Analyse de données

by
Rosalba Juarez Mosqueda

Data set :

Notre échantillon de 78 personnes

	Person	gender	Age	Height	pre.weight	Diet	weight6weeks
0	25		41	171	60	2	60.0
1	26		32	174	103	2	103.0
2	1	0	22	159	58	1	54.2
3	2	0	46	192	60	1	54.0
4	3	0	55	170	64	1	63.3
...	...	...	...	...	...	...	...
73	74	1	35	183	83	3	80.2
74	75	1	49	177	84	3	79.9
75	76	1	28	164	85	3	79.7
76	77	1	40	167	87	3	77.8
77	78	1	51	175	88	3	81.9

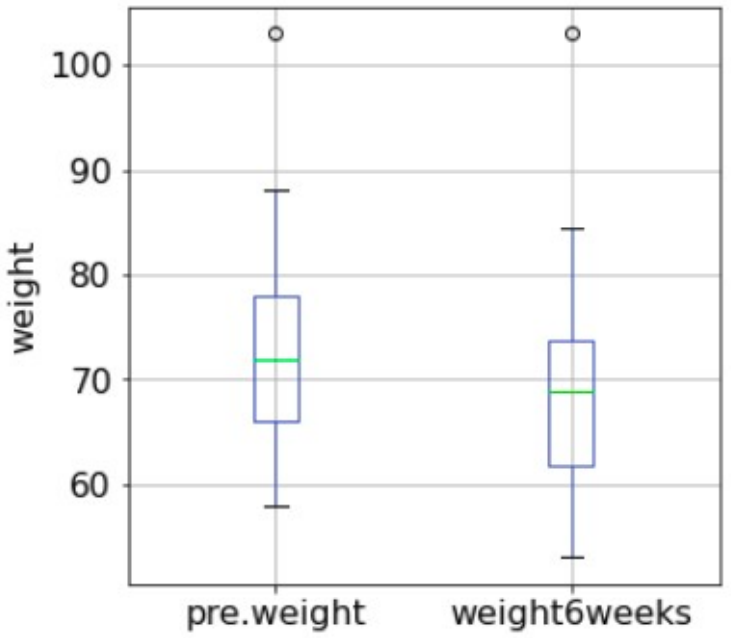
78 rows × 7 columns

Question 1

Premièrement, nous voulons savoir **si notre échantillon de 78 personnes a perdu du poids après 6 semaines de régime**, quel que soit le régime qu'elles suivent.

- $H_0 : \mu_{prépoids} = \mu_{poids6semaines}$
- $H_a : \mu_{prépoids} > \mu_{poids6semaines}$

Analyse exploratoire des données (EDA)



	Pre Weight	Weight after 6 weeks
Count	78.000000	78.000000
mean	**72.525641**	**68.680769**
Std	8.723344	8.924504
min	58.000000	53.000000
max	103.00000	103.000000
Q25 %	66.000000	61.850000
Q50 %	**72.000000**	**68.950000**
Q75 %	78.000000	73.825000

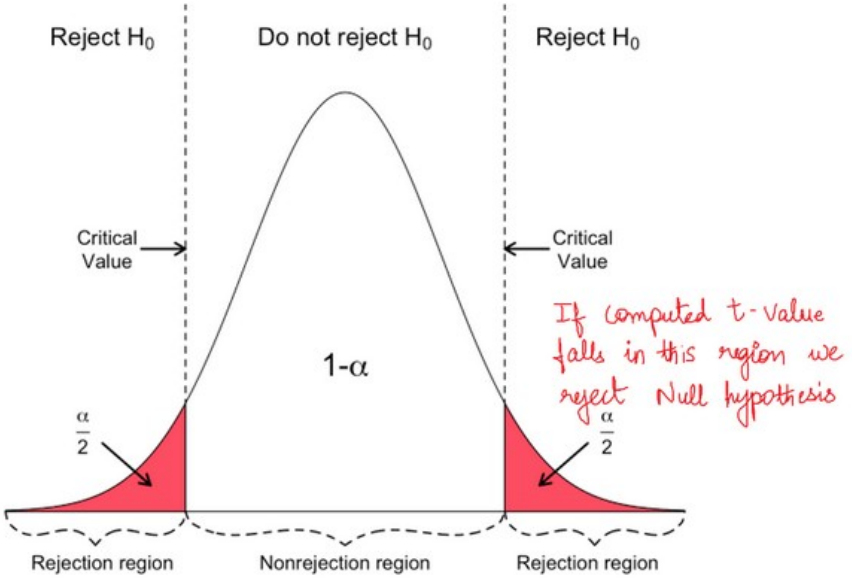
Les deux moyennes $\mu_{prepoids}$= 72.526 et $\mu_{poid6semaines}$ = 68.68, sont-elles significativement différentes au risque α = 0.05 ?

T-test : comparaison des moyennes de deux populations normales

$$H_0 : \mu_{prépoids} = \mu_{poids6semaines}$$

(avec avec un niveau de confiance de 95%)

$$t = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{(n1-1)\sigma1^2+(n2-1)\sigma2^2}{n1+n2-2}} \times \sqrt{\frac{1}{n1}+\frac{1}{n2}}}$$



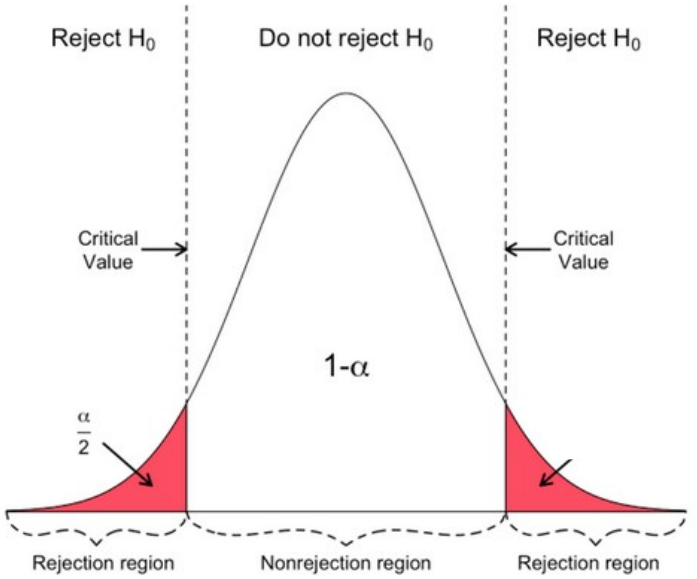
T-test : comparaison des moyennes de deux populations normales

$H_0 : \mu_{prépoids} = \mu_{poids6semaines}$

$H_a : \mu_{prépoids} > \mu_{poids6semaines}$

En utilisant les méthodes de scipy.stats :

	st.t.cdf(t,dof)	st.ttest_ind (a, b, 'two-sides')
t	2.720973 (input)	2.720973 (output)
dof = $(n_1+n_2)-2$	(78 - 78)-2 (input)	from dataframe (input)
Prob	0.996371	------
1-Prob	0.003629	------
p-value	**0.007258**	**0.007258**



Selon notre seuil établi pour un niveau de confiance de 95%, nous pouvons rejeter le H_0. Cela signifie que **s'il y a un effet, et que notre échantillon de 78 personnes a perdu du poids après 6 semaines de régime.**

Data set

Notre échantillon de 78 personnes

	Person	gender	Age	Height	pre.weight	Diet	weight6weeks
0	25		41	171	60	2	60.0
1	26		32	174	103	2	103.0
2	1	0	22	159	58	1	54.2
3	2	0	46	192	60	1	54.0
4	3	0	55	170	64	1	63.3
...	...	...	...	...	...	...	...
73	74	1	35	183	83	3	80.2
74	75	1	49	177	84	3	79.9
75	76	1	28	164	85	3	79.7
76	77	1	40	167	87	3	77.8
77	78	1	51	175	88	3	81.9

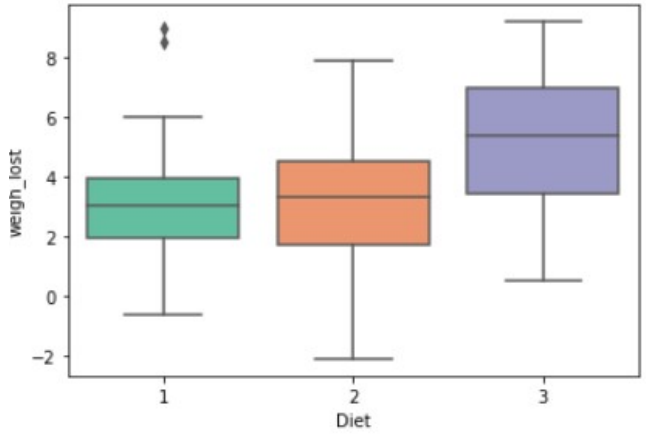
78 rows × 7 columns

Diet	Person	gender	Age	Height	pre.weight	weight6weeks
1	24	24	24	24	24	24
2	27	27	27	27	27	27
3	27	27	27	27	27	27

Analyse exploratoire des données (EDA)

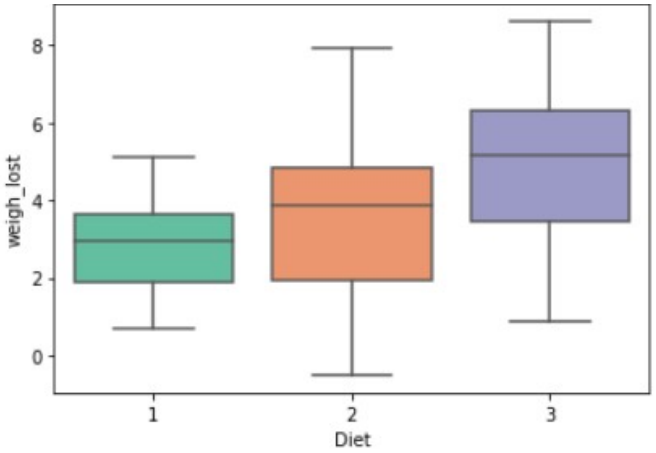
Nous voulons savoir **quel régime était le meilleur pour perdre du poids** ?

échantillon de 24 (1), 27 (2), et 27 (3) personnes



	Diet	mean	std	count
0	1	3.300000	2.240148	24
1	2	3.025926	2.523367	27
2	3	5.148148	2.395568	27

échantillon de 16 personnes par groupe



	Diet	mean	std	count
0	1	2.83750	1.314471	16
1	2	3.38125	2.294695	16
2	3	4.96875	2.071624	16

Les moyennes μ_1, μ_2 et μ_3 sont-elles significativement différentes au risque $\alpha = 0.05$?

ANOVA : comparaison des variances de deux populations normales

L'ANOVA à un facteur (one way) teste l'hypothèse nulle selon laquelle deux groupes ou plus ont la même moyenne de population.

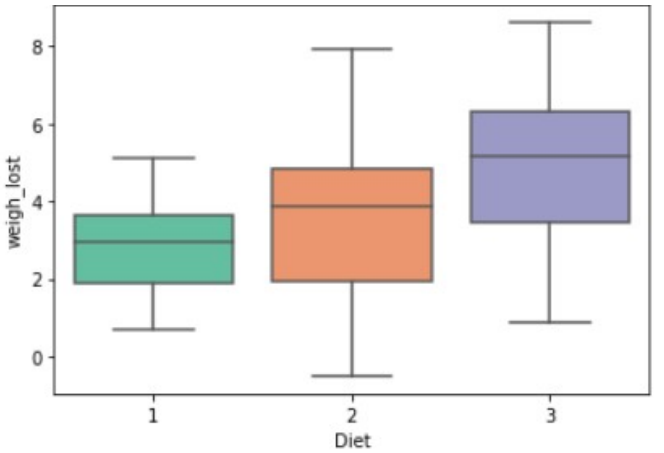
- $H_0 : \sigma^2_1 = \sigma^2_2 = \sigma^2_3$
-
- H_a : au moins une des variances est différent

ANOVA : comparaison des variances de deux populations normales

F = Variance entre les traitements / variance au sein des traitements

$$F = \sigma^2_B / \sigma^2_w$$

$\sigma^2_1 = 1.4967$ $\sigma^2_2 = 1.5885$ $\sigma^2_3 = 1.5478$



En utilisant les méthodes de scipy.stats et statsmodels :

	st.f_oneway (a,b,c)	ols(Y ~CX)
F	5.216122	5.216122
p-value	**0.009175**	**0.009175**

La pvalue de 0.0092, implique que, toujours **avec un niveau de confiance de 95 %, nous pouvons rejeter notre H_0 ; et par conséquent, s'il y a un effet significatif du type de régime sur la perte de poids,** même pour l'échantillon de 16 personnes par groupe.

NOTA : Les valeurs de F et p lors de l'utilisation de l'échantillon d'origine avec un ddl différent (D1 =24, D2= 27, D3 = 27) : F = 6.197447 et pvalue =0.003229

Test de Cochran : comparaison l'égalité de k variances de populations normales

Le nombre de degrés de liberté associé à chacune des estimations de ces variances doit être constant

$$g_{obs} = \sigma^2_{max} / \Sigma\sigma^2_i$$

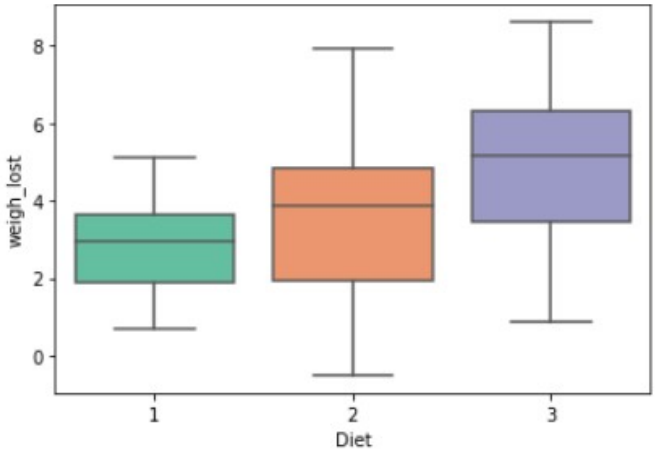
$$\sigma^2_2 = \sigma^2_{max} = 1.5885$$

$$\Sigma\sigma^2_i = + \sigma^2_1 + \sigma^2_2 + \sigma^2_3 = 4.6330$$

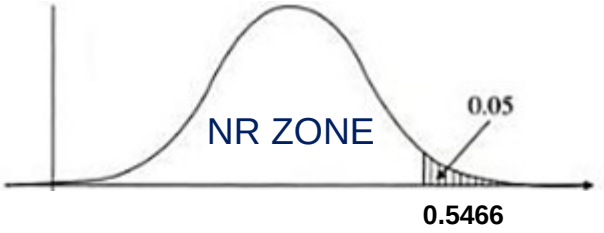
$$\mathbf{g_{obs} = 0.3428}$$

$$\mathbf{g_{(k=3,\ ddl=16)} = 0.5466}$$

$\sigma^2_1 = 1.4967$ $\sigma^2_2 = 1.5885$ $\sigma^2_3 = 1.5478$

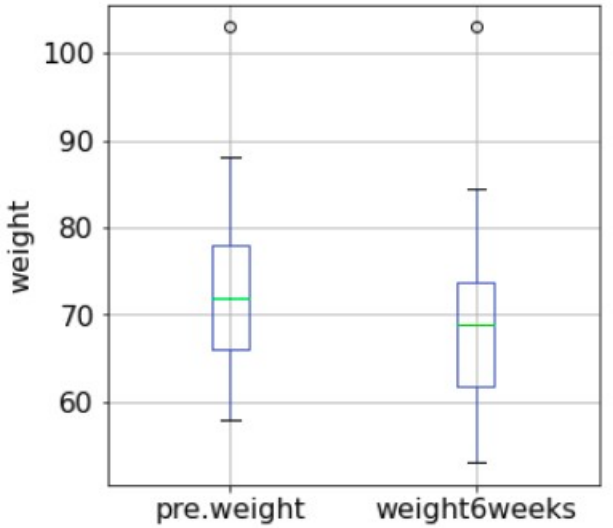


Puisque g_{obs} il tombe dans la zone de non-rejet, Avec un niveau de confiance de 95 %, nous ne pouvons pas rejeter notre H_0

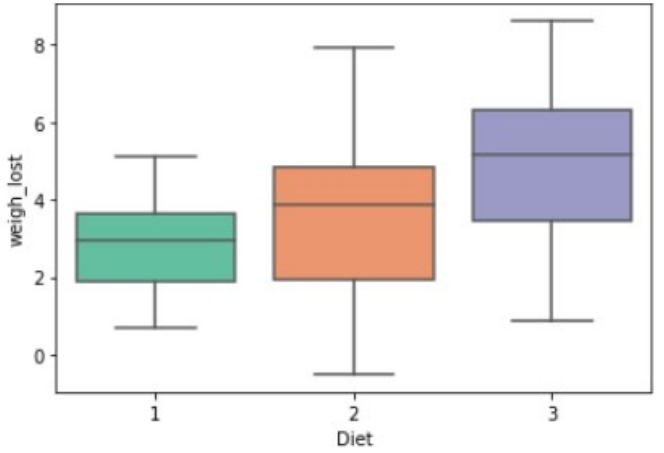


Conclusions

- Avec un niveau de confiance de 95%, **nous pouvons établir que l'échantillon de 78 personnes a perdu du poids après 6 semaines de régime.**



- Toujours avec un niveau de confiance de 95 %, **nous pouvons confirmer (sur la base de cet échantillon de 16 personnes par groupe) que il y a un régime (régime 3) mieux des deux autres .**



Merci !