L'analyse de la variance.

Dans cet article nous allons reprendre les fondements théoriques de l'**ANalyse de la VAriance** (ANOVA) à un facteur et présenter une courte application : comparaison des temps moyens de production. L'analyse de la variance **à un facteur** est une méthode très répandue pour comparer au moins **3 sous-échantillons** d'après leur moyenne. Pour ce faire, l'anova va comparer la variance inter- et intra-classe afin de **savoir si au moins deux des moyennes diffère significativement**. L'anova repose donc sur un test de Fisher-Snedecor de comparaison de variance. Pourquoi utiliser un test de variance pour étudier les différences de moyenne ? Simplement parce que la dispersion des données peut avoir deux origines :

- la **variabilité lié au facteur** : la variance factorielle ou inter-classe ;
- la **variabilité intrinsèque à chaque catégorie du facteur** : la variance résiduelle ou intra-classe - la part de variabilité restante une fois la variabilité factorielle soustraite à la variabilité totale.

Ainsi, on va chercher à savoir si la part de variance inter-classe est significativement supérieure à la part de variance intra-classe. On pourra alors conclure qu'au moins deux des moyennes sont globalement différente si la variabilité intra-class (résiduelle) est faible relativement à la part de variabilité inter-class (factorielle). Pour l'écrire différement, plus le rapport des variances inter- et intra-classe sera élevé plus on aura de chance de conclure que les moyennes de nos sous-échantillons sont signifcativement différentes de la moyenne générale. L'étude de la variabilité de nos sous-échantillons va donc nous permettre d'étudier les divergences de moyennes.

La statistique de test sera donc le rapport entre la variance inter-classe et la variance intra-classe. Pour établir la significativité statistique de ce test on va utiliser le test F de Fisher-Snedecor qui suit une loi de Fisher à $$K-1$$ et $$N-K$$ degrés de liberté.

$$ F_{k-1; N-k} = \frac{Variance_factorielle}{Variance_résiduelle} .$$

Après avoir fixé un risque de première espèce $$\alpha$$, en général $$5 %$$, si la statistique de test est supérieure à la valeur critique de la table de Fisher pour $$k-1$$ et $$N-k$$ degrés de liberté avec un seuil de significativité $$\alpha$$, on conclura qu'**au moins une des moyennes diffère significativement d'une autre**.

Le jeu des hypothèses est le suivant :

$$H_0$$: les moyennes des populations sont identiques ; $$H_1$$: Au moins deux populations parentes ne partagent pas la même moyenne.

Notez que ce test va être réalisé à partir d'un modèle linéaire. En effet, même si ca n'est pas évident comme ca, l'anova est un modèle linéaire de la forme suivante :

$$ Y_{ij} = \beta_{j} + U_{ij},$$ où $$Y_{ij}$$ est la réponse de l'observation $$i$$ du groupe $$j$$, $$\beta_{j}$$ le paramètre du groupe $$j$$ à estimer et $$U_{ij}$$ l'erreur associée à l'individu $$i$$ du groupe $$j$$. L'erreur est identiquement et indépendemment distribuée, $$U_{ij} \sim \mathcal{N}(0,,\sigma^{2})$$ - i.e. normalité et homoscedasticité sont requises en chaque point du facteur explicatif.

Précisément, $$\beta_{j}$$ correspond à la moyenne du groupe $$j$$ et $$U_{ij}$$ à l'écart entre l'observation $$i$$ et la moyenne du groupe $$j$$ auquel elle appartient. Avec ce modèle, pour un facteur à $$k$$ modalités, le jeu des hypothèses est donc le suivant :

- $$H_0$$: $$\beta_1 = \beta_2 = ... = \beta_k$$;
- $$H_1$$: $$\exists l \neq m : \beta_l \neq \beta_m$$.

Sous l'hypothèse nulle, le modèle est dit constant (ou blanc) et se distingue du modèle considéré puisque le paramètre $$\beta$$ n'est autre que la moyenne globale de la variable endogène. Alors, la statistique de test F va comparer les différences de valeurs prédites entre le modèle considéré et le modèle blanc par rapport à l'écart-type ; soit la somme des carrés factoriels (divisé par son degré de liberté, K-1) par rapport à la somme des carrés résiduels (divisé par son degré de liberté, n-k) ; soit le rapport entre la variance factorielle et la variance résiduelle.

> Pour plus de détail, consultez cet article.

Take home messages.

- L'anova est un modèle linéaire employé pour comparer les moyennes de plus de deux sous-échantillons (les sous-échantillons formés par une variable quantitative relativement à un facteur qualitatif à plus de 2 modalités) ;
- Les hypothèses de normalité, d'homoscédasticité et d'indépendance des résidus doivent être satisfaites (de façon équivalente, les données doivents être normalement distribuées et à variance homogène en chaque point du facteur) ;
- Pour étudier les différences de moyennes, on utilise le rapport de la variance inter- et intra-classe : la significativité statistique d'au moins une différence est établie avec un test F de Fisher-Snedecor.

Application : comparaison des temps moyens de production.

Aujourd'hui l'application va s'intéresser à la défaillance machine. Plus précisémment, on va chercher à savoir si les trois chaînes de production de notre usine fonctionne bien. Pour cela, on va comparer les temps de production de chacune d'entre elles. On a donc relevé 100 mesures du temps de production pour les 3 machines, ce de façon indépendante. On concluera à une défaillance machine si une machine a un temps de production moyen supérieur à une autre.

Notre application repose sur des données fictives dont voici un court aperçu.

```
    chaine_production   temps_production
0                   A          36.014622
1                   A          38.043625
2                   A          35.204505
3                   A          38.484877
4                   A          37.948544
```

On dénombre 300 observations indépendantes et deux variables, l'une quantitative continue (*temps_production*), l'autre qualitative à 3 modalités (*chaine_production*). Voici une courte description de ces variables : la variable *temps_production* à une moyenne de 56.87 minutes et un écart-type de 18.94 minutes. Le minimum est de 25.18 minutes et le maximum de 98.99 minutes. Concernant la variable *chaine_production*, chaque modalité compte 100 observations. L'ensemble de données ne comporte aucune valeur manquante, ni valeurs extrêmes.

```
count    300.000000
mean      56.871202
std       18.938976
min       25.181439
25%       38.636396
50%       55.862853
75%       75.503349
max       98.989768
Name: temps_production, dtype: float64
```

Passons à la réalisation de l'anova. En python vous avez deux possibilités pour appliquer une *anova*, soit vous utiliser une fonction *f_oneway* du module python *scipy.stats*, soit vous utiliser un modèle linéaire généralisé sur lequel vous appliquer l'anova avec le module *statsmodels*. Dans le premier cas, vous ne pourrez pas accéder à la **table d'analyse de la variance**, vous aurez comme retour la statistique de test et la pvalue. Voyons cela en pratique.

Dans un premier temps, nous allons ajuster le modèle linéaire afin de vérifier la statisfactions des hypothèses, afin de nous assurer de la validité des résultats.

```
# ajustement du modèle linéaire
model = ols('temps_production ~ C(chaine_production)', data=df).fit()

# si vous souhaitez afficher les résultats de la regression linéaire
# print(model.summary())

# on récupère les erreurs du modèle
residuals = model.resid
```

Vérification des conditions d'application de l'anova.

Normalité des résidus.

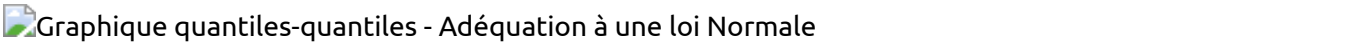
```
# on vérifie l'hypothèse de normalité avec un test d'adéquation
paramétrique

shapiro_test = stats.shapiro(residuals)
print("Test de Shapiro-Wilk :")
print(f"Statistique : {shapiro_test.statistic}, p-value :
{shapiro_test.pvalue}")

# on complète ce test d'une observation graphique
sm.qqplot(residuals, line='s')
plt.title('Q-Q Plot des résidus')
plt.show()
```

Voici les résultats de l'étude de la normalité des résidus.

```
Test de Shapiro-Wilk :
Statistique : 0.9907482862472534, p-value : 0.05559874325990677
```


Graphique quantiles-quantiles - Adéquation à une loi Normale

D'après le test de Sahpiro-Wilk nous n'avons pas suffisamment d'évidence pour rejeter l'hypothèse nulle de normalité des résidus. Le graphique quantiles-quantiles met en évidence la présence de quelques outliers ; globalement, l'adéquation à la loi Normale est bonne.

Homoscédasticité des résidus.

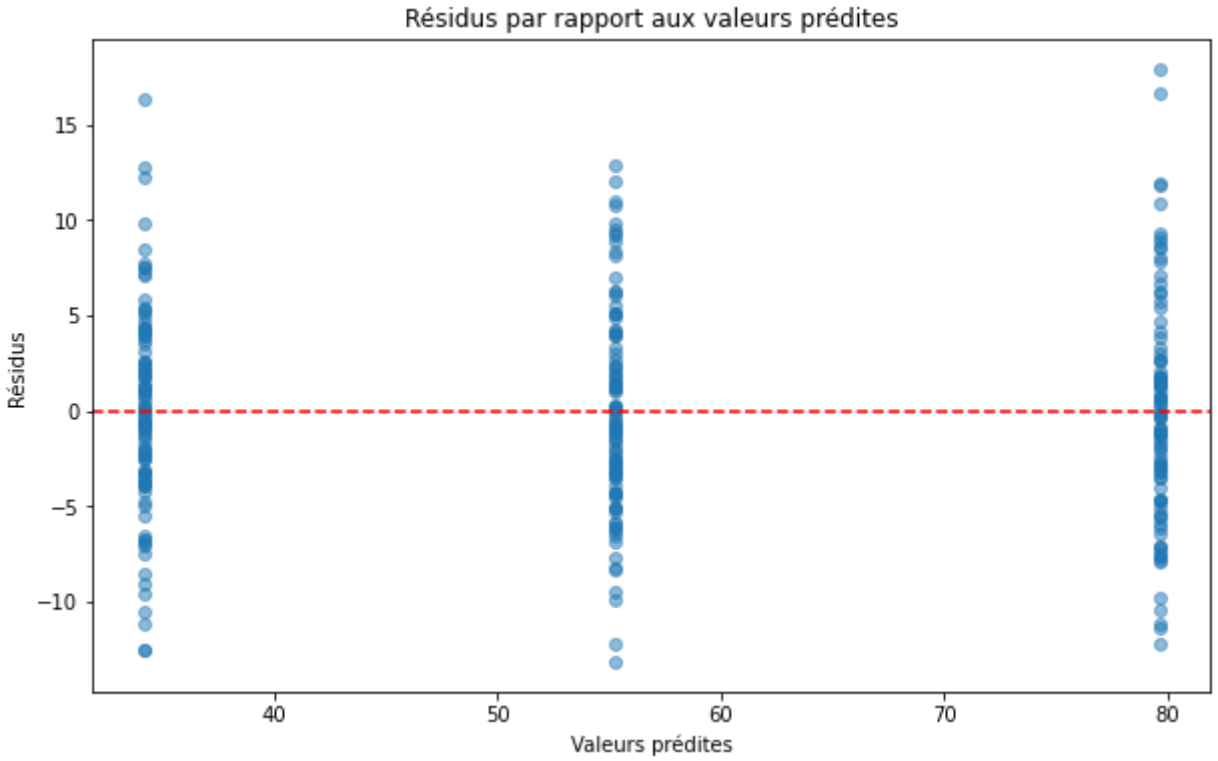
Pour vérifier l'homogénéité des variances on va utilsier le test de Breusch-Pagan ainsi qu'un graphique des résidus standardisés en fonction des valeurs prédites.

```python
# test de Breusch-Pagan pour l'homoscédasticité
bp_test = het_breuschpagan(residuals, model.model.exog)
labels = ['LM Stat', 'LM p-value', 'F Stat', 'F p-value']
bp_results = dict(zip(labels, bp_test))
print("Test de Breusch-Pagan :")
for key, value in bp_results.items():
    print(f"{key} : {value}")


# graphique des résidus standardisés vs valeurs ajustées
plt.figure(figsize=(10, 6))
plt.scatter(model.fittedvalues, residuals, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Résidus par rapport aux valeurs prédites')
plt.xlabel('Valeurs prédites')
plt.ylabel('Résidus')
plt.show()
```

Voici les résultats de l'étude de l'homogénéité des variances.

```
Test de Breusch-Pagan :
LM Stat : 0.9712142537018131
LM p-value : 0.6153235039150617
F Stat : 0.4823124847822556
F p-value : 0.6178368047840446
```



Le test de Breusch-Pagan confirme l'homogénéité des variances : on ne peut pas rejeter l'hypothèse nulle. Également, le graphique montre une distribution homogène des observations en chaque point du facteur explicatif.

Indépendance des données.

Il nous reste à vérifier l'indépendance des données, on peut utiliser un test de Durbin-Watson. Notez que l'indépendance des données est généralement garantie par la façon dont sont recueillies les données.

```
# test de Durbin-Watson
dw_statistic = durbin_watson(residuals)
print(f"Statistique de Durbin-Watson : {dw_statistic}")


Statistique de Durbin-Watson : 2.1853337338421053
```

La statistique de Durbin-Watson varie entre 0 et 4. Une valeur de 2 garatie l'indépendance des données.

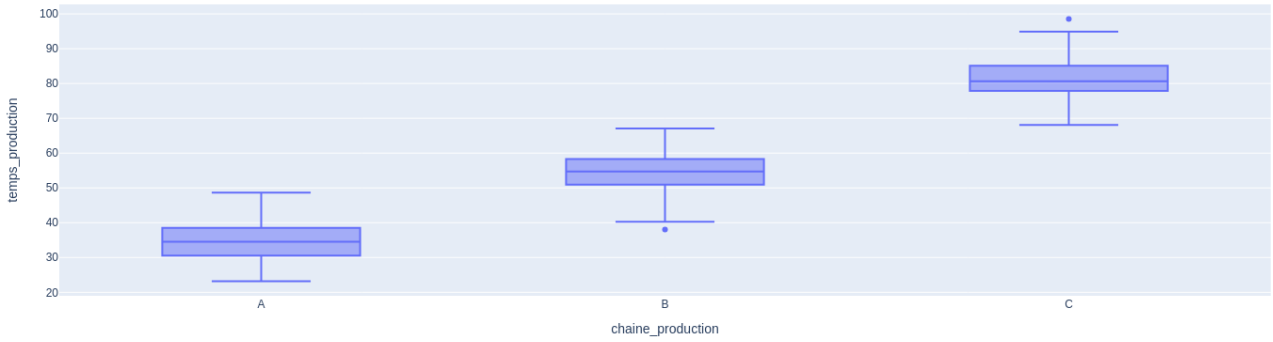
Nous avons donc vérifier l'ensemble des conditions d'application de l'anova. On peut maintenant passer à la lecture des résultats.

Inspection graphique de la validité des hypothèses.

Je disais précédemment que "normalité et homoscedasticité sont requises en chaque point du facteur explicatif." En effet, cela est équivalent à dire que les résidus sont normalement distribuée et à variance homogène. Pour vous en convaincre observer les diagrammes en boîte suivant : un diagramme pour chaque sous-échantillon.

```
import plotly.express as px

fig = px.box(data, x = "chaine_production", y="temps_production")
fig.show()
```



Sur ce graphique on peut observer la **distribution de la variable dépendante en chaque point du facteur explicatif**. Premièrement, on note que les distributions sont à variance homogène. Deuxièmement, les distributions sont normales malgré la présence d'un outlier parmis les chaînes de production B et C. Nous allons maintenant tester la significativité statistique de nos hypothèses.

```
# tests d'hypothèses en chaque modalité du facteur explicatif

# groupement des valeurs par chaîne de porduction
groups = df.groupby("chaine_production")["temps_production"].apply(list)

# normalité en chaque point du facteur explicatif
print("Normalité de la variable en chaque point du facteur explicatif")
print("---"*35)
for group in groups:
    statistic, pvalue = stats.shapiro(group)
    print(f"group has statistic {statistic} and pvalue {pvalue}")

print("---"*35)

# homoscédasticité en chaque point du facteur explicatif
from scipy.stats import bartlett
statistic, pvalue = bartlett(*groups, axis=0, nan_policy='propagate',
keepdims=False)
print("Homogénéité des variances avec le test de Bartlett : ")
print("---"*35)
print(f"The statistic is {statistic} and the pvalue {pvalue}")
```

Voici les résultats des tests statistiques.

```
Normalité de la variable en chaque point du facteur explicatif
---------------------------------------------------------------------------
```

```
------------------------------
group has statistic 0.9883927202932101 and pvalue 0.5375996303014767
group has statistic 0.9817484541115354 and pvalue 0.18149418636596398
group has statistic 0.9889764351567896 and pvalue 0.5827173779582634
------------------------------------------------------------------------------
------------------------------
Homogénéité des variances avec le test de Bartlett :
------------------------------------------------------------------------------
------------------------------
The statistic is 1.9816262409012764 and the pvalue 0.37127467721903507
```

Les données de la variable dépendante sont donc normalement distribuées en chaque point du facteur explicatif et à variance homogène. Les conclusions sont équivalente que celles des tests effectués sur les résidus du modèle linéaire qu'on a ajusté en début d'application. Pour cause, les deux assertions "les résidus sont normalement distribuées et à variance homogène" et "les données sont normalement distribuées et à variance homogène en chaque point du facteur explicatif" sont équivalentes.

> J'utilise toujours un boxplot conditionnel avant de réaliser les tests d'adéquation et d'homogénéité. Cela permet d'avoir un premeir visuel de la satisfaction des hypothèses d'application.

Anova - Calcul de la statistique de test F.

Voyons maintenant comment calculer notre statistique de test F. On sait qu'elle est le rapport de la variance factorielle et résiduelle. On doit donc calculer ces variances. Pour cela, on va utiliser les paramètres du modèle linéaire ajusté en début d'application.

```
# somme des carrés résiduels (SSE)
SSE = sum((model.resid) ** 2)

# somme des carrés totaux (SST) = Variabilité totale
SST = sum((df['temps_production'] - df['temps_production'].mean()) ** 2)

# somme des carrés factoriels (SSF)
SSF = SST - SSE

# degrés de liberté
dfn = model.df_model
dfd = model.df_resid

print(f"La somme des carrés factoriels {SSF} divisée par le degré de
liberté factoriels {dfn} donne la variance factorielle {SSF/dfn}.")
print(f"La somme des carrés résiduels {SSE} divisée par le degré de liberté
résiduel {dfd} donne la variance résiduelle {SSE/dfd}.")
```

```
    La somme des carrés factoriels 103876.91142829404 divisée par le degré
de liberté factoriels 2.0 donne la variance factorielle 51938.45571414702.
    La somme des carrés résiduels 9069.353797436608 divisée par le degré de
liberté résiduel 297.0 donne la variance résiduelle 30.53654477251383.
```

On peut maintenant calculer notre statistique de test F, c'est simplement le rapport de la variance factorielle et résiduelle.

```
# calcul de la statistique F
variance_factorielle = SSF/dfn
variance_residuelle = SSE/dfd
statistique_F = variance_factorielle/variance_residuelle

print(f"La statistique de test est : F={statistique_F}")
```

```
    La statistique de test est : F=1700.862232484705
```

On peut vérifier notre résultat grâce à la statistique F du modèle linéaire ajusté précédemment :

```
statistique_F = model.fvalue
print(f"La statistique F est : {statistique_F}")
```

```
    La statistique F est : 1700.8622324847042
```

Il ne nous reste plus qu'a comparer cette statistique au quotient de la table de Fisher pour 2 et 297 degrés de liberté. Avec python, on peut retrouver cette valeur critique grâce à la fonction *stats.f.ppf* du module *scipy*. Cette fonction prend en paramètre le **niveau de confiance** ($$1-\alpha$$), le nombre de **degrés de liberté** du nominateur (*dfn*) et du dénominateur (*dfd*).

```
# recherche quantile de la loi de Fisher
alpha = 0.05
quantile_critique = stats.f.ppf(1 - alpha, dfn, dfd)

print(f"Statistique F du modèle : {statistique_F}")
print(f"Quantile critique à {1-alpha} pour Fisher({int(dfn)}, {int(dfd)}) :
{quantile_critique}")

# règle de décision
if statistique_F > quantile_critique:
    print("La statistique F est significative, on rejette l'hypothèse
nulle.")
else:
    print("La statistique F n'est pas significative, on ne rejette pas
l'hypothèse nulle.")
```

```
    Statistique F du modèle : 1700.8622324847042
    Quantile critique à 0.95 pour Fisher(2, 297) : 3.0261533685653728
    La statistique F est significative, on rejette l'hypothèse nulle.
```

Si vous souhaitez trouver la probabilité d'observer la statistique de test - c'est-à-dire la **pvalue** - vous devez utiliser la **fonction cumulative** de la loi de Fisher (fonction de répartition cumulative). Précisément, on

cherche la probabilité d'observer une statistique de test au moins aussi grande que celle du test. Il faut donc trouver le complément de la probabilité retournée par la fonction de répartition cumulative, soit $$ \alpha = P(X \geq statistique_F) = 2*(1 - P(X \le statistique_F)) $$.

```
# calcul de la pvalue
pvalue = 2*(1 - stats.f.cdf(statistique_F, dfn, dfd))
print(f"La pvalue de la statistic de test F est : {pvalue}")
```

```
    La pvalue de la statistic de test F est : 2.220446049250313e-16
```

On soustrait la probabilité à 1 puisque le pvalue correspond à la probabilité que la statistique soit au moins aussi grande que la valeur observée. Puisque notre pvalue est inférieure à 5%, on conclut au rejet de l'hypothèse nulle : **nous avons suffisament d'évidence pour affirmer qu'au moins deux populations parentes ont des moyennes significativement différentes**.

Pour plus d'informations sur le test F de Fisher ou le calcul du quotient et de la pvalue, voici une référence.

Anova - table d'analyse de la variance.

Voici le table d'analyse de la variance. Elle présente d'une part la somme des carrés (**sum_sq**) factoriel et résiduels, d'autre part les nombres de degrés de liberté (**df**). Puis, on retrouve la statistique de test **F** et la pvalue associée à cette statistique (**PR(>F)**).

```
# application de l'ANOVA
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

```
                            sum_sq     df            F         PR(>F)
    C(chaine_production)  103876.911428    2.0  1700.862232  2.231586e-163
    Residual                9069.353797  297.0          NaN            NaN
```

Les résultats de la table d'analyse de la variance sont similaire à celles que nous avons calculés à la section précédente. Pour retrouver la statistique de test, nous l'avons vu, il vous suffit de diviser la somme des carrés avec leur degré de liberté respectif puis de faire le rapport de la variance factorielle et de la variance résiduelle.

$$Variance_factorielle = sum_sq(chaine_production) / df(chaine_production) = 103876.911428 / 2.00 = 51938.4557$$

$$Variance_résiduelle = sum_sq(Residual) / df(Residual) = 9069.353797436608 / 297.00 = 30.5365$$

$$statistique_F = \dfrac{variance_factorielle}{variance_residuelle} = 1700.862232$$

D'après la pvalue de la statistique de test, **il est peu probable que toutes les populations parentes aient une moyenne identique**. Autrement dit, il est probable qu'au moins une machine présente une défaillance compartivement à une autre. Pour savoir laquelle, il faut réaliser un test post-hoc afin de comparer tous les sous-échantillons deux-à-deux ou à un groupe de référence.

Voilà, nous sommes arrivés au bout de cette note sur l'analyse de la variance. Dans un prochain billet, nous verrons l'application des tests post-hoc tels que la procédure de Dunnett et de Tukey. À bientôt !

First tabs

{% tabs normalité %}

{% tab normalité python %}

```
from scipy.stats import shapiro
statistic, pvalue = shapiro(residuals)
print(f"Satistic {statistic}, pvalue : {pvalue}")
```

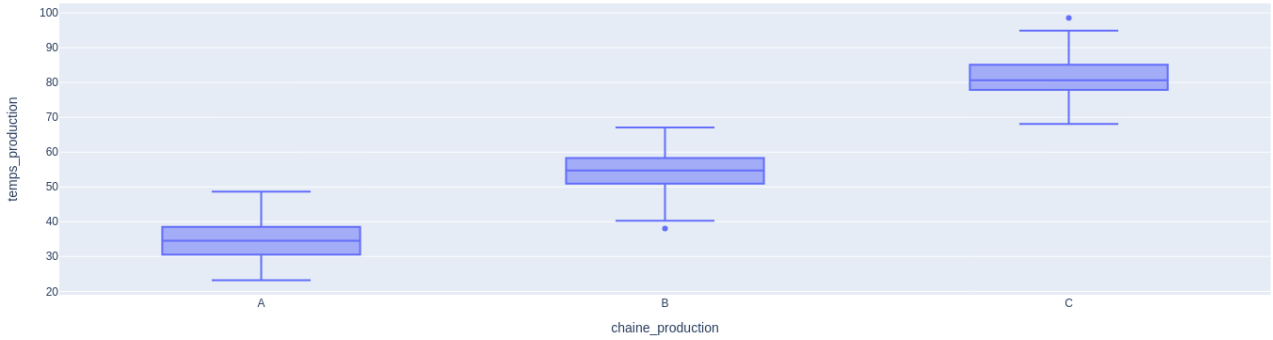
{% endtab %}

{% tab normalité r %}

```
df %>%
  group_by(chaine_production) %>%
  shapiro_test(temps_production)
```

Voici les résultats du test : La pvalue est supérieure au seuil de significativité : nous ne pouvons pas rejeter l'hypothèse nulle. {% endtab %}

{% tab normalité qqplot %}



{% endtab %}

{% endtabs %}