In [13]:
```
import warnings
warnings.filterwarnings("ignore")
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
```

In [214]:
```
data_df = pd.read_csv('haberman.csv')
```

1. Age of patient at time of operation (numerical & continous)
2. Patient's year of operation (year - 1900, numerical & continous)
3. Number of positive axillary nodes detected (numerical & ordinal)
4. Survival status (dependent, nominal, categorical and target variable)

```
#1 = the patient survived > 5 years after operation
#2 = the patient died within 5 year
```

Objective: Given the Age, year of operation and no.of positive axillary nodes detected, We need to predict the survival status.

In [8]:
```
print(data_df.shape)
data_df.head()
```

```
(306, 4)
```

Out[8]:

	age	year	nodes	status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1

Observation: The data set has 306 data points and 4 variables in total, 3 are independent and 1 is depenedent

In [4]:
```
data_df['status'].value_counts()   # target variable
```

Out[4]:
```
1    225
2     81
Name: status, dtype: int64
```

Observation: The data is imbalanced with ratio of 2.7777. The people survived are 2.777 times more than people died. The chance of surivial = 73.5%

In [40]:
```
data_df.isnull().sum()   # no null values in any of the columns
```

Out[40]:
```
age       0
year      0
nodes     0
status    0
dtype: int64
```

In [36]:
```
data_df.describe()   # get stats of data
```

Out[36]:

	age	year	nodes	status
count	306.000000	306.000000	306.000000	306.000000
mean	52.457516	62.852941	4.026144	1.264706
std	10.803452	3.249405	7.189654	0.441899
min	30.000000	58.000000	0.000000	1.000000
25%	44.000000	60.000000	0.000000	1.000000
50%	52.000000	63.000000	1.000000	1.000000
75%	60.750000	65.750000	4.000000	2.000000
max	83.000000	69.000000	52.000000	2.000000

Observations:

1. While highest no.of nodes is 52, 75% has nodes less than or equal to 4 and 25% has 0 nodes.
2. mean for nodes is 4 and median for the same is 1, this represents the data of nodes is skewed.
3. Other columns, age and year have almost equal mean and median represents that both columns are in normal distrubution.

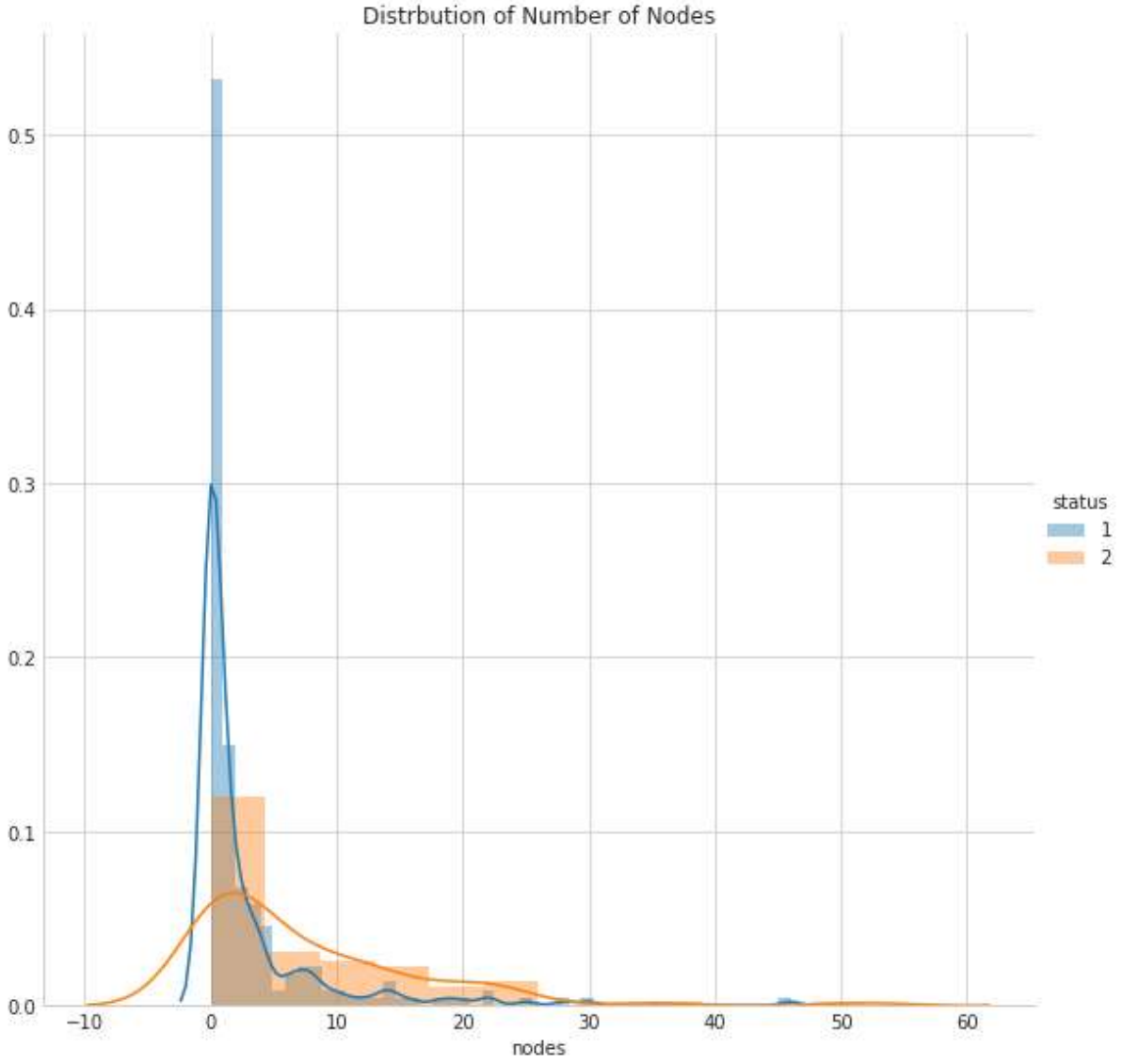
In [46]:
```
print(data_df.dtypes)
data_df = data_df.astype('int8')   # for memory optimization as maxvalue in dataframe is only 83.
data_df.dtypes
```

```
age       int64
year      int64
nodes     int64
status    int64
dtype: object
```

Out[46]:
```
age       int8
year      int8
nodes     int8
status    int8
dtype: object
```

Univariate Analysis

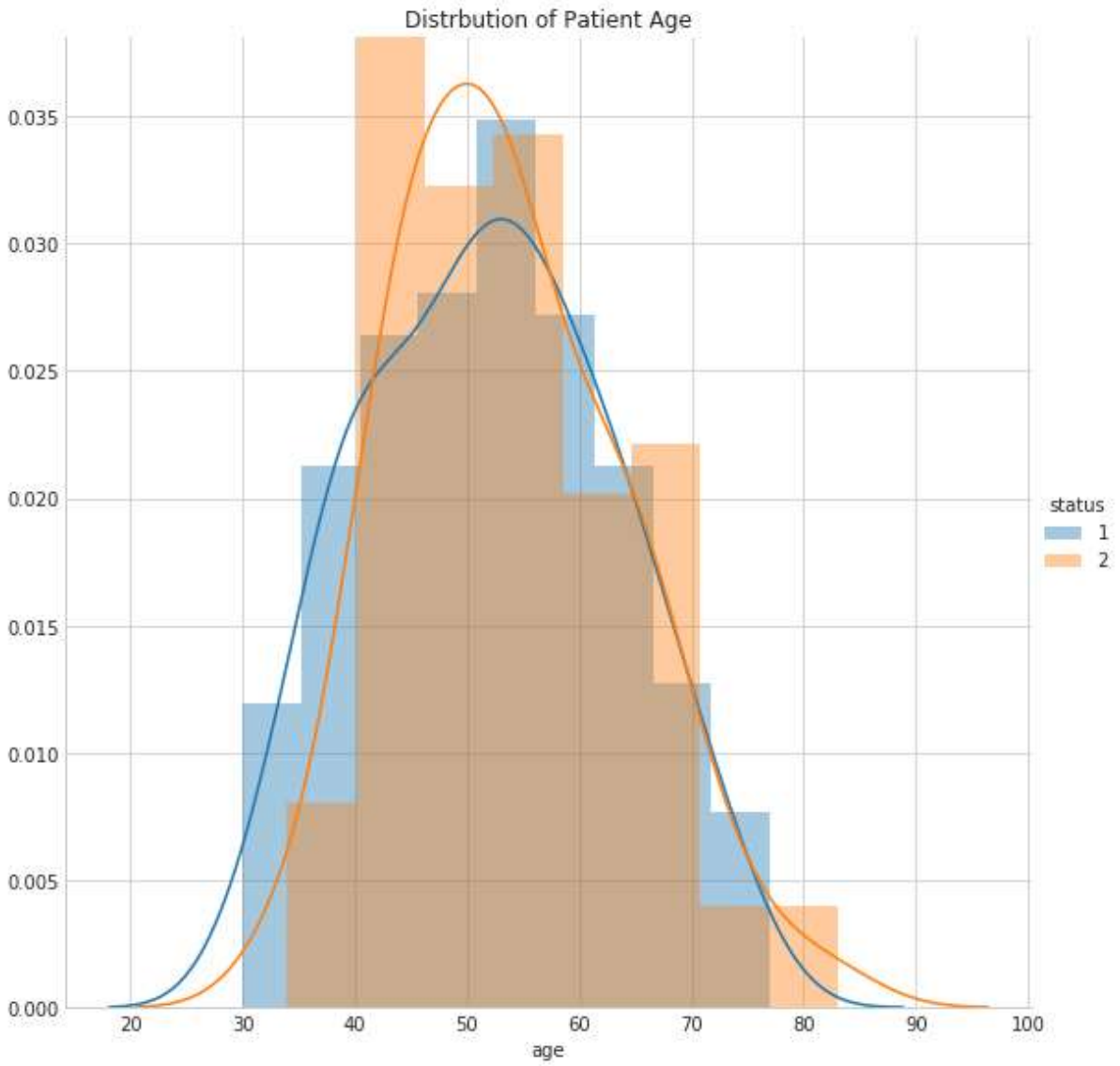
In [133]:
```
sns.FacetGrid(data_df, hue="status", size=8) \
   .map(sns.distplot, "nodes") \
   .add_legend();
plt.title('Distrbution of Number of Nodes')
plt.show();
```



1.In case of 0 nodes, the probability to survive more than 5 years in more.

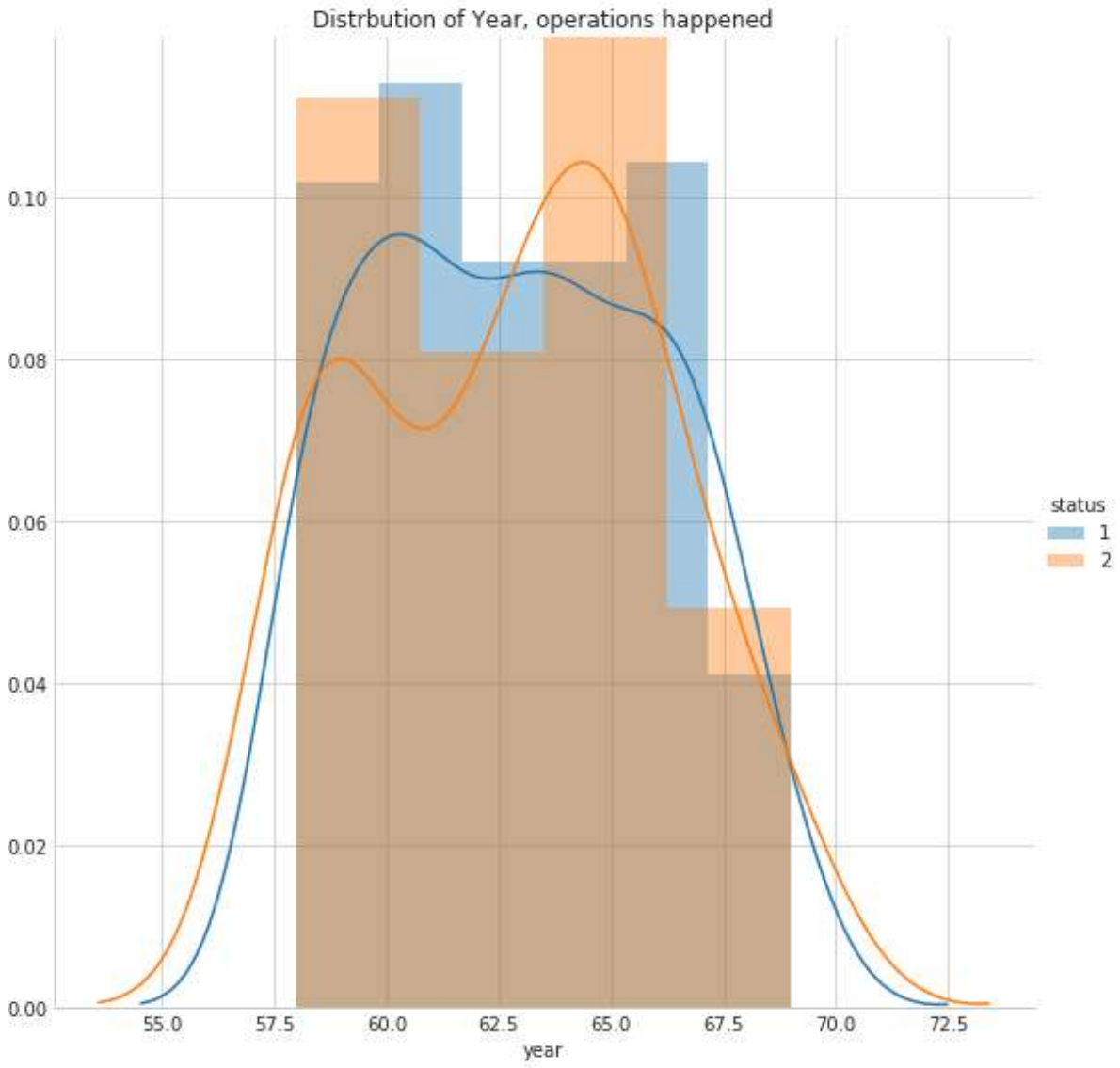
2.For nodes > 1, the probability of dying with in 5 years is more.

In [134]:
```
sns.FacetGrid(data_df, hue="status", size=8) \
   .map(sns.distplot, "age") \
   .add_legend();
plt.title('Distrbution of Patient Age')
plt.show();
```



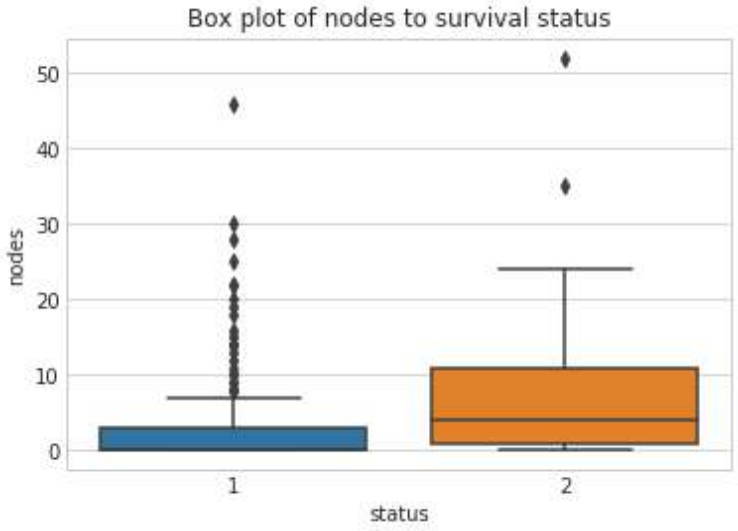
1. Deaths are highest in the age range of 40 to 50 and 60 to 70.
2. Most people less than 40 year of age survived.(Probability is more)

In [135]:
```
sns.FacetGrid(data_df, hue="status", size=8) \
   .map(sns.distplot, "year") \
   .add_legend();
plt.title('Distrbution of Year, operations happened')
plt.show();
```



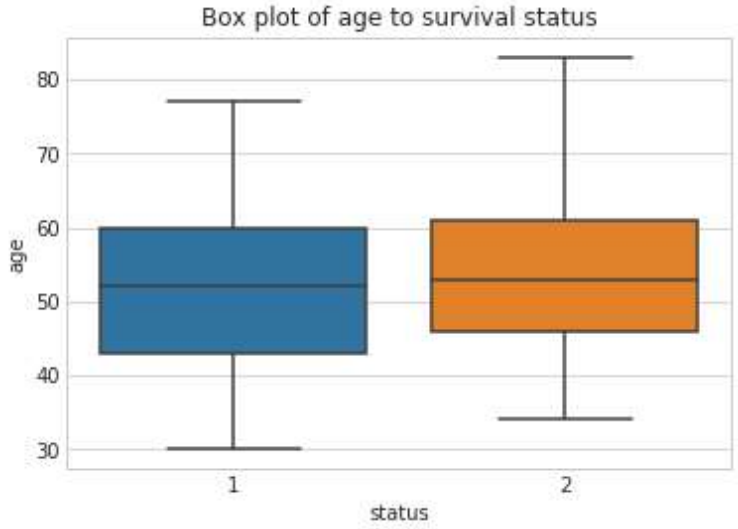
1. With almost similar distributions, year seams to be less important variable.
2. Deaths for year 63 to 66 and 58 to 59 are more than the survivals

In [179]:
```
sns.boxplot(x='status', y='nodes', data=data_df)
plt.title('Box plot of nodes to survival status')
plt.show()
```

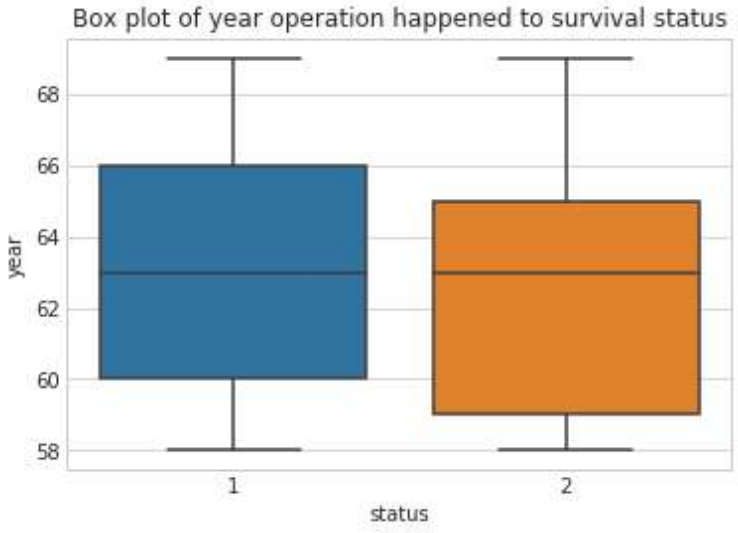


Observations: There seem to be more outliers where no.of nodes being high and survived. Nodes > 30 are very rare.

In [180]:
```
sns.boxplot(x='status', y='age', data=data_df)
plt.title('Box plot of age to survival status')
plt.show()
```



In [181]:
```
sns.boxplot(x='status', y='year', data=data_df)
plt.title('Box plot of year operation happened to survival status')
plt.show()
```



Observations: There is lot of overlapping b/w both survived and not survived cases when compared using age and year.With only these features, status cannot be predicted accurately, making the no.of nodes most important feature.

PDFs and CFDs

In [33]:
```
status1 = data_df[ data_df['status'] == 1]
status2 = data_df[ data_df['status'] == 2]
```

In [128]:
```
status1.describe()
```

Out[128]:

	age	year	nodes	status
count	225.000000	225.000000	225.000000	225.0
mean	52.017778	62.862222	2.791111	1.0
std	11.012154	3.222915	5.870318	0.0
min	30.000000	58.000000	0.000000	1.0
25%	43.000000	60.000000	0.000000	1.0
50%	52.000000	63.000000	0.000000	1.0
75%	60.000000	66.000000	3.000000	1.0
max	77.000000	69.000000	46.000000	1.0

In [129]:
```
status2.describe()
```

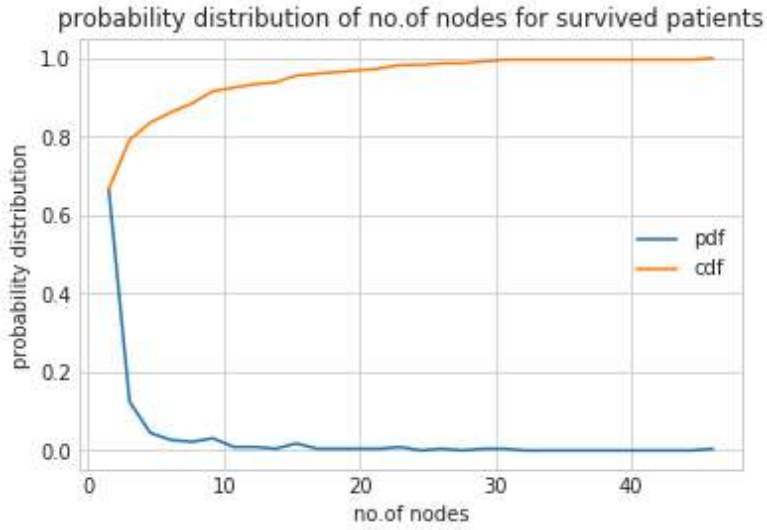
Out[129]:

	age	year	nodes	status
count	81.000000	81.000000	81.000000	81.0
mean	53.679012	62.827160	7.456790	2.0
std	10.167137	3.342118	9.185654	0.0
min	34.000000	58.000000	0.000000	2.0
25%	46.000000	59.000000	1.000000	2.0
50%	53.000000	63.000000	4.000000	2.0
75%	61.000000	65.000000	11.000000	2.0
max	83.000000	69.000000	52.000000	2.0

Observations:

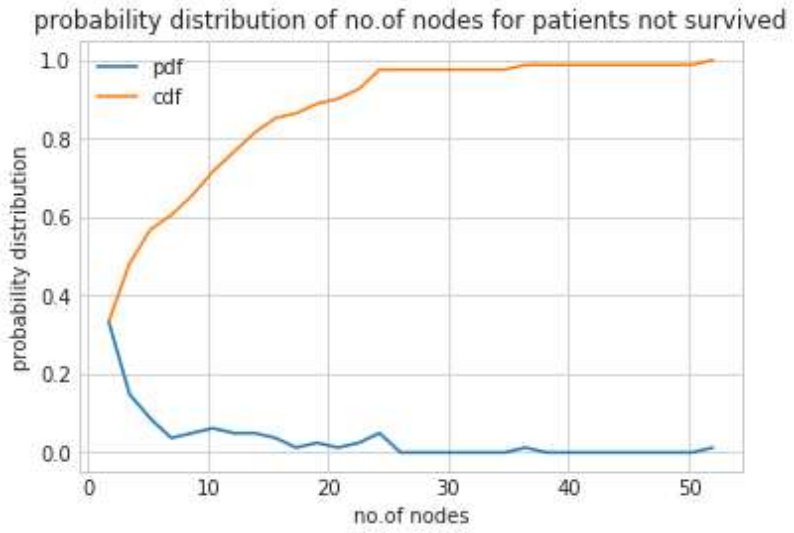
1. The measure of central tendancies didnt vary for the "year" variable for both cases.
2. There is slight change in median "age" of who survived and who is not.
3. The median of "no.of nodes" has seen significant change in case of who didnt survived when compared to who survived

In [178]:
```
# nodes are most important of all three

counts, bin_edges = np.histogram(status1['nodes'], bins=30,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)

plt.legend(['pdf', 'cdf'])
plt.xlabel('no.of nodes')
plt.ylabel('probability distribution')
plt.title( 'probability distribution of no.of nodes for survived patients')
plt.show();
```



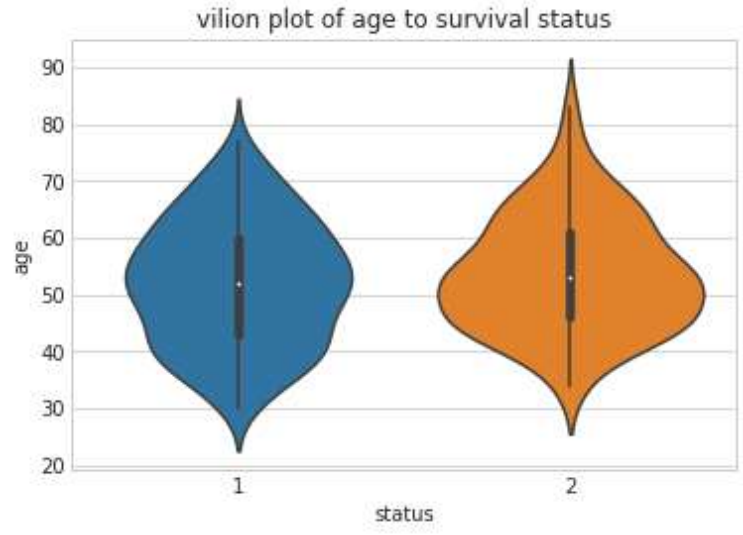
In [177]:
```
#versicolor
counts, bin_edges = np.histogram(status2['nodes'], bins=30,
                                 density = True)
pdf = counts/(sum(counts))
cdf = np.cumsum(pdf)
plt.plot(bin_edges[1:],pdf)
plt.plot(bin_edges[1:], cdf)
plt.xlabel('no.of nodes')
plt.ylabel('probability distribution')
plt.title( 'probability distribution of no.of nodes for patients not survived')
plt.legend(['pdf', 'cdf'])
```

Out[177]: <matplotlib.legend.Legend at 0x7f83380ad8d0>

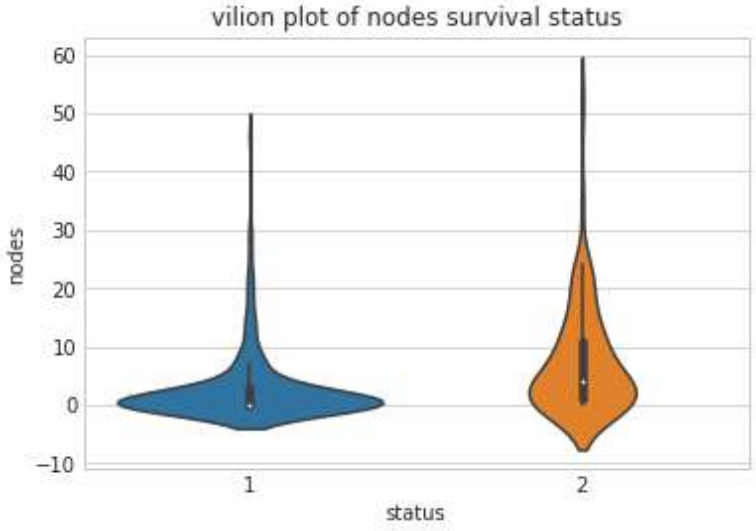


Observations: From the above two pdfs, The probability for surivival in more in case of nodes = 0 (1st bin) and in all remaining cases (nodes > 1) the probability of survival is less than or equal to non-survival.

In [183]:
```
sns.violinplot(x="status", y="age", data=data_df, size=6)
plt.title('vilion plot of age to survival status')
plt.show()
```

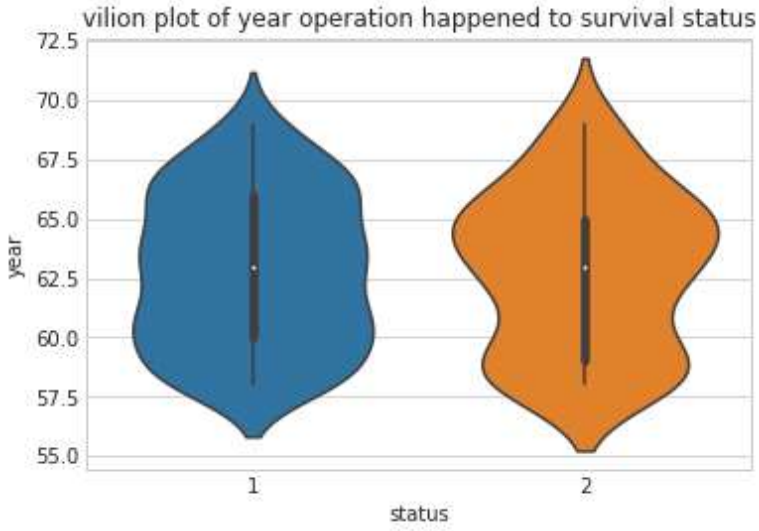


In [184]:
```
sns.violinplot(x="status", y="nodes", data=data_df, size=6)
plt.title('vilion plot of nodes survival status')
plt.show()
```



1. The probability for status 1 is more for nodes = 0
2. The Probability for status 2 is more for nodes > 1

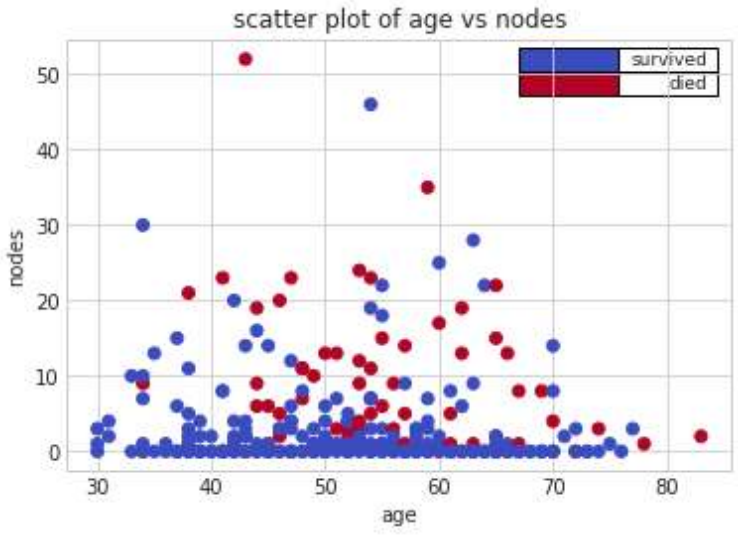
In [185]:
```
sns.violinplot(x="status", y="year", data=data_df, size=6)
plt.title('vilion plot of year operation happened to survival status')
plt.show()
```



1. More non-survival cases happened in the year 65 and lowest in the year 61.

BiVariate Analysis

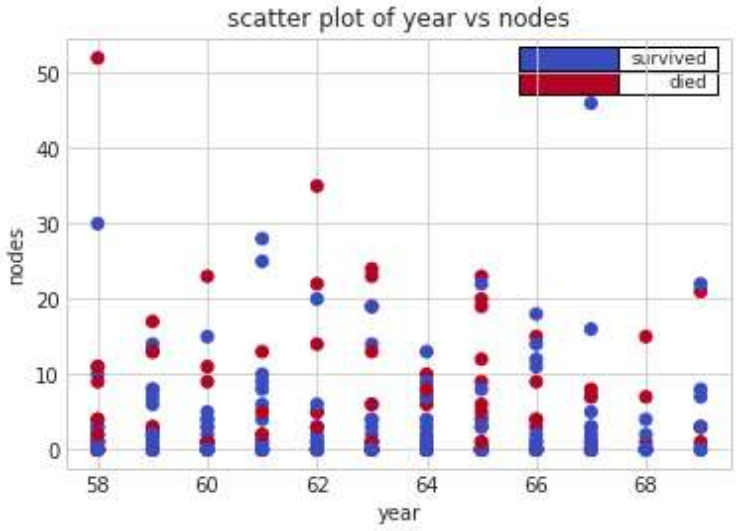
In [172]:
```
cmap = plt.cm.get_cmap("coolwarm",2)
plt.scatter(x = data_df['age'], y = data_df['nodes'], c = data_df['status'], la
bel = data_df['status'],
            cmap = cmap)
plt.xlabel('age')
plt.ylabel('nodes')
unique_classes = [1,2]
labels = ['survived', 'died']
plt.table(cellText=[[x] for x in labels], loc='best',
          colWidths=[0.15],rowColours=cmap(np.array(unique_classes)-1))
plt.title( 'scatter plot of age vs nodes')
plt.show()
```



Observations:

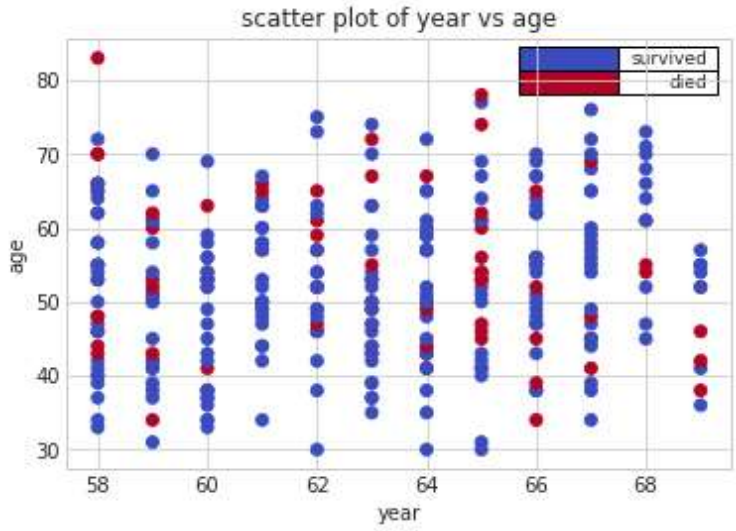
1. All people with zeros nodes survived. if no.of nodes = 0, status = 1.
2. People of 30 to 85 age group are present and no.of nodes is not correlated with age.

In [173]:
```
plt.scatter(x = data_df['year'], y = data_df['nodes'], c = data_df['status'], c
map = 'coolwarm' )
plt.xlabel('year')
plt.ylabel('nodes')
unique_classes = [1,2]
labels = ['survived', 'died']
plt.table(cellText=[[x] for x in labels], loc='best',
          colWidths=[0.15],rowColours=cmap(np.array(unique_classes)-1))
plt.title( 'scatter plot of year vs nodes')
plt.show()
```



Observations: nodes and year are also not correlated

In [174]:
```
plt.scatter(x = data_df['year'], y = data_df['age'], c = data_df['status'], cma
p = 'coolwarm' )
plt.xlabel('year')
plt.ylabel('age')
unique_classes = [1,2]
labels = ['survived', 'died']
plt.table(cellText=[[x] for x in labels], loc='best',
          colWidths=[0.15],rowColours=cmap(np.array(unique_classes)-1))
plt.title( 'scatter plot of year vs age')
plt.show()
```



Observations: most no.of non-surivial operations happend in the years 58 - 60 & 65-66

In [222]:
```
pp_df = data_df.copy()
pp_df['status'] = pp_df['status'].replace({1:'survived', 2: 'not-survived'})
sns.pairplot(pp_df, hue = 'status', vars = ['age', 'year', 'nodes'], size=4)
plt.title("Pair plots")
plt.show()
```

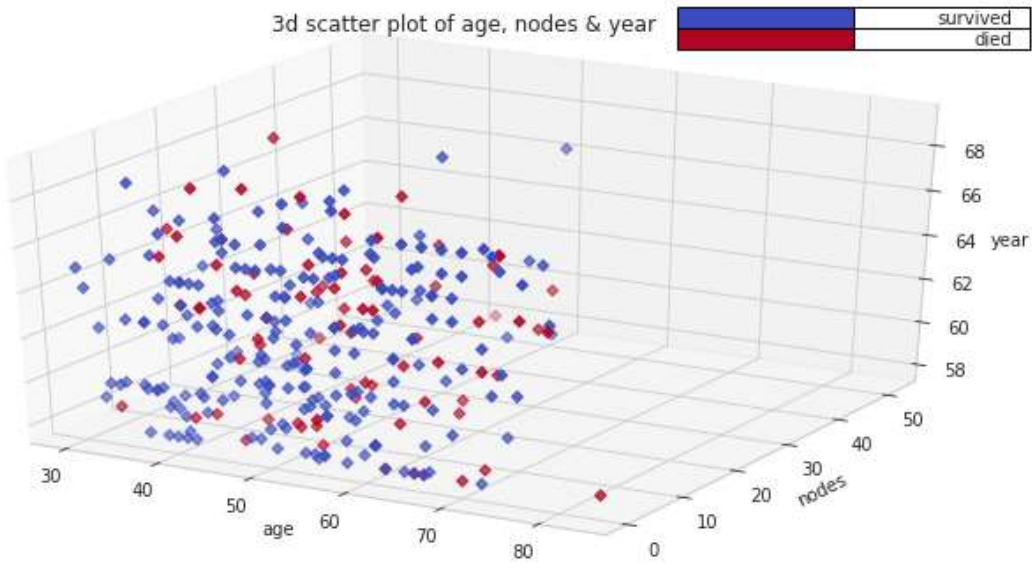
Observations:

1. nodes is the most important feature of all three

MultiVariate Analysis

In [182]:
```
from mpl_toolkits import mplot3d
fig = plt.figure(figsize = (12,6))
ax = plt.axes(projection='3d')
ax.scatter3D(data_df['age'], data_df['nodes'], data_df['year'], c = data_df['st
atus'], cmap='coolwarm', marker='D')
ax.set_xlabel('age')
ax.set_ylabel('nodes')
ax.set_zlabel('year')
plt.title('3d scatter plot of age, nodes & year')
unique_classes = [1,2]
labels = ['survived', 'died']
plt.table(cellText=[[x] for x in labels], loc='best',
          colWidths=[0.15],rowColours=cmap(np.array(unique_classes)-1))
```

Out[182]: <matplotlib.table.Table at 0x7f8339025978>



Observations: The data is not linearly separable, may be using decision trees could be better option.

Final Conclusions:

1. From the data we can say that operations are majorly performed on people age between 35 and 70 from year vs Age scatter plot.
2. There is good no.of people with nodes equal to 0 and they more likely to surivive as well.
3. Nodes above 30 are very rare in number.
4. Patients with age > 40 and nodes > 10 are more likely to not survive more than 5 years.
5. Most operations happend in the years 58 - 60 & 65-66 were unsuccessfull.
6. Patients within range of age 45-65 and had node > 1 are more likely to die. and those less than age 45 are more likely to survive though having nodes > 1
7. The patients with nodes > 1 and < 25 are the majority of patients who died.

Feature Importance:

1. nodes is most important feature in this dataset, as who had Axil node >= 1 those are more likely to die.
2. Age is also somewhat important feature as patients who aged less than 45 are likely to survive inspite of having node >= 1 and people with Age > 40 and nodes > 10 are not likely to survive.

In [117]:
```
# import plotly.plotly as py
# import plotly.graph_objs as go
```

In [116]:
```
# trace = go.Scatter3d(x=data_df['age'].values,
#     y=data_df['nodes'].values,
#     z=data_df['year'].values,
#     mode='markers',
#     marker=dict(size=12,line=dict(color='rgba(217, 217, 217, 0.14)',width=0.
5),
#     opacity=0.8))

# data = [trace]
# py.iplot(data)
```