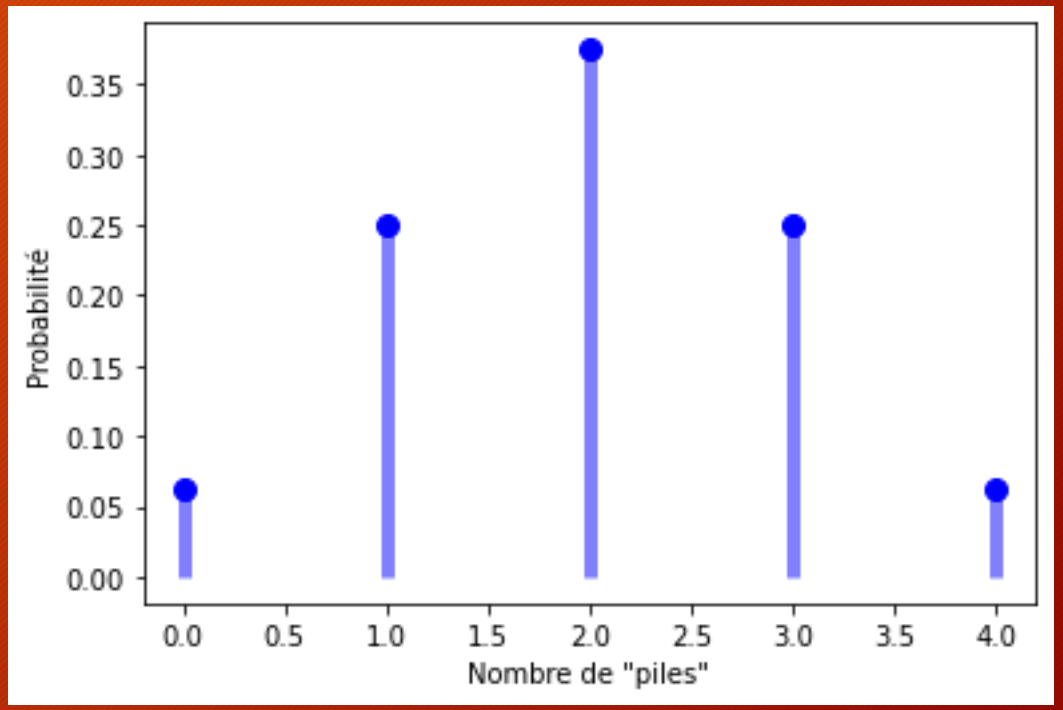
Distributions dans la librairie scipy.stats

- Beaucoup de distributions définies (avec des paramètres) sont disponibles dans la librairie *scipy.stats*
  - https://docs.scipy.org/doc/scipy/reference/stats.html
- Il est possible de choisir la bonne distribution théorique (avec fonction définie) qui va correspondre le mieux à une distribution de données, selon le contexte
- Nous en verrons quelques-unes et les fonctions pour les calibrer et/ou explorer

1- Distributions discrètes définies

Distribution (loi) binomiale

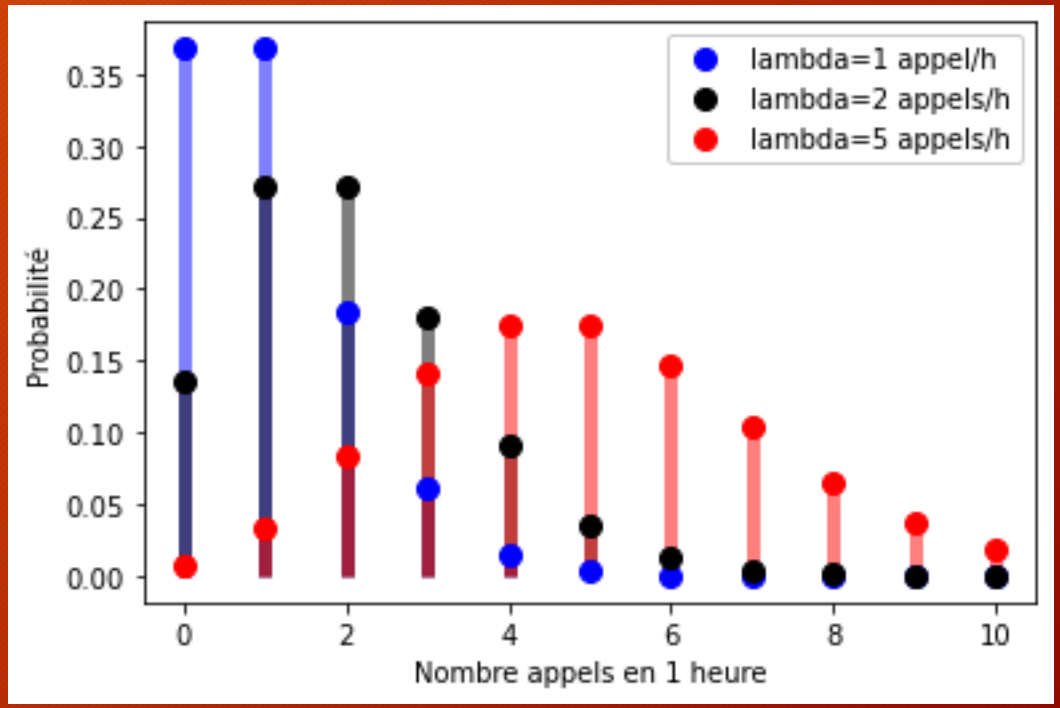
- Résultat binaire: « succès ou échec », « oui ou non », lorsque « n » expériences sont réalisées
- Exemple: tirer une pièce de monnaie 4 fois: succès d'avoir piles 0 fois, 1 fois, 2 fois, 3 fois ou 4 fois.
  - *import numpy as np*
  - *import matplotlib.pyplot as plt*
  - *import scipy.stats as sts*
  - *n=4*
  - *p=1/2*
  - *x=np.linspace(0, n, n+1)*
  - *probabilités=sts.binom.pmf(x, n, p)*



1- Distributions discrètes définies

Distribution (loi) de Poisson

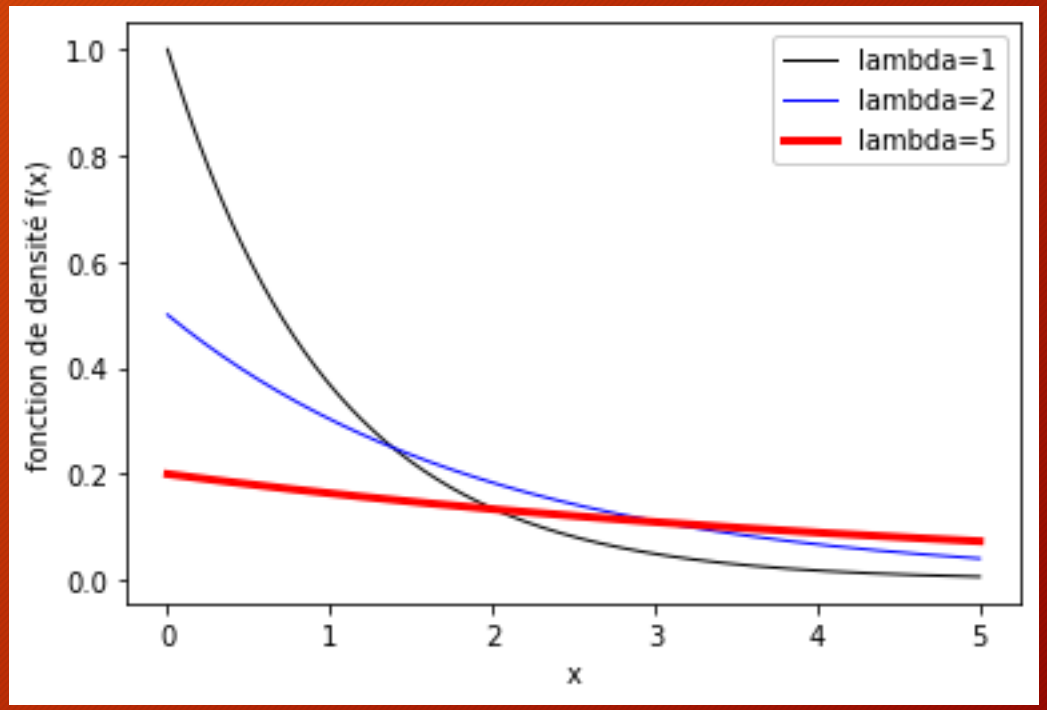
- Distribution discrète, décrivant par exemple le nombre d'évènements qui vont arriver par unité de temps, selon un taux moyen (ex: nombre d'appels téléphoniques en 1 heure = λ)
  - *lambdaa=5*
  - *x=np.linspace(0, 10, 10+1)*
  - *probabilités=sts.poisson.pmf(x,lambdaa)*



1- Distributions continues définies

Distribution (loi) exponentielle

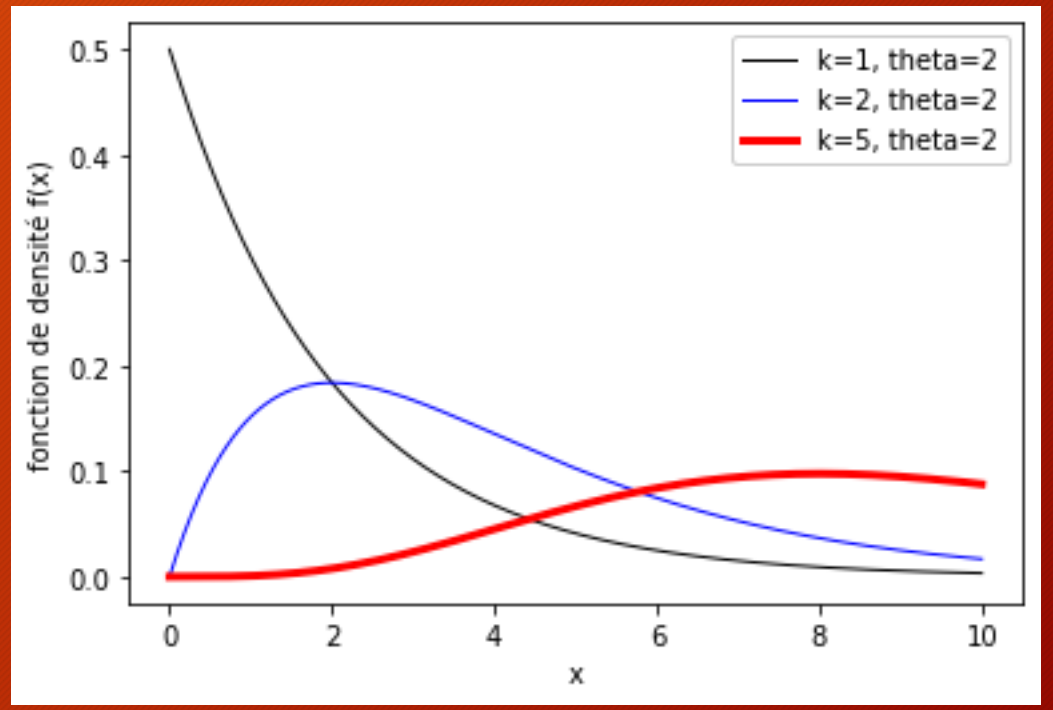
- Souvent utilisée pour mesurer les intervalles de temps entre des événements, des temps de service ou des analyses de bris (maintenance, en bris de pièces par heure).
- Utilise également un taux λ (ex: un taux d'évènement par heure)
  - *d=1000*
  - *MIN=0*
  - *MAX=5*
  - *lambdaa=[0,1,2]*
  - *grille_x = np.linspace(MIN, MAX, d)*
  - *pdf=sts.expon.pdf(grille_x,2)*



1- Distributions continues définies

Distribution (loi) gamma

- Utilisée couramment dans le domaine des assurances, des sciences, de l'ingénierie et de la finance (durées d'événements ou temps entre des événements, etc).
- Utilize deux paramètres: k et θ
  - *d=1000*
  - *MIN=0*
  - *MAX=10*
  - *k=[1,2,5]*
  - *theta=[2]*
  - *grille_x = np.linspace(MIN, MAX, d)*
  - *pdf=sts.gamma.pdf(grille_x,2,scale=2)*



1- Distributions discrètes et continues définies

- E10–1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)

Âge	Poids (kg)	Grandeur (cm	Genre
21	65.6	174	Homme
23	71.8	175.3	Homme
28	80.7	193.5	Homme
23	72.6	186.5	Homme
22	78.8	187.2	Homme
21	74.8	181.5	Homme
26	86.4	184	Homme
27	78.4	184.5	Homme
23	62	175	Homme
21	81.6	184	Homme

- Source des données: Heinz G, Peterson LJ, Johnson RW, Kerk CJ. 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education 11(2)

https://www.openintro.org/data/index.php?data=bdims

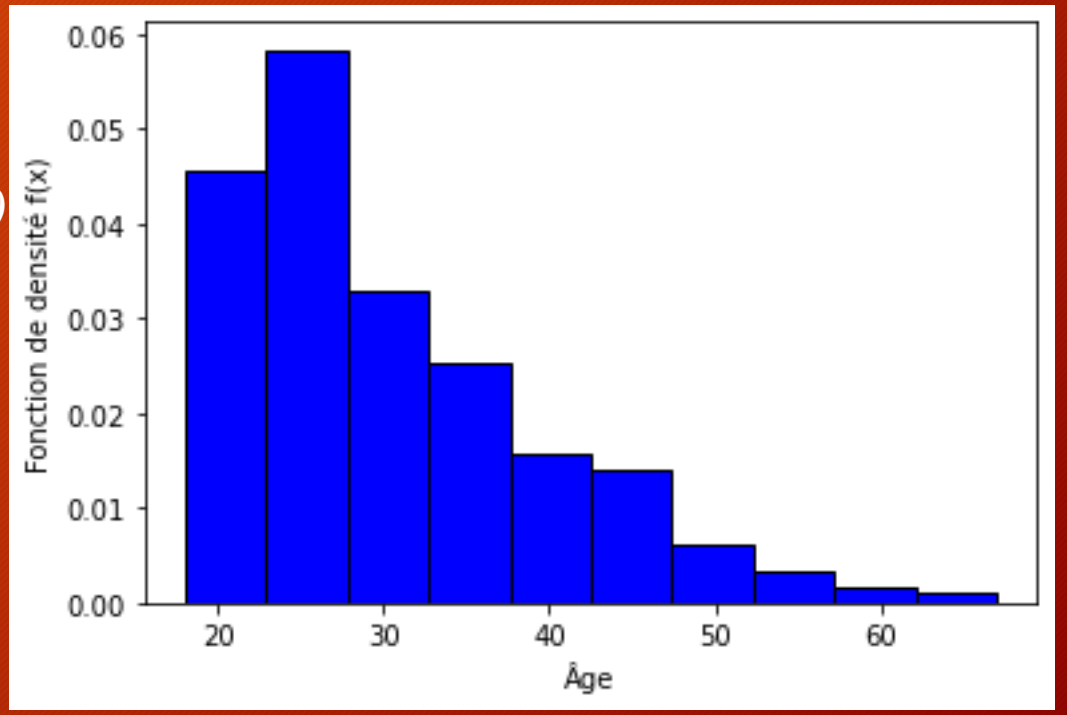
1- Distributions discrètes et continues définies

- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Les données sont chargées dans Spyder.
  - La distribution de la variable « Âge » est affichée
    - *import numpy as np*
    - *import pandas as pd*
    - *import matplotlib.pyplot as plt*
    - *import scipy.stats as sts*
    - *import statsmodels.api as stm*
    - *import statsmodels.stats.weightstats as ws*
    - *import math*

1- Distributions discrètes et continues définies

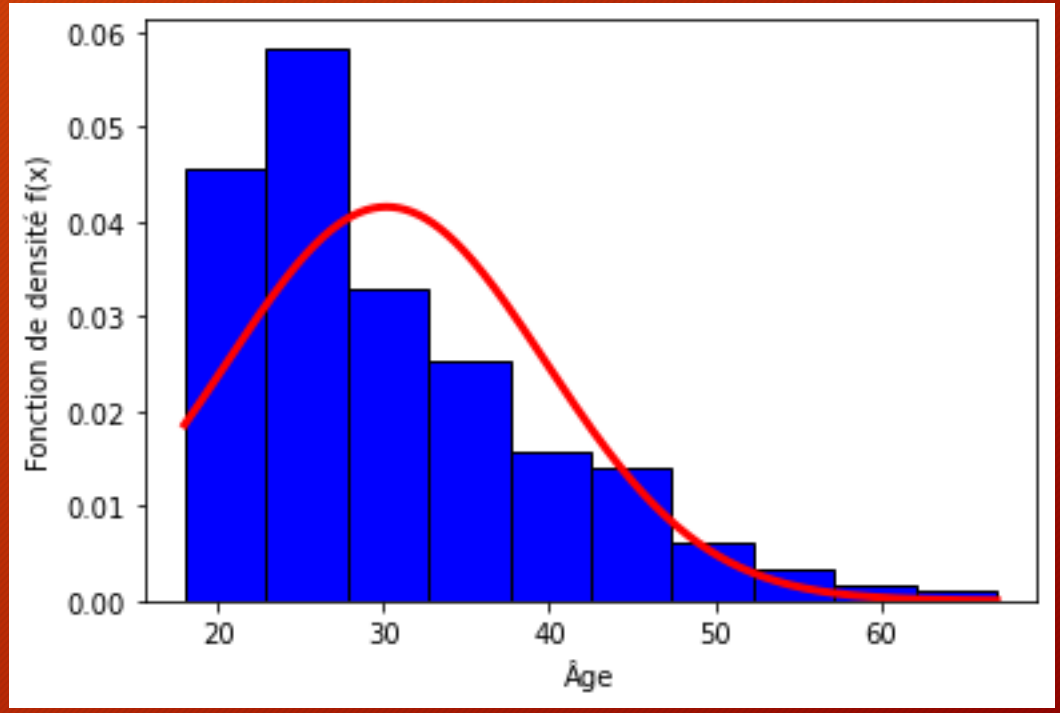
- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Les données sont chargées dans Spyder.
  - La distribution de la variable « Âge » est affichée
    - *donnee = pd.read_csv('PersonnesActivesv0r2.csv')*
    - *stats=donnee.describe()*
    - *dimensions=donnee.shape*
    - *nomsvariables = pd.DataFrame(donnee.columns)*
    - *Variable=donnee["Âge"]*
    - *ax=Variable.plot.hist(density=True, bins = 10, color = 'blue', edgecolor = 'black')*
    - *ax.set_xlabel("Âge")*
    - *ax.set_ylabel("Fonction de densité f(x)")*

La distribution ne semble pas suivre une normale



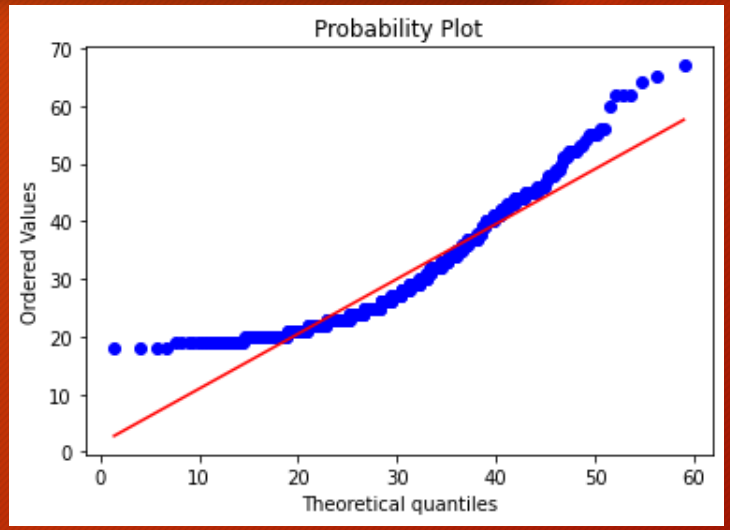
1- Distributions discrètes et continues définies

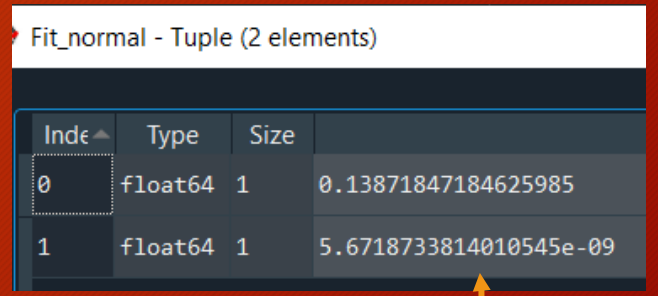
- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Essai pour calibrer (« fitter ») une distribution normale
    - *d=1000*
    - *grille_x = np.linspace(Variable.min(), Variable.max(), d)*
    - *dx=(Variable.max()-(Variable.min()))/(d-1)*
    - *mu, sigma = sts.norm.fit(Variable.values)*
    - *param=sts.norm.fit(Variable.values)*
    - *pdf = sts.norm.pdf(grille_x, mu, sigma)*
    - *ax=Variable.plot.hist(density=True, bins = 10, color = 'blue', edgecolor = 'black')*
    - *ax.set_xlabel("Âge")*
    - *ax.plot(grille_x, pdf, linewidth=3, color = 'red')*
    - *ax.set_ylabel("Fonction de densité f(x)")*



1- Distributions discrètes et continues définies

- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Essai pour calibrer (« fitter ») une distribution normale
    - *sts.probplot(Variable.values, dist=sts.norm(mu, sigma), plot=plt.figure().add_subplot(111))*
    - *Fit_normal = sts.kstest(Variable,'norm',param)*

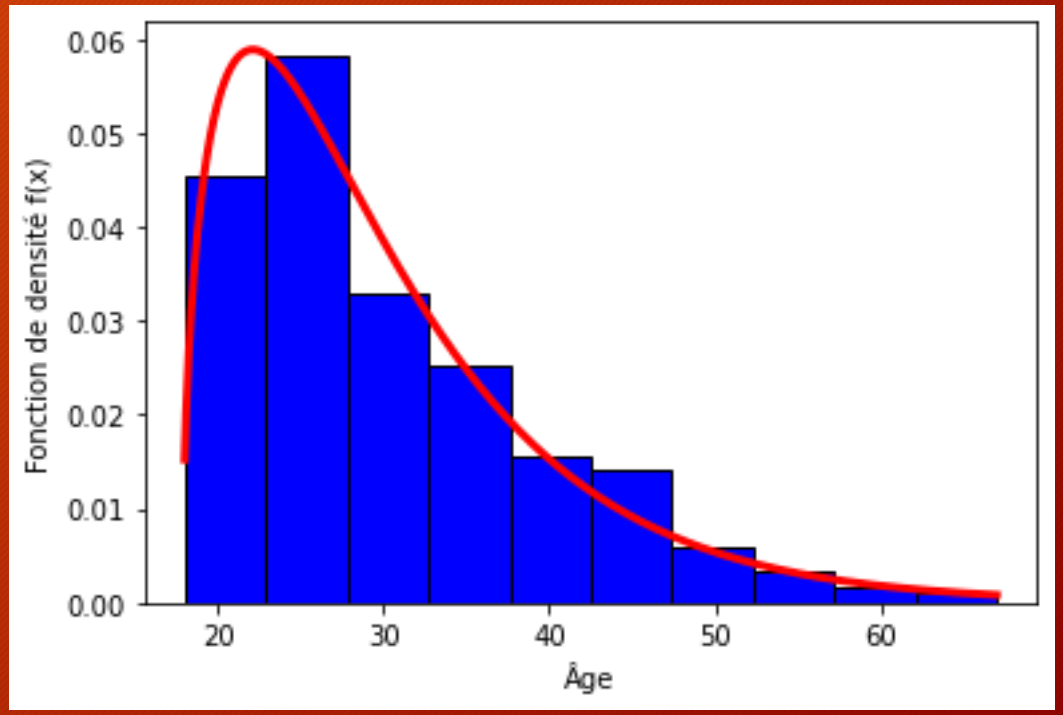




p-value basse: il y a donc une différence entre la distribution calibrée et une distribution normale

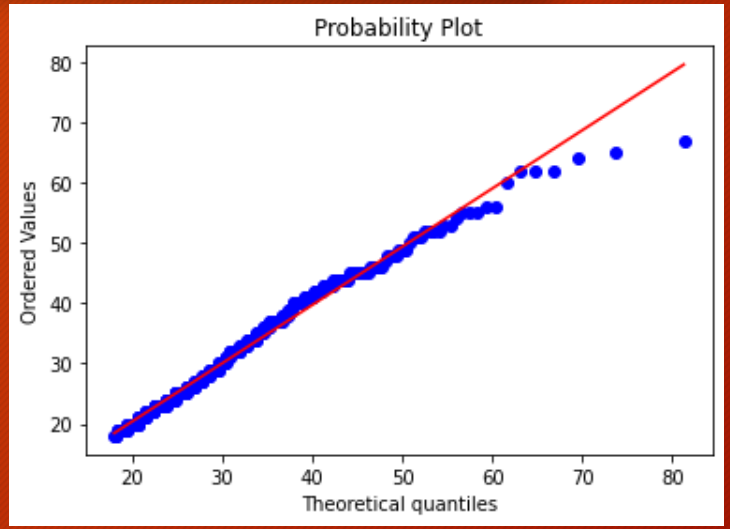
1- Distributions discrètes et continues définies

- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Essai pour calibrer (« fitter ») une distribution gamma
    - *d=1000*
    - *grille_x = np.linspace(Variable.min(), Variable.max(), d)*
    - *dx=(Variable.max()-(Variable.min()))/(d-1)*
    - *shape,loc,scale=sts.gamma.fit(Variable.values,loc=0.1)*
    - *param=sts.gamma.fit(Variable.values)*
    - *pdf = sts.gamma.pdf(grille_x,shape,loc,scale)*
    - *ax=Variable.plot.hist(density=True, bins = 10, color = 'blue', edgecolor = 'black')*
    - *ax.set_xlabel("Âge")*
    - *ax.plot(grille_x, pdf, linewidth=3, color = 'red')*
    - *ax.set_ylabel("Fonction de densité f(x)")*



1- Distributions discrètes et continues définies

- E10-1 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Essai pour calibrer (« fitter ») une distribution gamma
    - *sts.probplot(Variable.values, dist=sts.gamma(shape,loc,scale), plot=plt.figure().add_subplot(111))*
    - *Fit_gamma = sts.kstest(Variable.values,'gamma',param)*



Fit_gamma - Tuple (2 elements)

Inde	Type	Size	
0	float64	1	0.05384750411322797
1	float64	1	0.10191060839031796

p-value haute: il y a donc un moins de différence entre distribution calibrée et une distribution gamma

1- Distributions discrètes et continues définies

- Exercice L10 - #1
  - Vous avez les données d'une étude réalisée sur les habitudes de consommation (marketing). Les données ont été adaptées de la source originale.

ID	Âge	Statut Marital	Revenus	Enfants	Adolescents	Date	Temps depuis dernier achat	Vins ($/2sem)	Fruits ($/2sem
5524	63	Célibataire	58138	0	0	2012-09-04	58	635	88
2174	66	Célibataire	46344	1	1	2014-03-08	38	11	1
4141	55	Conjoint de fait	71613	0	0	2013-08-21	26	426	49
6182	36	Conjoint de fait	26646	1	0	2014-02-10	26	11	4
5324	39	Marié	58293	1	0	2014-01-19	94	173	43
7446	53	Conjoint de fait	62513	0	1	2013-09-09	16	520	42
965	49	Divorcé	55635	0	1	2012-11-13	34	235	65
6177	35	Marié	33454	1	0	2013-05-08	32	76	10
4855	46	Conjoint de fait	30351	1	0	2013-06-06	19	14	0
5899	70	Conjoint de fait	5648	1	1	2014-03-13	68	28	0

https://www.kaggle.com/rodsaldanha/arketing-campaign?select=marketing_campaign.xlsx

https://pxhere.com/fr/photo/1440159

1- Distributions discrètes et continues définies

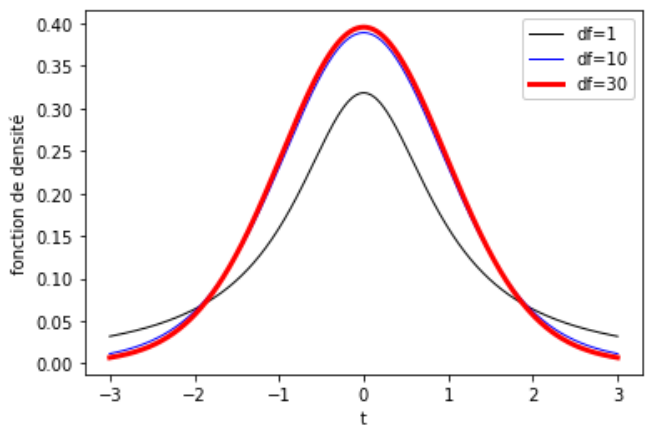
- Exercice L10 - #1
  - Pour la variable associée aux ventes en ligne (« Achats web »), tenter de calibrer 3 distributions: normale, exponentielle et gamma.
  - Vérifier s'il y a un bon « fit ».
  - Laquelle semble le mieux fonctionner?

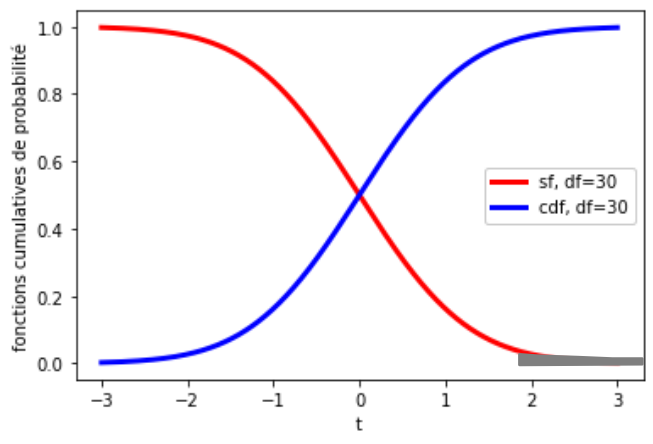
https://pxhere.com/fr/photo/1440159

2- Distributions pour l'inférence statistique

Distribution t (Loi de « Student »)

- Basée sur une loi gamma: a une forme de cloche comme la loi normale
- Cependant la forme dépend du nombre de degrés de liberté (« df »), qui est fonction du nombre d'observations (souvent « n-1 »)
- Très utilisée pour les tests d'hypothèses et les intervalles de confiance (ex: sur des paramètres, des moyennes)
- Est préférée quand la variance de la population (par rapport à un échantillon) n'est pas connue, ou pour de petits échantillons (n < 30)
- Pour un nombre d'observations très grand, elle tend vers la distribution normale



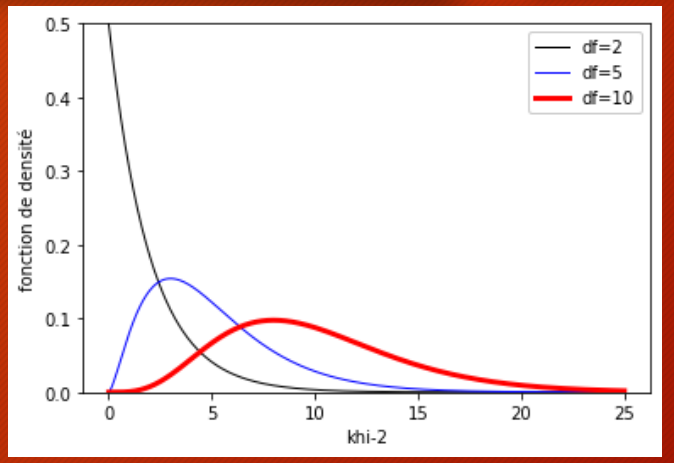


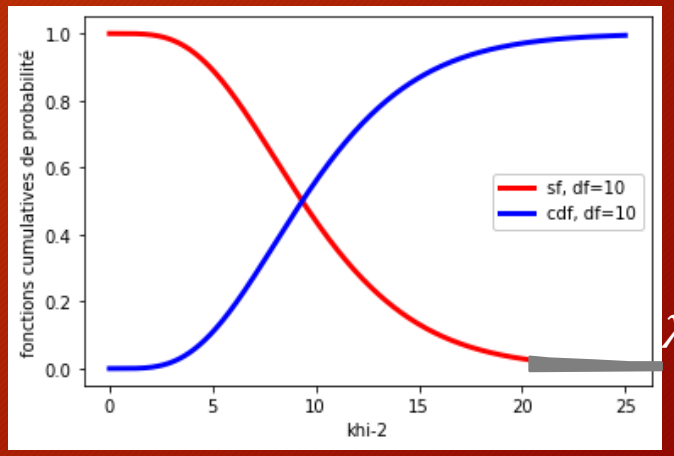
$t_{\alpha/2,\mathrm{df}}$

2- Distributions pour l'inférence statistique

Distribution du khi-carré

- Permet d'évaluer comment des données s'éloignent de leurs valeurs espérées (usuelles).
- Sa forme dépend également du nombre de degrés de liberté (« df »), qui est fonction du nombre d'observations (souvent « n-1 »)
- Souvent utilisée dans un test d'hypothèse, pour comparer la variance d'un échantillon de données à la variance espérée (connue).
- Est utilisée également pour établir un intervalle de confiance autour de la variance d'un échantillon.



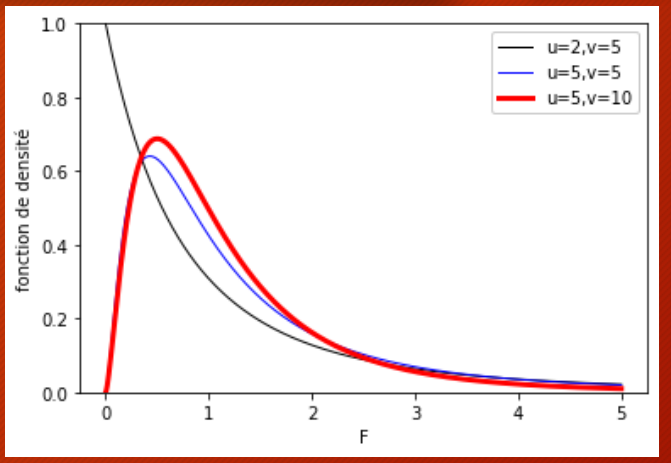


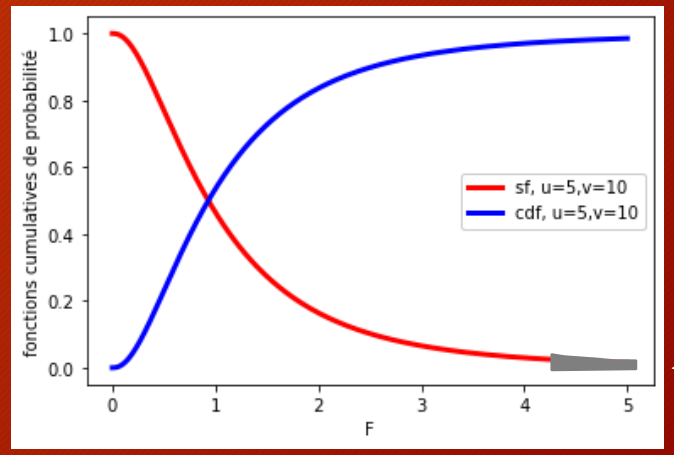
$\chi^2_{\alpha/2,\mathrm{df}}$

2- Distributions pour l'inférence statistique

Distribution F

- Permet de comparer des variances d'échantillons
- Permet de comparer notamment les moyennes de différents groupes pour une variable, quand celle-ci est répartie en différents traitements/catégories/groupes
  - ANOVA: permet de comparer la variabilité intra-groupe par rapport à la variabilité inter-groupe
  - Permet de savoir à quel point les différences entre moyennes de groupes sont significatives (par rapport au hasard)
- Sa forme dépend également du nombre de degrés de liberté (« df ») des variances considérées (ex: u et v), qui sont fonction du nombre d'observations ou du nombre de groupes/paramètres
- Peut être utilisée pour comparer la variance de deux échantillons.





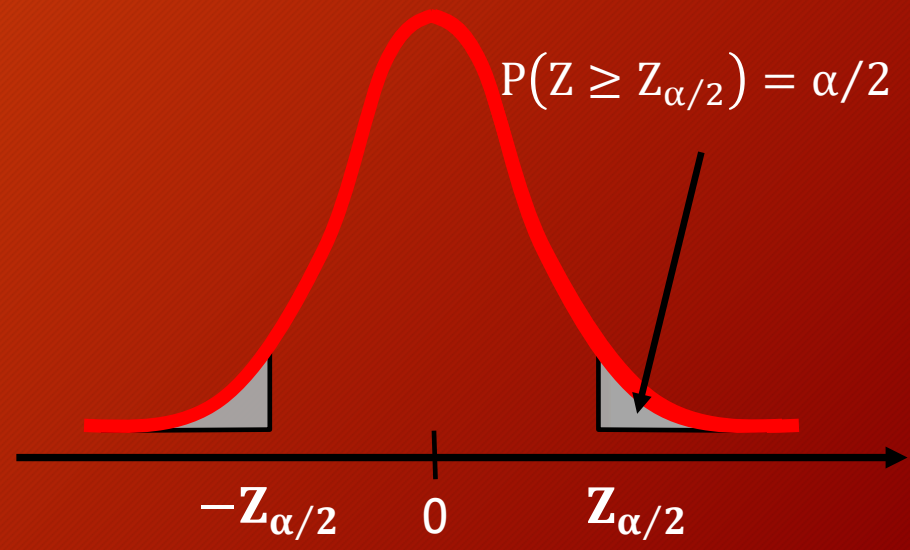
$F_{\alpha/2,u,v}$

3- Tests d'hypothèses

- Les tests d'hypothèse reposent sur le fait que certaines statistiques (ex: les moyennes réduites) sont normalement distribuées
- Les tests d'hypothèses comportent souvent une hypothèse nulle H0 (ex: pas de différence entre 2 valeurs), et une hypothèse requérant une preuve forte (à un niveau de confiance α souvent de 5% ou de 1%).
  - H0 (hypothèse nulle): Pas de différence entre des résultats (ex: dû à la chance)
  - H1: Différence significative entre des résultats

p-value: probabilité de ne pas pouvoir rejeter H0

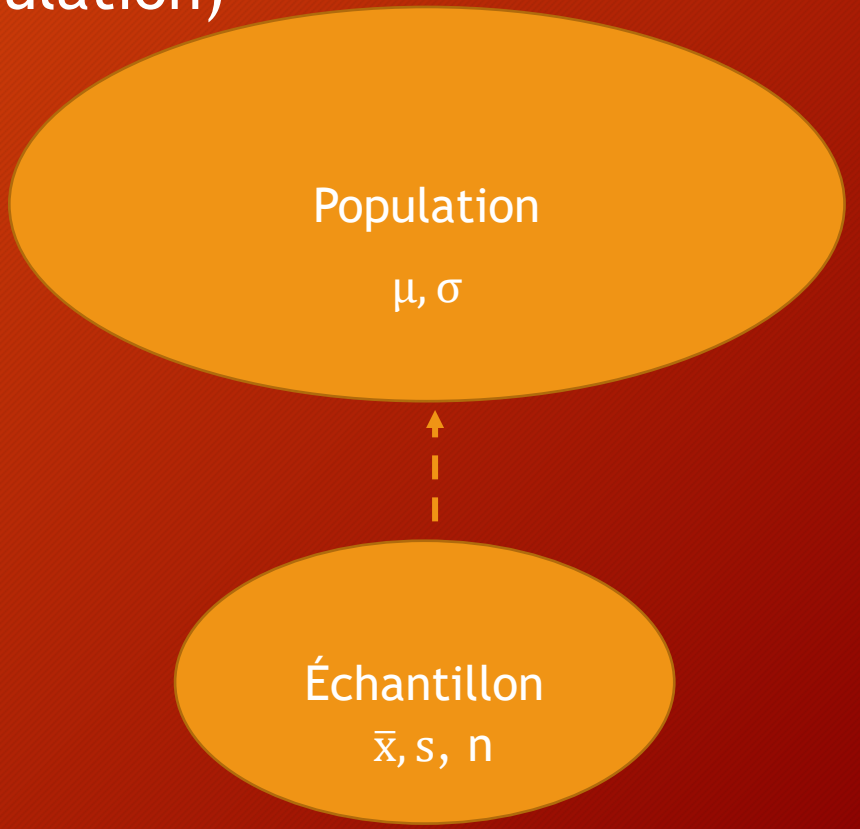
Plus cette valeur est petite (ex: en bas de 0.05), plus l'hypothèse H1 devient probable (et plus la différence devient significative)



3- Tests d'hypothèses – Moyenne d'une distribution

- Pour un test d'hypothèse sur un échantillon (vs population)
  - Population normale
  - Condition: σ^2 connue
  - Hypothèses:
    - H0 (hypothèse nulle): $\mu = \mu_0$
    - H1: $\mu \neq \mu_0$

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$



Rejette H0 (il y a une différence) si..

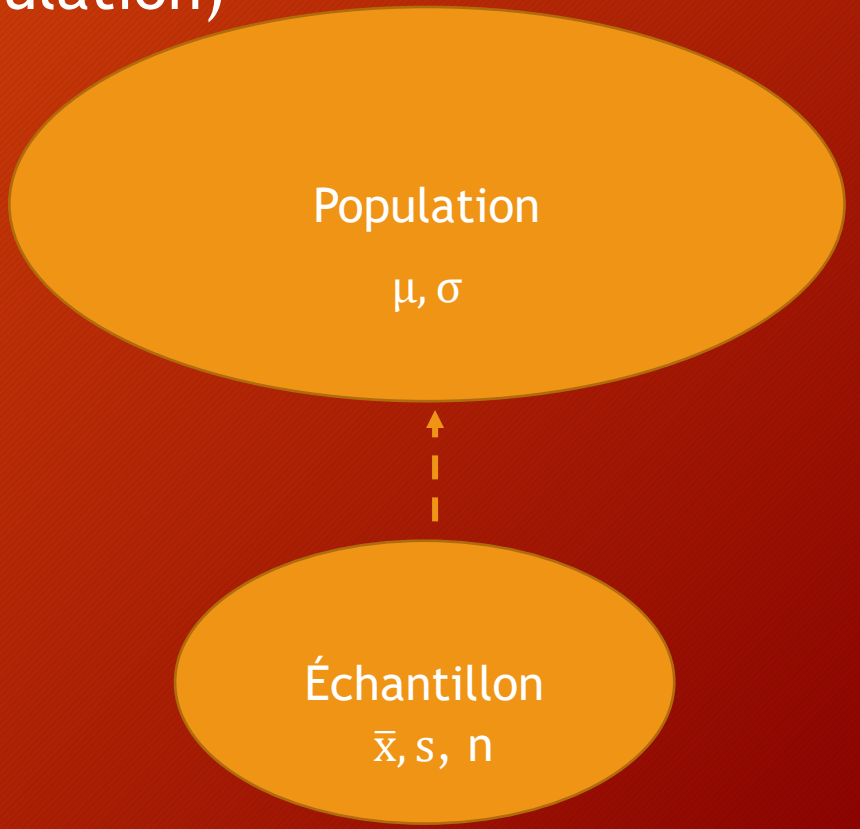
$\text{abs}(Z_0) > Z_{\alpha/2}$ ou CV

$p - value < \alpha\ (ex: 0.05\ ou\ 5\%)$

3- Tests d'hypothèses – Moyenne d'une distribution

- Pour un test d'hypothèse sur un échantillon (vs population)
  - Population normale
  - Condition: σ^2 inconnue
  - Hypothèses:
    - H0 (hypothèse nulle): $\mu = \mu_0$
    - H1: $\mu \neq \mu_0$

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



Rejette H0 (il y a une différence) si..

$abs(t_0) > t_{\alpha/2,n-1}$ ou CV

$p - value < \alpha\ (ex: 0.05\ ou\ 5\%)$

3- Tests d'hypothèses – Moyenne d'une distribution

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)

Âge	Poids (kg)	Grandeur (cm	Genre
21	65.6	174	Homme
23	71.8	175.3	Homme
28	80.7	193.5	Homme
23	72.6	186.5	Homme
22	78.8	187.2	Homme
21	74.8	181.5	Homme
26	86.4	184	Homme
27	78.4	184.5	Homme
23	62	175	Homme
21	81.6	184	Homme

- Source des données: Heinz G, Peterson LJ, Johnson RW, Kerk CJ. 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education 11(2)

https://www.openintro.org/data/index.php?data=bdims

3- Tests d'hypothèses – Moyenne d'une distribution

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Exemple: deux autres études indépendantes ont été réalisées sur le sujet dans la population générale:
    - Cas 1 - Poids moyen des hommes de 84 kg, variance = 100 kg^2
    - Cas 2 - Poids moyen des hommes de 79 kg
  - Est-ce qu'il y a une différence significative entre le poids moyen de ces études et le poids moyens obtenus dans le jeu de données?

3- Tests d'hypothèses – Moyenne d'une distribution

- E10-2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 1
  - Les librairies et les données sont chargées dans Spyder
    - *import numpy as np*
    - *import pandas as pd*
    - *import matplotlib.pyplot as plt*
    - *import scipy.stats as sts*
    - *import statsmodels.api as stm*
    - *import statsmodels.stats.weightstats as ws*
    - *import math*

    - *donnee = pd.read_csv('PersonnesActivesv0r2.csv')*
    - *stats=donnee.describe()*
    - *dimensions=donnee.shape*
    - *nomsvariables = pd.DataFrame(donnee.columns)*

donnee - DataFrame

Index	Âge	Poids (kg)	Grandeur (cm)	Genre
0	21	65.6	174	Homme
1	23	71.8	175.3	Homme
2	28	80.7	193.5	Homme
3	23	72.6	186.5	Homme
4	22	78.8	187.2	Homme
5	21	74.8	181.5	Homme
6	26	86.4	184	Homme
7	27	78.4	184.5	Homme
8	23	62	175	Homme
9	21	81.6	184	Homme
10	23	76.6	180	Homme
11	22	83.6	177.8	Homme

3- Tests d’hypothèses – Moyenne d’une distribution

- E10-2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 1
  - Aller chercher les données d’intérêt et calculer les statistiques sur ces données
    - *PoidsHommes=donnee[(donnee["Genre"] == "Homme")]["Poids (kg)"]*
    - *stats_PoidsHommes=PoidsHommes.describe()*
    - *n_H=PoidsHommes.shape[0]*
    - *X_barre_H_Poids=PoidsHommes.mean()*
    - *s_H_Poids=PoidsHommes.std()*

stats_PoidsHommes - S

Index	Poids (kg)
count	247
mean	78.1445
std	10.5129
min	53.9
25%	70.95
50%	77.3
75%	85.5
max	116.4

3- Tests d'hypothèses – Moyenne d'une distribution

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 1
  - Il est alors possible d'effectuer le test d'hypothèse (variance connue)
    - *mhu0H=84*
    - *sigmaHPoids=math.sqrt(100)*
    - *Z0=(X_barre_H_Poids-mhu0H)/(sigmaHPoids/(math.sqrt(n_H)))*
    - *CV1 = sts.norm.isf(0.05/2)*
    - *p_value_calc1=sts.norm.sf(abs(Z0))*2*

$\mathrm{abs}(Z_0) = 9.2 > CV1 = 1.96$

$p - value$ = 3.494e-20 < 0.05

Rejette H0

La différence entre la moyenne proposée par l'étude (84 kg) et celle du jeu de données (78.14 kg) est significative

3- Tests d'hypothèses – Moyenne d'une distribution

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 2
  - Il est alors possible d'effectuer le test d'hypothèse (variance inconnue)
    - *mhu0H=79*
    - *t0=(X_barre_H_Poids-mhu0H)/(s_H_Poids/(math.sqrt(n_H)))*
    - *CV2 = sts.t.isf(0.05/2,n_H-1)*
    - *p_value_calc2=sts.t.sf(abs(t0),df=(n_H-1))*2*
    - *pvalue2=sts.ttest_1samp(PoidsHommes,79)*

$\mathrm{abs}(t_0) = 1.279$ < CV2 = 1.96

$p - value$ = 0.20 > 0.05

Ne peut pas rejeter H0

La différence entre la moyenne proposée par l'étude (79 kg) et celle du jeu de données (78.14 kg) n'est pas significative

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

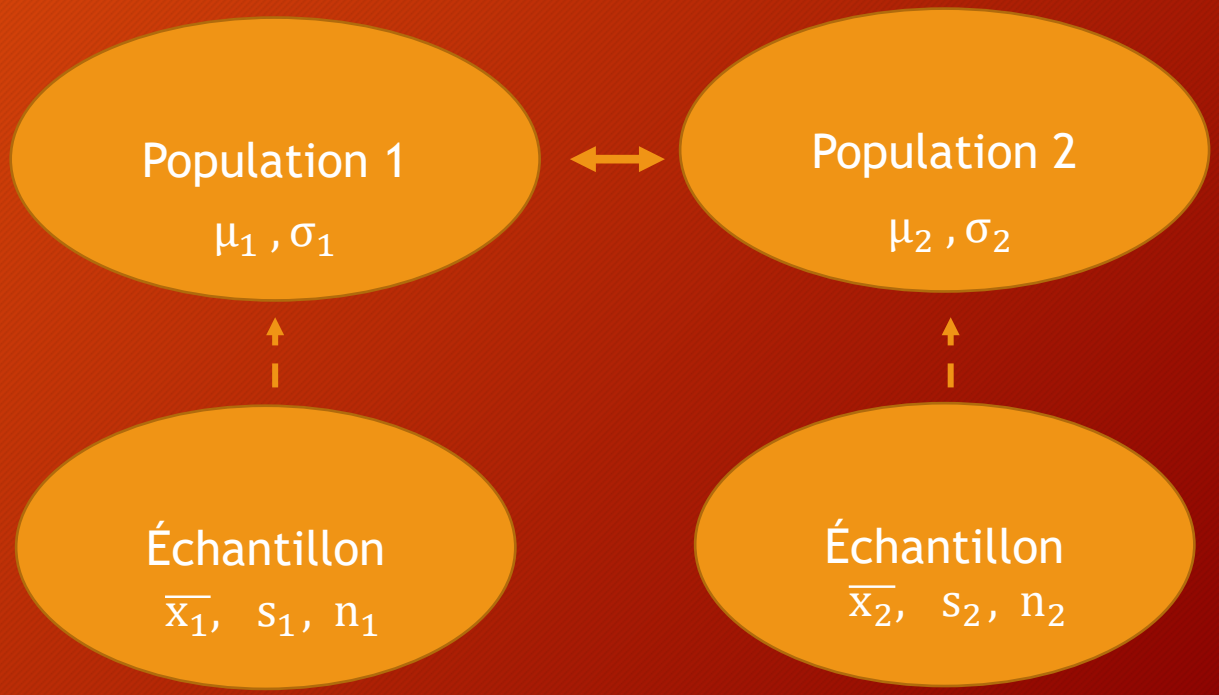
- Pour un test d'hypothèse sur deux échantillons/populations
  - Populations normales indépendantes
  - Condition: $\sigma_1{}^2$ et $\sigma_2{}^2$ connues
  - Hypothèses:
    - H0 (hypothèse nulle): $\mu_1 = \mu_2$
    - H1: $\mu_1 \neq \mu_2$

$$Z_0 = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}}$$

Rejette H0 (il y a une différence) si..

$\text{abs}(Z_0) > Z_{\alpha/2}$ ou CV

$p - value < \alpha\ (ex: 0.05\ ou\ 5\%)$



3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

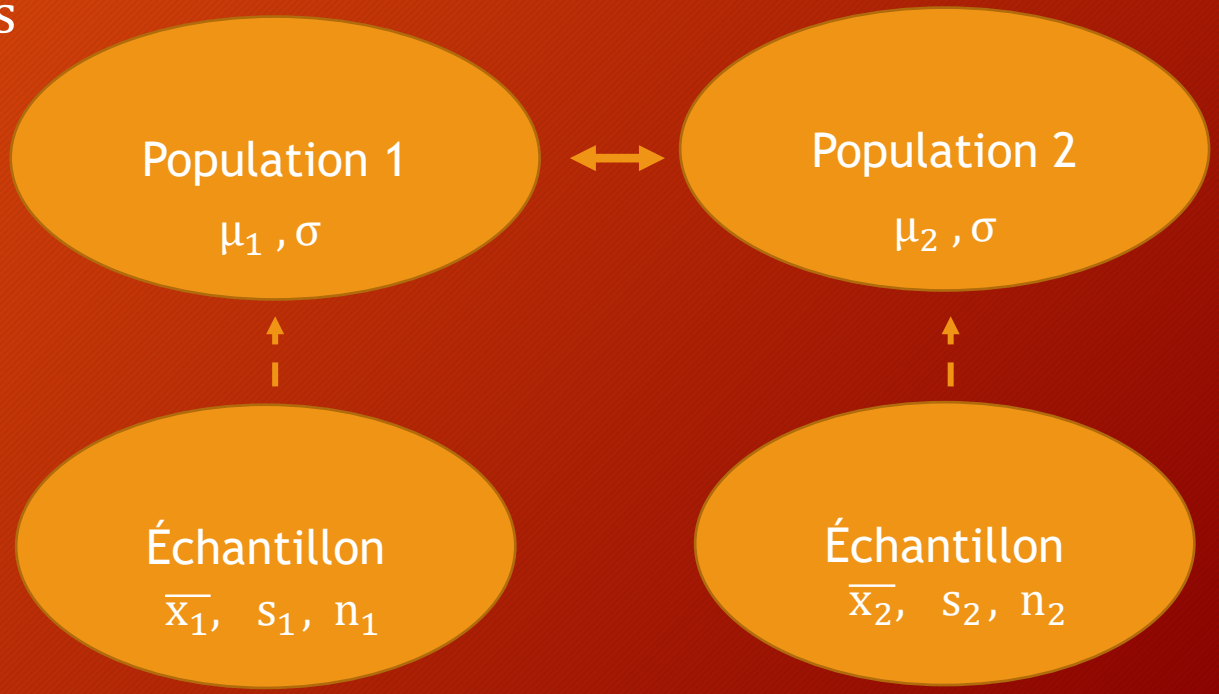
- Pour un test d'hypothèse sur deux échantillons/populations
  - Populations normales indépendantes
  - Condition: $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$ mais inconnues
  - Hypothèses:
    - H0 (hypothèse nulle): $\mu_1 = \mu_2$
    - H1: $\mu_1 \neq \mu_2$

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_P = \sqrt{\frac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}}$$

Rejette H0 (il y a une différence) si..

$abs(t_0) > t_{\alpha/2, n_1+n_2-2}$ ou CV

$p - value < \alpha\ (ex: 0.05\ ou\ 5\%)$



3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

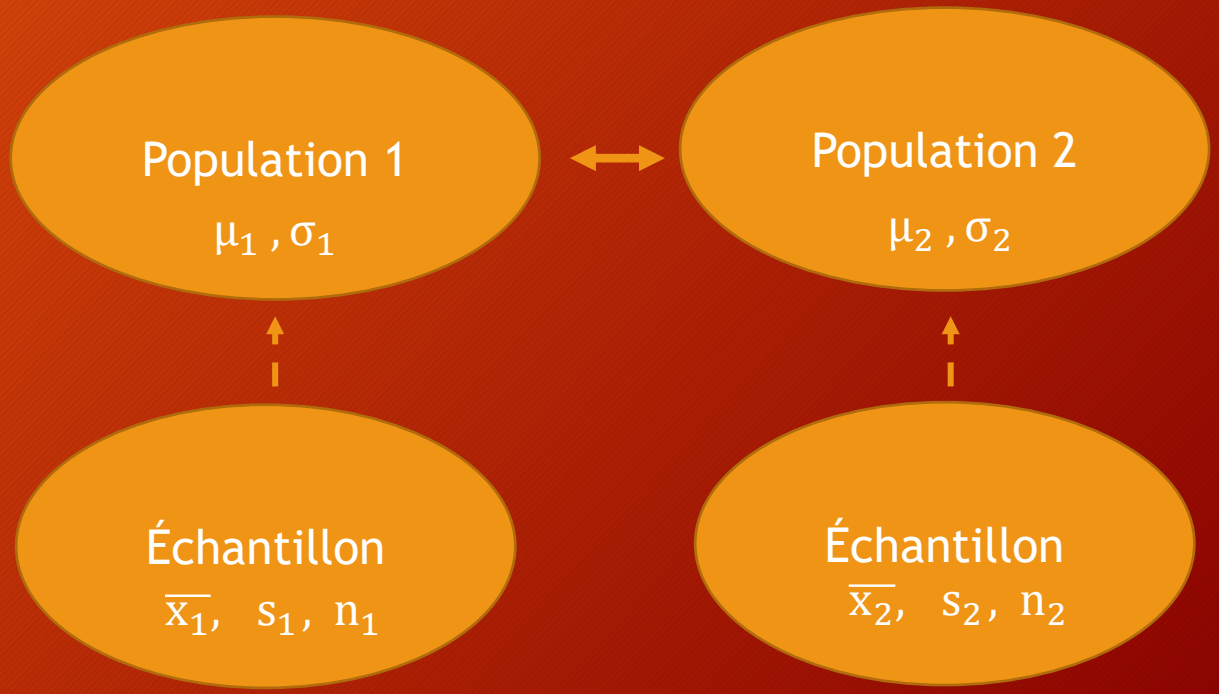
- Pour un test d'hypothèse sur deux échantillons/populations
  - Populations normales indépendantes
  - Condition: $\sigma_1{}^2 \neq \sigma_2{}^2$ et inconnues
  - Hypothèses:
    - H0 (hypothèse nulle): $\mu_1 = \mu_2$
    - H1: $\mu_1 \neq \mu_2$

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}}} \qquad \nu = \frac{\left(\frac{s_1{}^2}{n_1} + \frac{s_2{}^2}{n_2}\right)^2}{\frac{\left(s_1{}^2/n_1\right)^2}{n_1 + 1} + \frac{\left(s_2{}^2/n_2\right)^2}{n_2 + 1}} - 2$$

Rejette H0 (il y a une différence) si..

$\text{abs}(t_0) > t_{\alpha/2,\nu}$ ou CV

$p - value < \alpha\ (ex: 0.05\ ou\ 5\%)$



3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)

Âge	Poids (kg)	Grandeur (cm	Genre
21	65.6	174	Homme
23	71.8	175.3	Homme
28	80.7	193.5	Homme
23	72.6	186.5	Homme
22	78.8	187.2	Homme
21	74.8	181.5	Homme
26	86.4	184	Homme
27	78.4	184.5	Homme
23	62	175	Homme
21	81.6	184	Homme

- Source des données: Heinz G, Peterson LJ, Johnson RW, Kerk CJ. 2003. Exploring Relationships in Body Dimensions. Journal of Statistics Education 11(2)

https://www.openintro.org/data/index.php?data=bdims

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Voudrait savoir s'il y a, au point de vue statistique, une différence significative de grandeur (en moyenne) entre les hommes et les femmes dans ce jeu de données (différence entre ces 2 populations)
  - Dans un premier temps, nous allons assumer que les variances sont connues
    - Un médecin vous dit que par expérience, chez les populations d'athlètes qu'il a observées:

$$\sigma_H = 7.4 \text{ cm} \qquad \sigma_F = 6.2 \text{ cm}$$

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles)
  - Cas 3 - Dans un premier temps, nous allons assumer que les variances sont connues
    - Un médecin vous dit que par expérience, chez les populations d'athlètes qu'il a observées:

      $\sigma_H = 7.4\ \text{cm}$ $\sigma_F = 6.2\ \text{cm}$
  - Cas 4 – Nous allons assumer que les variances sont inconnues, mais à peu près égales
  - Cas 5 – Nous allons assumer qu'on ne peut pas supposer que les variances sont égales (et qu'elles sont inconnues)

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10-2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 3
  - Les librairies et les données sont chargées dans Spyder
    - *GrandeursHommes=donnee[(donnee["Genre"] == "Homme")]["Grandeur (cm)"]*
    - *GrandeursFemmes=donnee[(donnee["Genre"] == "Femme")]["Grandeur (cm)"]*
    - *Stats_H=GrandeursHommes.describe()*
    - *Stats_F=GrandeursFemmes.describe()*

    - *X_barre_H_Grand=GrandeursHommes.mean()*
    - *X_barre_F_Grand=GrandeursFemmes.mean()*
    - *s_H_Grand=GrandeursHommes.std()*
    - *s_F_Grand=GrandeursFemmes.std()*
    - *n_H=GrandeursHommes.shape[0]*
    - *n_F=GrandeursFemmes.shape[0]*

Hommes

Stats_H - Series

Index	Grandeur (cm)
count	247
mean	177.745
std	7.18363
min	157.2
25%	172.9
50%	177.8
75%	182.65
max	198.1

Femmes

Stats_F - Series

Index	Grandeur (cm)
count	260
mean	164.872
std	6.5446
min	147.2
25%	160
50%	164.5
75%	169.5
max	182.9

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 3
  - Il est alors possible d'effectuer le test d'hypothèse (variances connues)
    - *sigmaH=7.4*
    - *sigmaF=6.2*
    - *Z0=(X_barre_H_Grand-X_barre_F_Grand)/math.sqrt(sigmaH**2/n_H+sigmaF**2/n_F)*
    - *CV3 = sts.norm.isf(0.05/2)*
    - *p_value_calc3=sts.norm.sf(abs(Z0))*2*
    - *print(p_value_calc3)*

$\mathrm{abs}(Z_0) = 21.17 > \mathrm{CV3} = 1.96$

$p - value$ = 1.587e-99 < 0.05

Rejette H0

La différence entre les deux moyennes (grandeurs des hommes et des femmes dans ce jeu de données) est significative

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 4
  - Il est alors possible d'effectuer le test d'hypothèse (variances inconnues mais égales)
    - *Sp=math.sqrt(((n_H-1)*s_H_Grand**2+(n_F-1)*s_F_Grand**2)/(n_H+n_F-2))*
    - *t0=(X_barre_H_Grand-X_barre_F_Grand)/(Sp*math.sqrt(1/n_H+1/n_F))*
    - *CV4 = sts.t.isf(0.05/2,n_H+n_F-2)*
    - *p_value_calc4=sts.t.sf(abs(t0),df=(n_H+n_F-2))*2*
    - *pvalue4 = sts.ttest_ind(GrandeursHommes,GrandeursFemmes)*

$\mathrm{abs}(t_0) = 21.11 > \mathrm{CV4} = 1.965$

$p - value$ = 2.234e-71 < 0.05

Rejette H0

La différence entre les deux moyennes (grandeurs des hommes et des femmes dans ce jeu de données) est significative

3- Tests d'hypothèses – Comparaison de moyennes (2 échantillons)

- E10–2 - Jeu de données (étude) sur les indices biométriques des personnes actives (données partielles) – Cas 5
  - Il est alors possible d'effectuer le test d'hypothèse (variances inconnues mais égales)
    - *v=(s_H_Grand**2/n_H+s_F_Grand**2/n_F)**2/((s_H_Grand**2/n_H)**2/(n_H+1)+(s_F_Grand**2/n_F)**2/(n_F+1))-2*
    - *t0=(X_barre_H_Grand-X_barre_F_Grand)/math.sqrt(s_H_Grand**2/n_H+s_F_Grand**2/n_F)*
    - *CV5 = sts.t.isf(0.05/2,v)*
    - *p_value_calc5=sts.t.sf(abs(t0),df=(v))*2*

$\mathrm{abs}(t_0) = 21.06 > \mathrm{CV5} = 1.965$

$p - value$ = 7.828e-71 < 0.05

Rejette H0

La différence entre les deux moyennes (grandeurs des hommes et des femmes dans ce jeu de données) est significative

3- Tests d'hypothèses

- Exercice L10 - #2
  - Vous avez les données d'une étude réalisée sur les habitudes de consommation (marketing). Les données ont été adaptées de la source originale.

ID	Âge	Statut Marital	Revenus	Enfants	Adolescents	Date	Temps depuis dernier achat	Vins ($/2sem)	Fruits ($/2sem
5524	63	Célibataire	58138	0	0	2012-09-04	58	635	88
2174	66	Célibataire	46344	1	1	2014-03-08	38	11	1
4141	55	Conjoint de fait	71613	0	0	2013-08-21	26	426	49
6182	36	Conjoint de fait	26646	1	0	2014-02-10	26	11	4
5324	39	Marié	58293	1	0	2014-01-19	94	173	43
7446	53	Conjoint de fait	62513	0	1	2013-09-09	16	520	42
965	49	Divorcé	55635	0	1	2012-11-13	34	235	65
6177	35	Marié	33454	1	0	2013-05-08	32	76	10
4855	46	Conjoint de fait	30351	1	0	2013-06-06	19	14	0
5899	70	Conjoint de fait	5648	1	1	2014-03-13	68	28	0

https://www.kaggle.com/rodsaldanha/arketing-campaign?select=marketing_campaign.xlsx

https://pxhere.com/fr/photo/1440159

3- Tests d’hypothèses

- Exercice L10 - #2
  - Vous voulez comparer les revenus des personnes ayant acheté pour plus de 50$ en joaillerie (en 2 semaines) par rapport aux autres personnes.
  - Créez 2 séries de données pour la variable « Revenus » (selon la condition d’avoir acheté pour plus de 50$ de joaillerie ou non).
  - Question 1:
    - Vérifiez si les 2 séries de données (après traitement) sont presque normalement distribuées. Si c’est le cas, calibrer une distribution normale sur chaque série.

https://pxhere.com/fr/photo/1440159

3- Tests d'hypothèses

- Exercice L10 - #2
  - Question 2:
    - Un collègue (Marcus) suggère que les personnes achetant pour plus de 50$ de joaillerie en 2 semaines gagnent en moyenne 100 000$.
    - Une autre collègue (Chantale) suggère que les personnes achetant pour plus de 50$ de joaillerie en 2 semaines gagnent en moyenne 66 500$.
    - Quel estimé de la moyenne serait statistiquement valable, considérant les données (à $\alpha = 5\%$)?

https://pxhere.com/fr/photo/1440159

3- Tests d’hypothèses

- Exercice L10 - #2
  - Question 3:
    - Vérifiez si statistiquement il y a une différence dans la moyenne des revenus entre les personnes qui achètent pour 50$ et plus de joaillerie et ceux qui achètent pour moins de 50$ (aux 2 semaines).

https://pxhere.com/fr/photo/1440159

4- Évaluation formative

- Problème L10 - #2
  - Barème
    - Télécharger les données du fichier CSV (5%)
    - Pré-traitement (5%)
    - Q1 (20%)
      - Calibration des fonctions normales (15%)
      - Vérification de la normalité (ex: QQ-plot) (5%)
    - Q2 (40%)
      - Calcul des statistiques/métriques pour les tests d'hypothèses (ex: Z_0, t_0) (20%)
      - Calcul des « p-values » (10%)
      - Interprétation des résultats (10%)

https://pxhere.com/fr/photo/1440159

4- Évaluation formative

- Problème L10 - #2
  - Barème
    - Q3 (30%)
      - Calcul des statistiques/métriques pour les tests d'hypothèses (ex: Z_0, t_0) (10%)
      - Calcul des « p-values » (10%)
      - Interprétation des résultats (10%)
    - Dans chaque cas: 80% pour la démarche (code, choix des fonctions), et 20% pour le résultat numérique.

https://pxhere.com/fr/photo/1440159

Références

- Médiagraphie
  - Probabilités et statistiques pour ingénieurs (éd. française), Chenelière Éducation (2005), par Hines, Montgomery, Goldsman et Borror
  - Practical Statistics for Data Scientists: 50 Essential Concepts, May 28th 2017 by Peter Bruce (Author), Andrew Bruce (Author).
- Sites web
  - https://pygot.wordpress.com/2018/06/28/hypothesis-testing-in-python/
  - https://blog.minitab.com/en/understanding-statistics-and-its-application/what-should-i-do-if-my-data-is-not-normal-v2