In [20]:
```
# Sujet 6 : Autour du Paradoxe de Simpson
# Parcours : Jupyter Notebook
# ------------------------------------------------
```

In [22]:
```
# Importations
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Chargement des données
file_path = 'Subject6_smoking.csv'
df = pd.read_csv(file_path)

# Aperçu des données
df.head()
```

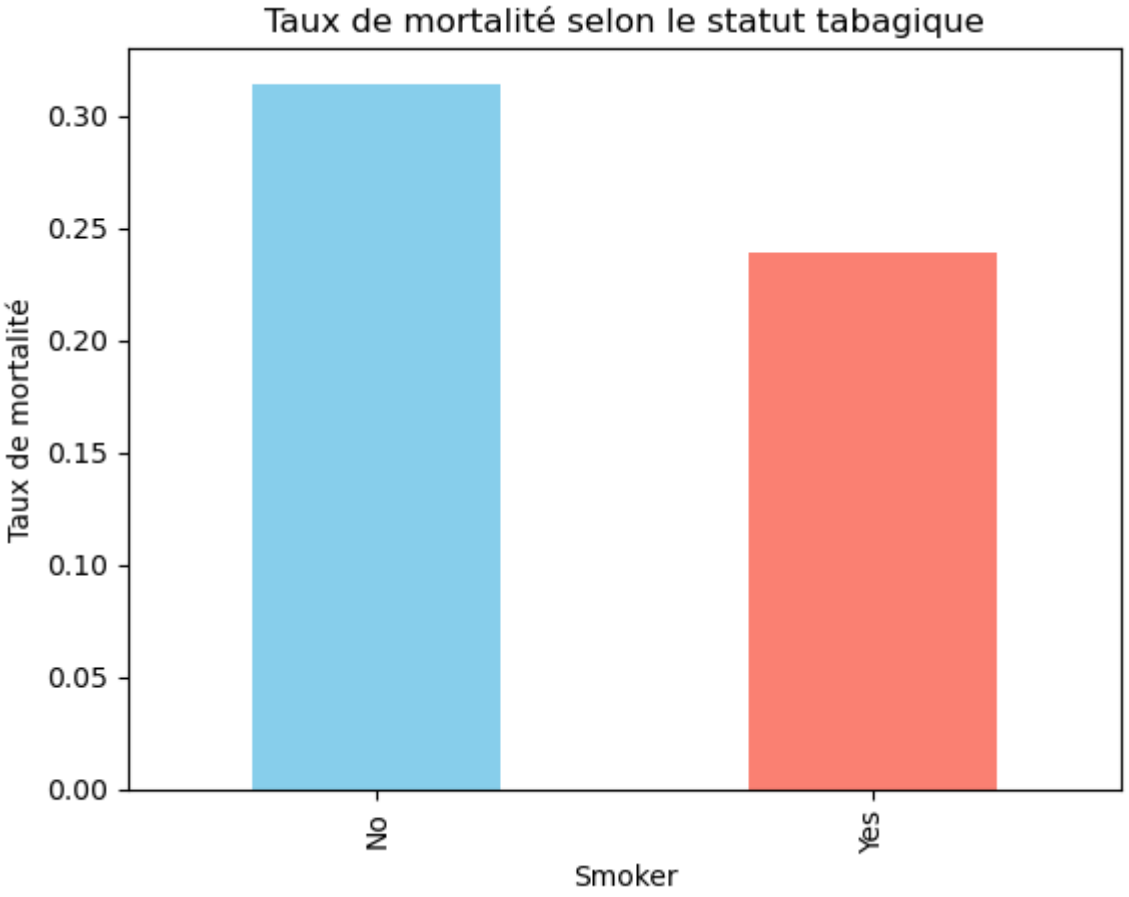
Out[22]:

	Smoker	Status	Age
0	Yes	Alive	21.0
1	Yes	Alive	19.3
2	No	Dead	57.5
3	No	Alive	47.1
4	Yes	Alive	81.4

In [24]:
```
# ------------------------------------------------
# Question 1 : Tableau vivantes/décédées selon le tabagisme
# ------------------------------------------------
```

In [26]:
```
# Tableau de contingence
contingency_table = pd.crosstab(df['Smoker'], df['Status'])

# Taux de mortalité
mortality_rates = contingency_table.apply(lambda row: row['Dead'] / (row['Alive'] + row['Dead'

# Affichage
print("Tableau de contingence:\n", contingency_table)
print("\nTaux de mortalité:\n", mortality_rates)

# Graphique
mortality_rates.plot(kind='bar', color=['skyblue', 'salmon'])
plt.ylabel('Taux de mortalité')
plt.title('Taux de mortalité selon le statut tabagique')
plt.show()
```

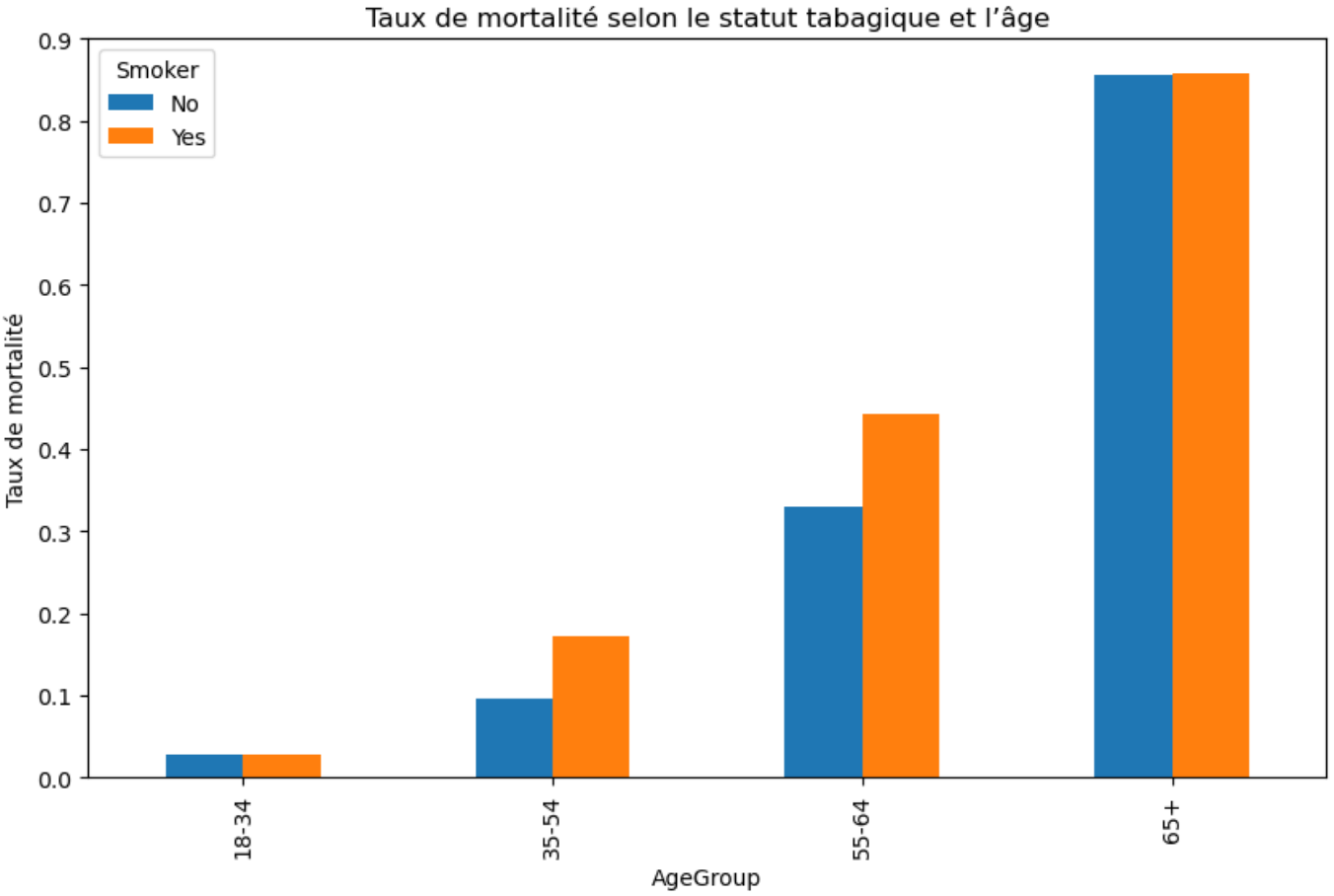
```
Tableau de contingence:
 Status  Alive  Dead
Smoker
No        502   230
Yes       443   139

Taux de mortalité:
 Smoker
No     0.314208
Yes    0.238832
dtype: float64
```



In [28]:
```
# ------------------------------------------------
# Question 2 : En introduisant les classes d'âge
# ------------------------------------------------
```

In [30]:
```
# Création de la colonne 'AgeGroup'
age_bins = [18, 34, 54, 64, 120]
age_labels = ['18-34', '35-54', '55-64', '65+']
df['AgeGroup'] = pd.cut(df['Age'], bins=age_bins, labels=age_labels)

# Nouveau tableau
contingency_table_age = pd.crosstab([df['Smoker'], df['AgeGroup']], df['Status'])

# Taux de mortalité par groupe
mortality_rates_age = contingency_table_age.apply(lambda row: row['Dead'] / (row['Alive'] + ro

# Affichage
print("Tableau de contingence par âge:\n", contingency_table_age)
print("\nTaux de mortalité par groupe:\n", mortality_rates_age)

# Graphique
mortality_rates_age.unstack(0).plot(kind='bar', figsize=(10,6))
plt.ylabel('Taux de mortalité')
plt.title('Taux de mortalité selon le statut tabagique et l'âge')
plt.legend(title='Smoker')
plt.show()
```

```
Tableau de contingence par âge:
 Status           Alive  Dead
Smoker AgeGroup
No     18-34       212     6
       35-54       180    19
       55-64        81    40
       65+          28   165
Yes    18-34       172     5
       35-54       196    41
       55-64        64    51
       65+           7    42

Taux de mortalité par groupe:
 Smoker  AgeGroup
No      18-34       0.027523
        35-54       0.095477
        55-64       0.330579
        65+         0.854922
Yes     18-34       0.028249
        35-54       0.172996
        55-64       0.443478
        65+         0.857143
dtype: float64
```

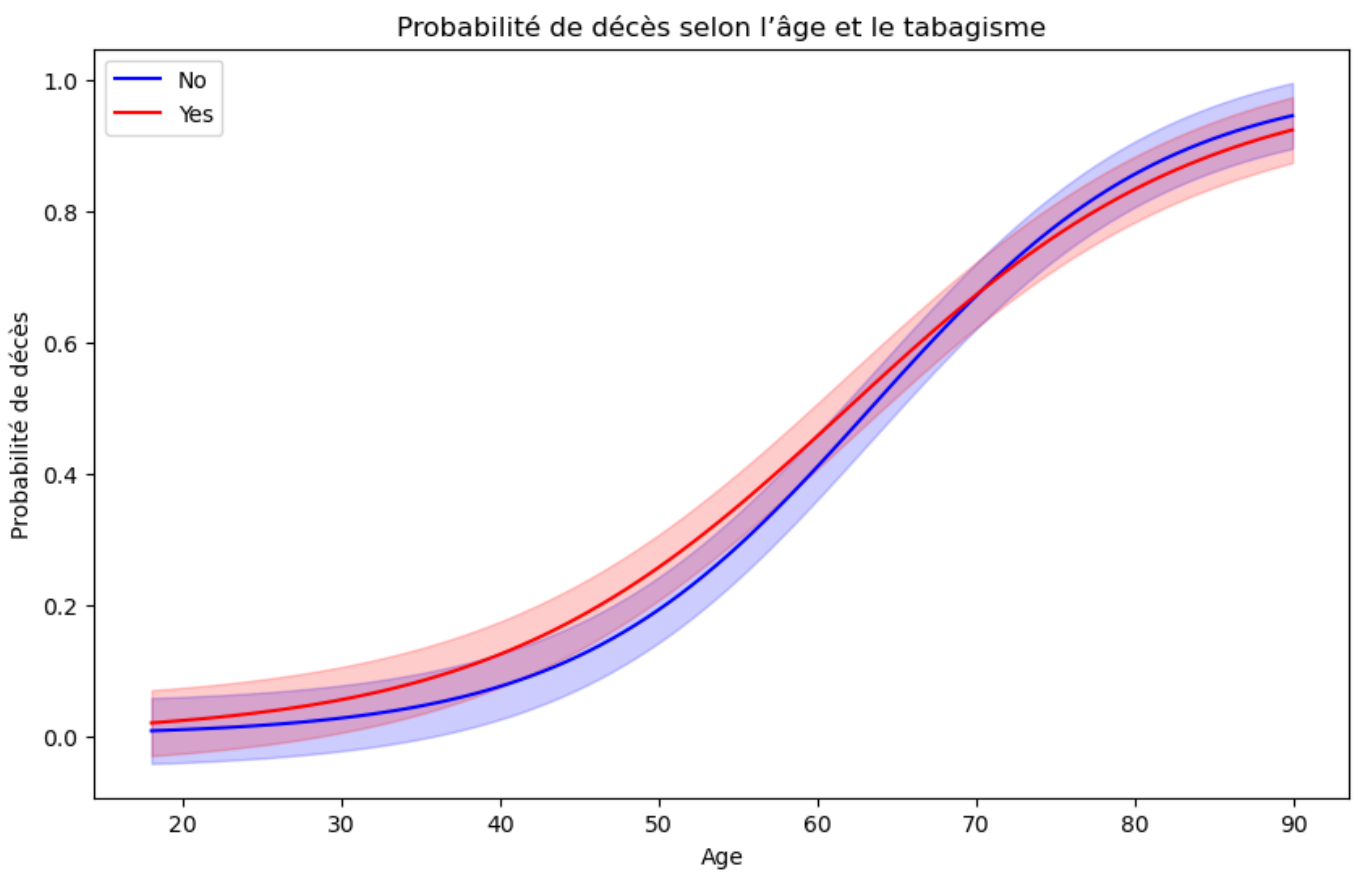


In [36]:
```
# ------------------------------------------------
# Question 3 : Régression logistique Status ~ Age
# ------------------------------------------------
```

In [38]:
```
# Création variable Death_bin (0=Alive, 1=Dead)
df['Death_bin'] = df['Status'].apply(lambda x: 1 if x == 'Dead' else 0)

# Définir les bons labels présents dans les données
smoker_values = df['Smoker'].unique()
smoker_label = smoker_values[0]
non_smoker_label = smoker_values[1]

# Régression pour les fumeuses
model_smoker = smf.logit('Death_bin ~ Age', data=df[df['Smoker'] == smoker_label]).fit()
```

```python
# Régression pour les non-fumeuses
model_non_smoker = smf.logit('Death_bin ~ Age', data=df[df['Smoker'] == non_smoker_label]).fit

# Affichage résumés
print(model_non_smoker.summary())
print(model_smoker.summary())

# Prédiction
ages = np.linspace(df['Age'].min(), df['Age'].max(), 100)

# Prédiction pour chaque groupe
pred_non_smoker = model_non_smoker.predict(pd.DataFrame({'Age': ages}))
pred_smoker = model_smoker.predict(pd.DataFrame({'Age': ages}))

# Graphique
plt.figure(figsize=(10,6))
plt.plot(ages, pred_non_smoker, label=f'{non_smoker_label}', color='blue')
plt.plot(ages, pred_smoker, label=f'{smoker_label}', color='red')
plt.fill_between(ages, pred_non_smoker - 0.05, pred_non_smoker + 0.05, color='blue', alpha=0.2
plt.fill_between(ages, pred_smoker - 0.05, pred_smoker + 0.05, color='red', alpha=0.2)
plt.xlabel('Age')
plt.ylabel('Probabilité de décès')
plt.title('Probabilité de décès selon l’âge et le tabagisme')
plt.legend()
plt.show()
```

```
Optimization terminated successfully.
         Current function value: 0.412727
         Iterations 7
Optimization terminated successfully.
         Current function value: 0.354560
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:              Death_bin   No. Observations:                  732
Model:                          Logit   Df Residuals:                      730
Method:                           MLE   Df Model:                            1
Date:                Sat, 26 Apr 2025   Pseudo R-squ.:                  0.4304
Time:                        18:54:29   Log-Likelihood:                -259.54
converged:                       True   LL-Null:                       -455.62
Covariance Type:            nonrobust   LLR p-value:                 2.808e-87
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -6.7955      0.479    -14.174      0.000      -7.735      -5.856
Age            0.1073      0.008     13.742      0.000       0.092       0.123
==============================================================================
                           Logit Regression Results
==============================================================================
Dep. Variable:              Death_bin   No. Observations:                  582
Model:                          Logit   Df Residuals:                      580
Method:                           MLE   Df Model:                            1
Date:                Sat, 26 Apr 2025   Pseudo R-squ.:                  0.2492
Time:                        18:54:29   Log-Likelihood:                -240.21
converged:                       True   LL-Null:                       -319.94
Covariance Type:            nonrobust   LLR p-value:                 1.477e-36
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -5.5081      0.466    -11.814      0.000      -6.422      -4.594
Age            0.0890      0.009     10.203      0.000       0.072       0.106
==============================================================================
```



In [ ]:
```
# ------------------------------------------------
# Remarques :
# Le paradoxe de Simpson apparaît ici car l'âge moyen diffère selon le groupe de fumeuses et r
# Les fumeuses étaient souvent plus jeunes que les non-fumeuses.
# Ainsi, à âge égal, les fumeuses peuvent avoir une mortalité plus forte,
# mais globalement (sans correction sur l’âge) elles semblent avoir une mortalité plus faible.
# ------------------------------------------------

# Fin du notebook
```

In [ ]:
```
```