In [7]:

```
#Pairplot
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
haberman=pd.read_csv('C:/Users/Dharanidhar/Desktop/MINIPJ/Personal/Habermann.csv')
sns.set_style('whitegrid')
#g=sns.pairplot(haberman,hue='Survival Status',size=3)
g.fig.suptitle("PAIRPLOT ON SURVIVAL STATUS")
sns.pairplot(haberman, vars=["Number of Axillary nodes", "AGE","Year of Operation"],hue='Survival Status',size=3)#Removed Survival Status as suggested
plt.show()
```

CONCLUSIONS: 1.AGE, Year of Operation and Number of Axillary nodes are not a good factors to depend on as, many of the graphs do not help us segregate the data and also that there way too many cases to plot.

2.Survival Status is the best factor to depend on to segregate data.

3.The people who are no more within 5 years of operation and those who are alive even after 5 years of operation is clearly identified in many of the graphs.

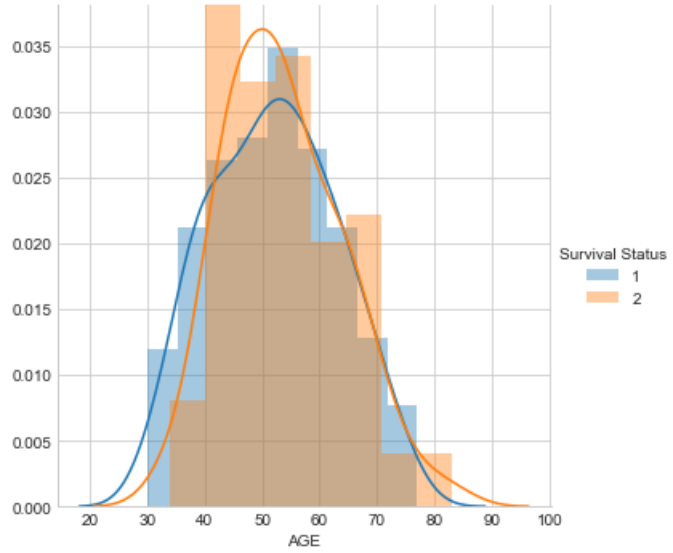
4.We can use if-else satements to easily seggregate the data.

5.(a)When we see the graph drawn between Number of auxillary nodes and Age, we can notice that the women who are aged between 40-60 have survived after 5 years of operation but the younger(below 40) and elder(above 60) did not have good chances to live. (b)Generally the number of nodes infected are between 0-10 but there are a considerable number of cases where more than 10 nodes have been infected(with highest number as 52 nodes). 6.(a)When we see the graph between Year of operation and Number of axillary nodes, we can see that in 1958 we had the case where one of the patients had 52 axillary nodes effected. 7.(a)When we see the graph between Year of Operation and Age, most number of patients are of the age 40-70. (b)The number of patients who lived 5 years after the operation are slightly less than the number of people who died before 5 years.

In [46]:

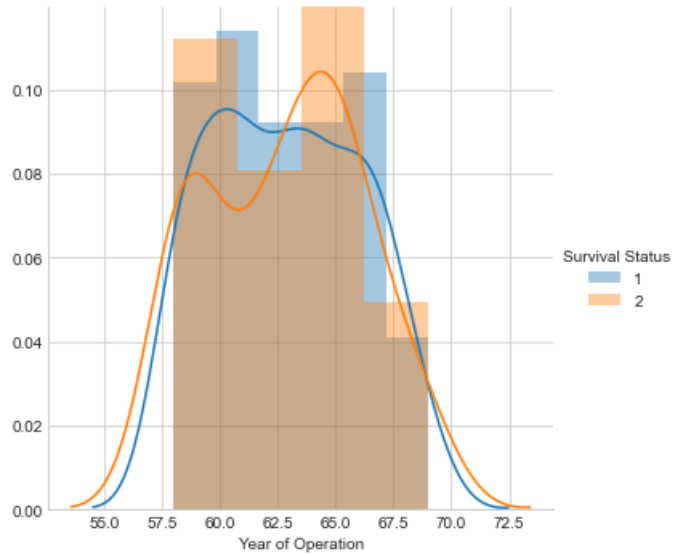
In [40]:

```
#HISTOGRAMS AND PDFs---Y axis is counts
import warnings
warnings.filterwarnings("ignore", category=UserWarning)
sns.FacetGrid(haberman, hue="Survival Status", size=5) \
   .map(sns.distplot, "AGE") \
   .add_legend();
plt.show();
```



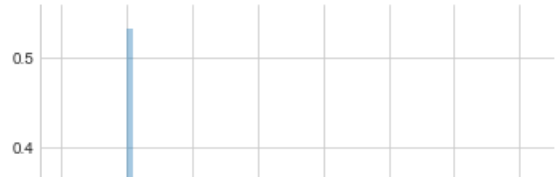
In [6]:

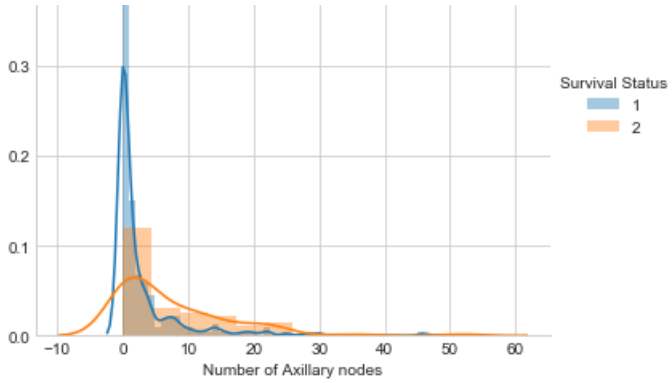
```
warnings.filterwarnings("ignore", category=UserWarning)
sns.FacetGrid(haberman, hue="Survival Status", size=5) \
   .map(sns.distplot, "Year of Operation") \
   .add_legend();
plt.show();
```



In [7]:

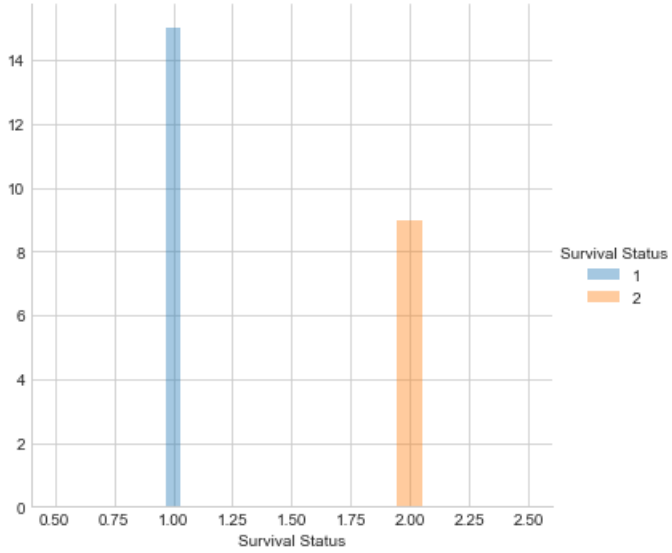
```
warnings.filterwarnings("ignore", category=UserWarning)
sns.FacetGrid(haberman, hue="Survival Status", size=5) \
   .map(sns.distplot, "Number of Axillary nodes") \
   .add_legend();
plt.show();
```





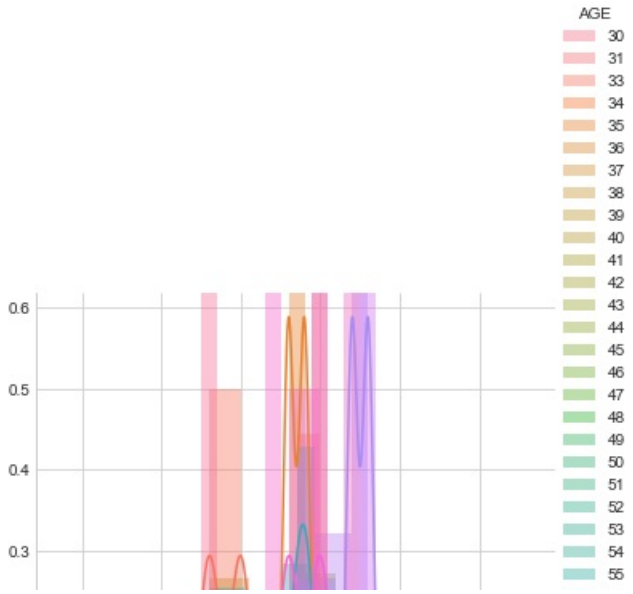
In [10]:

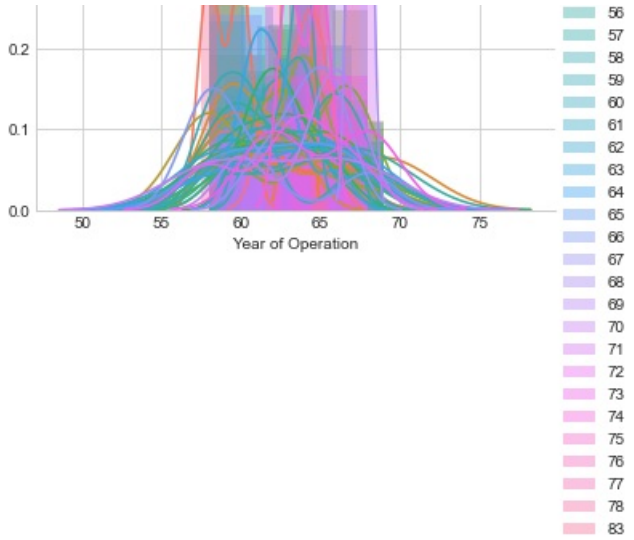
```
warnings.filterwarnings("ignore", category=UserWarning)
warnings.filterwarnings("ignore", category=RuntimeWarning)
sns.FacetGrid(haberman, hue="Survival Status", size=5) \
   .map(sns.distplot, "Survival Status") \
   .add_legend();
plt.show();
```



In [11]:

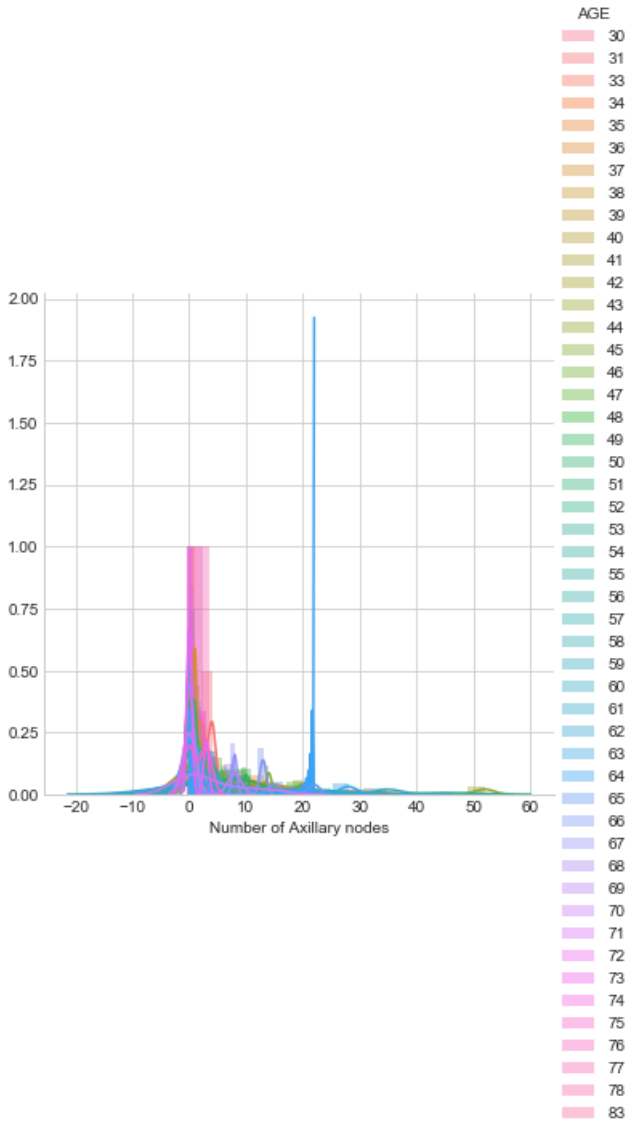
```
warnings.filterwarnings("ignore", category=UserWarning)
sns.FacetGrid(haberman, hue="AGE", size=5) \
   .map(sns.distplot, "Year of Operation") \
   .add_legend();
plt.show();
```





In [12]:

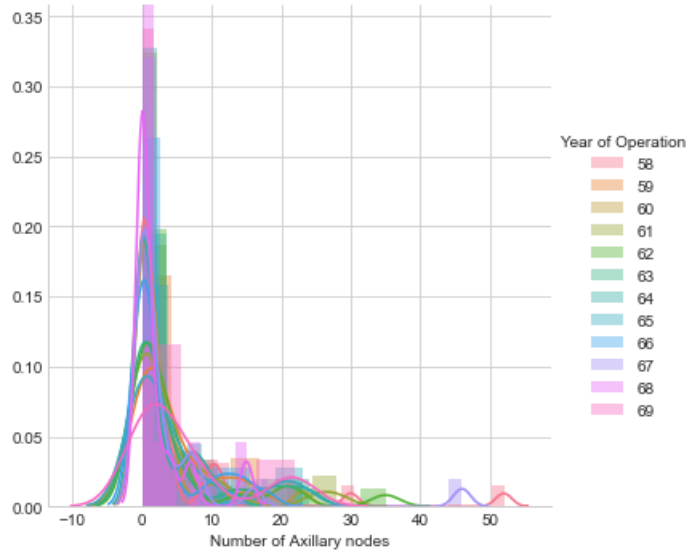
```
warnings.filterwarnings("ignore", category=UserWarning)
warnings.filterwarnings("ignore", category=RuntimeWarning)
sns.FacetGrid(haberman, hue="AGE", size=5) \
   .map(sns.distplot, "Number of Axillary nodes") \
   .add_legend();
plt.show();
```



In [58]:

```
import warnings
```

```
warnings.filterwarnings("ignore", category=UserWarning)
sns.FacetGrid(haberman, hue="Year of Operation", size=5) \
   .map(sns.distplot, "Number of Axillary nodes") \
   .add_legend();
plt.show();
```

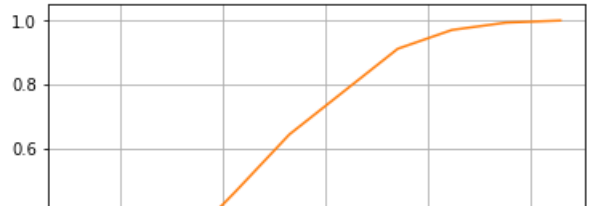


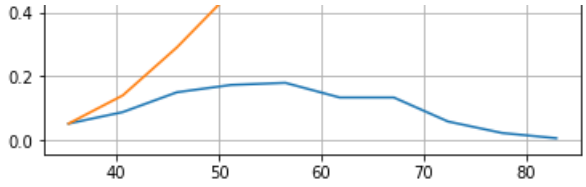
CONCLUSIONS: 1.Except for the "SURVIVAL STATUS" factor, none of the factors can be depended on for proper separation of data as all the graphs except for it are very much intersecting with each other. 2.Normal if-else statement can be used to properly segregate the data. 3.(a)With respect to AGE the pdf or histogram for the women who survived after 5 years of operation is higher. (b)It tells us that the number of people who are alive after 5 years of operation are mainly of the age 50 and those who are not alive are the people of age 53(most no of cases from the pdf ....as more no of cases the higher is the pdf) 4.From the next graph between Year of Operation and Survival Status we can see that most of the successfull operations happened between the years 1963-1964 and thoe which took place between 1959-1962 have failed the most. 5.In the next graph between number of axillary nodes and survival status we can see that,most of the women whose number of axillary nodes which did not get effected at all(number of axillary nodes =0) could not survive after 5 years of operation and almost 80% of the women did not have any auxillary nodes effected, but th rest 20% of the women survived. 6.Most of the women who came for treatment have the number of axillary nodes between 0-26. 7.All the other PDFs and Histograms are not so useful for drawing conclusions.

In [52]:

```
#CDF y axis------>percentage/percentiles blue pdf
plt.grid()
counts, bin_edges = np.histogram(haberman['AGE'], bins=10, density = True)
pdf = counts/(sum(counts))
print("PDFs",pdf)
print()
print("BIN EDGES",bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:],cdf)

counts, bin_edges = np.histogram(haberman['AGE'], bins=20, density = True)
pdf = counts/(sum(counts))
#plt.plot(bin_edges[1:],pdf)
#sns.add_legend();

plt.show();
```

```
PDFs [0.05228758 0.08823529 0.1503268  0.17320261 0.17973856 0.13398693
 0.13398693 0.05882353 0.02287582 0.00653595]

BIN EDGES [30.  35.3 40.6 45.9 51.2 56.5 61.8 67.1 72.4 77.7 83. ]
```

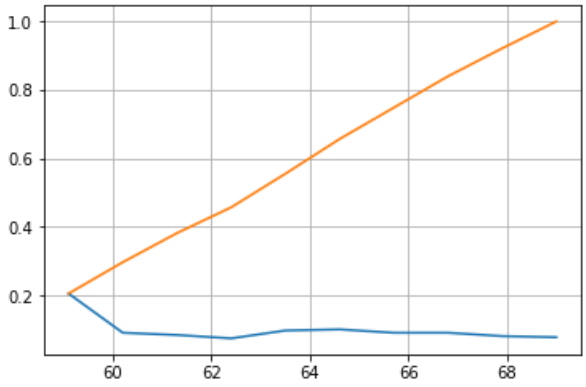




In [45]:

```
plt.grid()
counts, bin_edges = np.histogram(haberman['Year of Operation'], bins=10, density = True)
pdf = counts/(sum(counts))
print("PDFs",pdf)
print()
print("BIN EDGES",bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)


counts, bin_edges = np.histogram(haberman['Year of Operation'], bins=20, density = True)
pdf = counts/(sum(counts))
#plt.plot(bin_edges[1:],pdf)

plt.show();
```

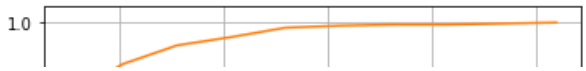
```
PDFs [0.20588235 0.09150327 0.08496732 0.0751634  0.09803922 0.10130719
 0.09150327 0.09150327 0.08169935 0.07843137]

BIN EDGES [58.  59.1 60.2 61.3 62.4 63.5 64.6 65.7 66.8 67.9 69. ]
```

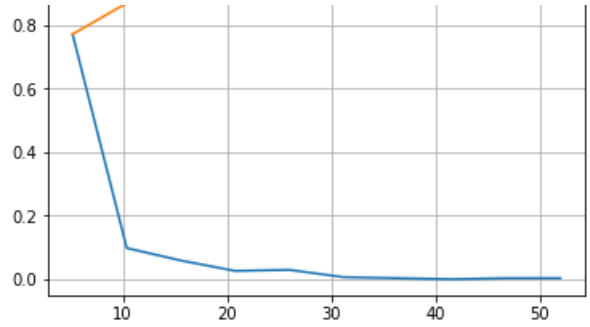


In [46]:

```
plt.grid()
counts, bin_edges = np.histogram(haberman['Number of Axillary nodes'], bins=10, density = True)
pdf = counts/(sum(counts))
print("PDFs",pdf)
print()
print("BIN EDGES",bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)


counts, bin_edges = np.histogram(haberman['Number of Axillary nodes'], bins=20, density = True)
pdf = counts/(sum(counts))
#plt.plot(bin_edges[1:],pdf);

plt.show();
```

```
PDFs [0.77124183 0.09803922 0.05882353 0.02614379 0.02941176 0.00653595
 0.00326797 0.         0.00326797 0.00326797]

BIN EDGES [ 0.   5.2 10.4 15.6 20.8 26.  31.2 36.4 41.6 46.8 52. ]
```

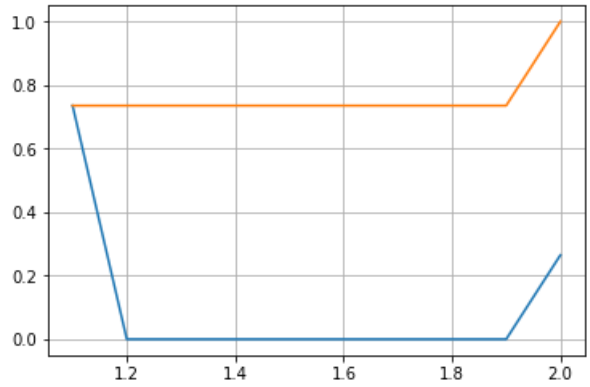




In [48]:

```
#CDF
#one=haberman[haberman["Survival Status"]=="One"]
#two=haberman[haberman["Survival Status"]=="Two"]
plt.grid()
counts, bin_edges = np.histogram(haberman['Survival Status'], bins=10,density = True)
pdf = counts/(sum(counts))
print("PDFs",pdf)
print()
print("BIN EDGES",bin_edges)
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)


counts, bin_edges = np.histogram(haberman['Survival Status'], bins=20,density = True)
pdf = counts/(sum(counts))
#plt.plot(bin_edges[1:],pdf);

plt.show();
```

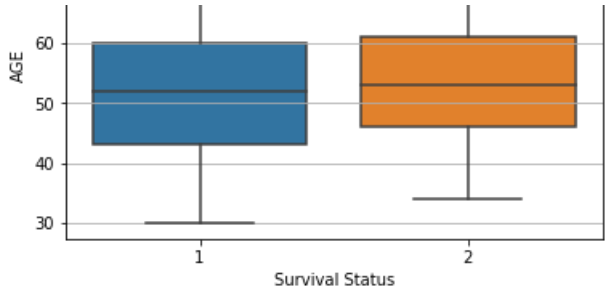
```
PDFs [0.73529412 0.         0.         0.         0.         0.
 0.         0.         0.         0.26470588]

BIN EDGES [1.  1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2. ]
```



CONCLUSIONS:1) Every graph consists of 2 lines where the blue represents PDF of the patients who have died before 5 years after the operation and green is the PDF for the patients ho are alive after 5 years of operation and the orange line represents cdf of the hue taken in each case. 2)They can be written in normal if-else statements also. 3)in the first CDF we can see that, approximately 75% of the women who got operated are below age of 60 years. 4)From the second CDF we can see that 60% of the operations were performed before 1964. 5)78% of the women have axillary nodes below 10 6)Rest of the graphs are not much of a use.

In [57]:

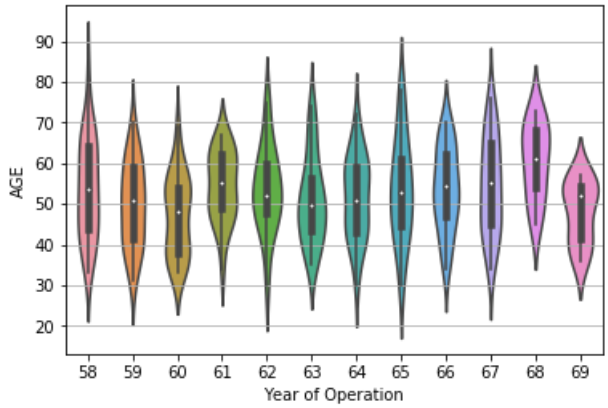
```
#BOXPLOT
plt.grid()
sns.boxplot(x='Survival Status',y='AGE', data=haberman)
plt.show()
```



CONCLUSIONS: 1.The box plot with X axis: AGE and Y axis: Survival Status is the only box plot which makes sense. 2.Percentiles 25th 50th 75th percentile value 3.For those who could not live after 5 years after operation: 25% of the women were below 43 years old 50% of the women were below 52 years old 75% of the women were below 60 years old 4.For those who could live after 5 years after operation: 25% of the women were below 45 years old 50% of the women were below 53 years old 75% of the women were below 61 years old

In [58]:

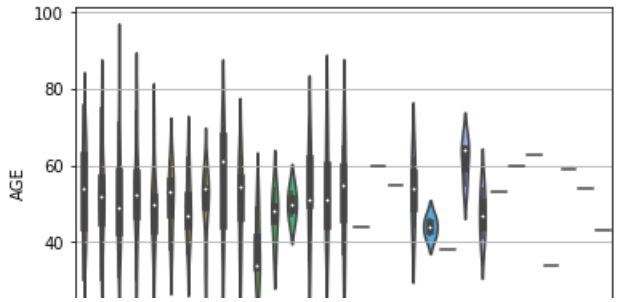
```
#VIOLIN PLOT
plt.grid()
sns.violinplot(x='Year of Operation',y='AGE', data=haberman)
plt.show()
```

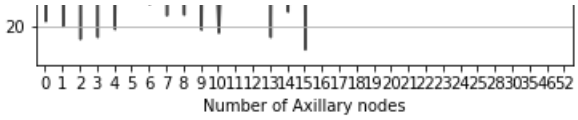


From the above graph we can say that: 1)In the year 1958, the 25th percentile is 43 years, 50th percentile is 53 years and 75th percentile is 64 years 2)In the year 1959, the 25th percentile is 41 years, 50th percentile is 50 years and 75th percentile is 59 years 3)In the year 1960, the 25th percentile is 39 years, 50th percentile is 48 years and 75th percentile is 52 years 4)In the year 1961, the 25th percentile is 49 years, 50th percentile is 55 years and 75th percentile is 62 years 5)In the year 1962, the 25th percentile is 48 years, 50th percentile is 51 years and 75th percentile is 59 years 6)In the year 1963, the 25th percentile is 42 years, 50th percentile is 49 years and 75th percentile is 58 years 7)In the year 1964, the 25th percentile is 42 years, 50th percentile is 50 years and 75th percentile is 59 years 8)In the year 1965, the 25th percentile is 44 years, 50th percentile is 51 years and 75th percentile is 61 years 9)In the year 1966, the 25th percentile is 48 years, 50th percentile is 52 years and 75th percentile is 62 years 10)In the year 1967, the 25th percentile is 44 years, 50th percentile is 52 years and 75th percentile is 63 years 11)In the year 1968, the 25th percentile is 52 years, 50th percentile is 60 years and 75th percentile is 65 years 12)In the year 1969, the 25th percentile is 38 years, 50th percentile is 50 years and 75th percentile is 52 years So in all these years, 25% of the women who got opearated are 44 years old 50% of the women who got opearated are 52 years old 75% of the women who got opearated are 60 years old

In [61]:

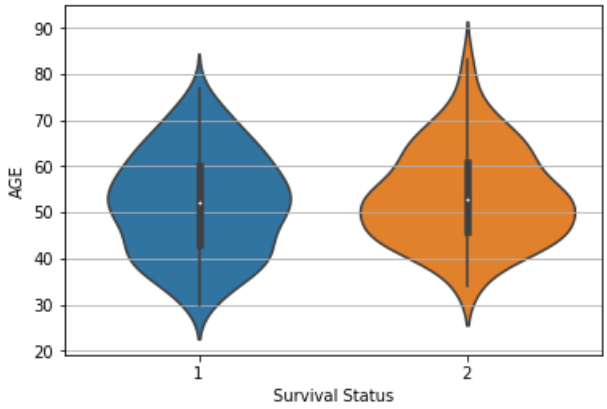
```
plt.grid()
sns.violinplot(x='Number of Axillary nodes',y='AGE', data=haberman)
plt.show()
```





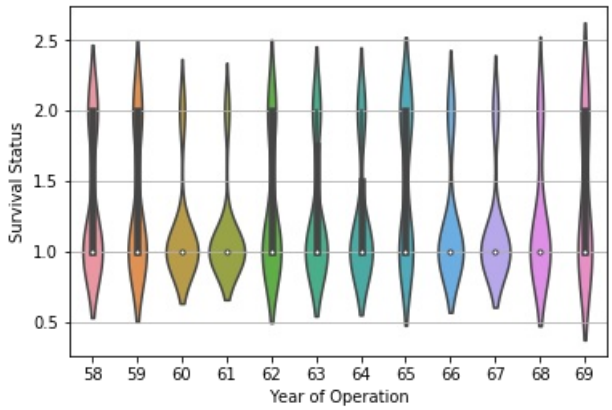
In [62]:

```
plt.grid()
sns.violinplot(x='Survival Status',y='AGE', data=haberman)
plt.show()
```



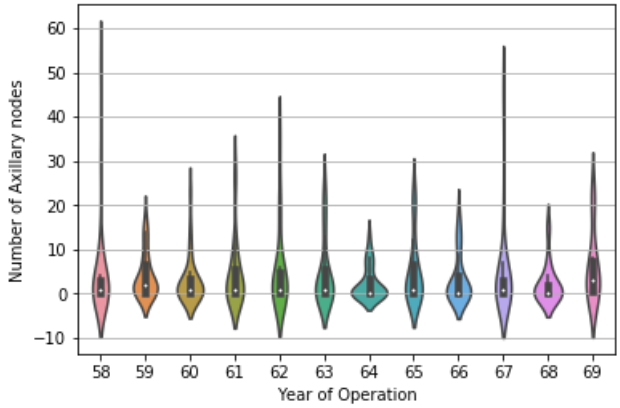
In [63]:

```
plt.grid()
sns.violinplot(x='Year of Operation',y='Survival Status', data=haberman)
plt.show()
```



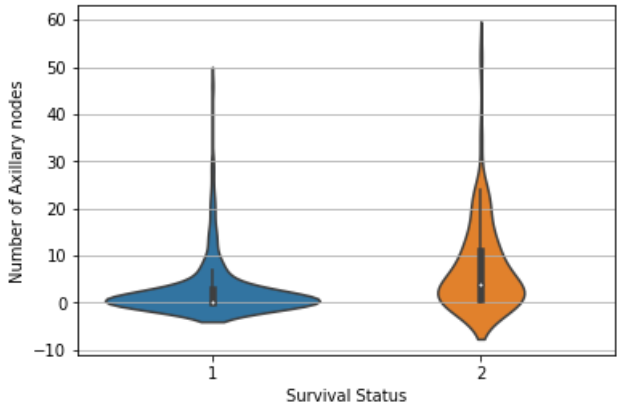
In [64]:

```
plt.grid()
sns.violinplot(x='Year of Operation',y='Number of Axillary nodes', data=haberman)
plt.show()
```



In [65]:

```
plt.grid()
sns.violinplot(x='Survival Status',y='Number of Axillary nodes', data=haberman)
plt.show()
```



CONCLUSIONS: 1)The Violin plots with A)X Axis:Survival Status and Y Axis:Number of Axillary nodes B)X Axis:Year of operation and Y Axis:Number of Axillary Nodes c)X Axis:Year of Operation and Y Axis:Survival Status are the best graphs to analyze the data as it clearly diffrentiates the graphs based on the data. 2)histogram pdf + box plot , to find the shape of the curve which helps us to identify whether the curve is gaussian or not. In the Black box(box plot)white dot-->50th percentile, bottom 25 top 75, pointed ends are the whiskers and pdf is the side curve.