1: importing required files.

In [4]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(color_codes=True)
from scipy.stats import skew,ttest_ind
from statsmodels.stats.proportion import proportions_ztest
```

In [5]:

```
insurance=pd.read_csv("insurance.csv")
```

2: displaying the data as dataframe

In [6]:

```
insurance.head()
```

Out[6]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [7]:

```
insurance['children']=insurance['children'].astype(str)
# This conversion was done as children is considered as categorical in later part of question.
```

3a. shape of the data

In [8]:

```
print("The shape of data is ",insurance.shape)
```

```
The shape of data is  (1338, 7)
```

3b. Data type of each attribute

In [9]:

```
print("The data type of each attribute is as follows")
print(insurance.info())
```

```
The data type of each attribute is as follows
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
age         1338 non-null int64
sex         1338 non-null object
bmi         1338 non-null float64
children    1338 non-null object
smoker      1338 non-null object
region      1338 non-null object
```

```
charges     1338 non-null float64
dtypes: float64(2), int64(1), object(4)
memory usage: 73.3+ KB
None
```

3c. Checking for missing values

In [10]:

```
insurance.isna()
```

Out[10]:

	age	sex	bmi	children	smoker	region	charges
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...
1333	False	False	False	False	False	False	False
1334	False	False	False	False	False	False	False
1335	False	False	False	False	False	False	False
1336	False	False	False	False	False	False	False
1337	False	False	False	False	False	False	False

1338 rows × 7 columns

Ans 3c. In the above table everthing is false and in answer of 3b we see that the total entries is equal to each column entries . This indicates absence of null values.

3d. 5 point summary of numerical attributes

In [11]:

```
print("The 5 point summary that is the minimun,25%,50%,75%,maximum of numerical attributes is as f
ollows")
insurance.describe()
```

```
The 5 point summary that is the minimun,25%,50%,75%,maximum of numerical attributes is as follows
```

Out[11]:

	age	bmi	charges
count	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	13270.422265
std	14.049960	6.098187	12110.011237
min	18.000000	15.960000	1121.873900
25%	27.000000	26.296250	4740.287150
50%	39.000000	30.400000	9382.033000
75%	51.000000	34.693750	16639.912515
max	64.000000	53.130000	63770.428010

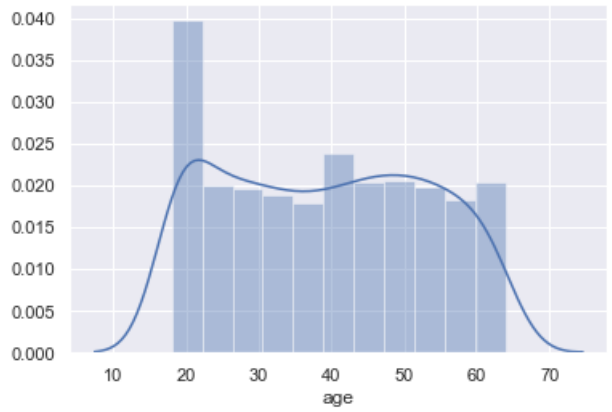
3e. DISTRIBUTION OF NUMERICAL ATTRIBUTES

In [12]:

```
sns.distplot(insurance["age"])
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b2e8a6c8>
```

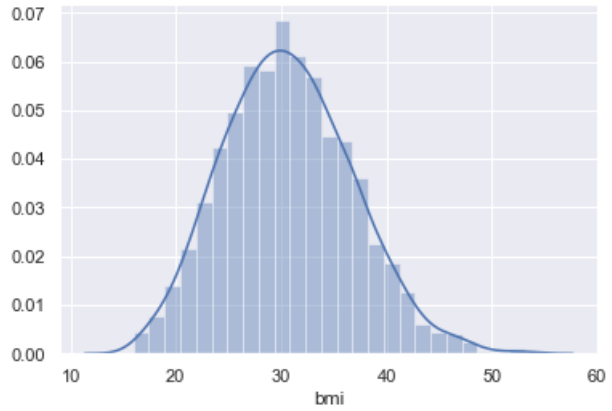


In [13]:

```
sns.distplot(insurance["bmi"])
```

Out[13]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4983948>
```

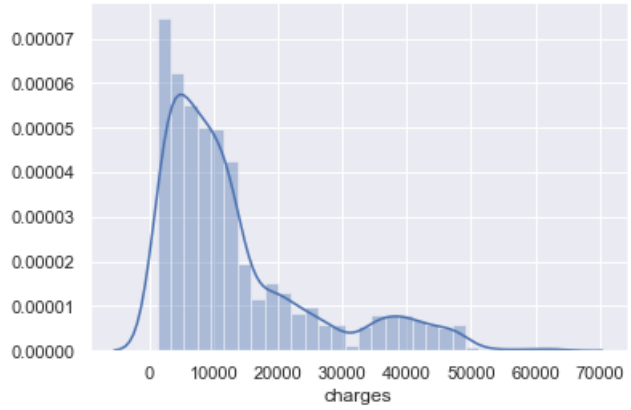


In [14]:

```
sns.distplot(insurance["charges"])
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4a0f648>
```



3f: Measuring of skewness of bmi, age and charges columns

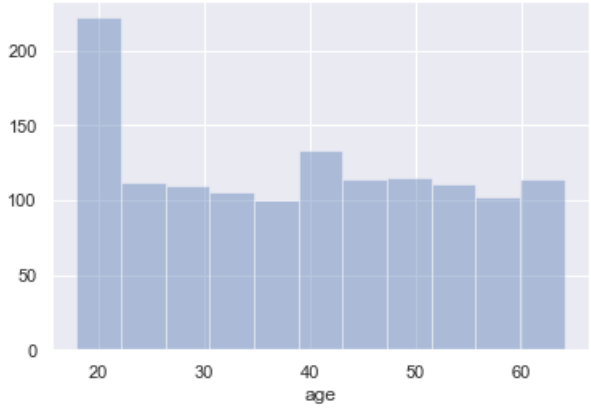
In [15]:

```
print("skewness of age = ",insurance['age'].skew())
sns.distplot(insurance["age"],kde=False)
```

```
skewness of age =  0.05567251565299186
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4b24c08>
```



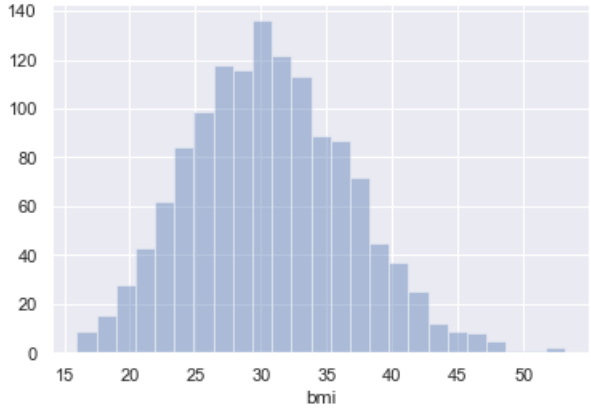
In [16]:

```
print("Skewness of bmi = ",insurance['bmi'].skew())
sns.distplot(insurance["bmi"],kde=False)
```

```
Skewness of bmi =  0.2840471105987448
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4b88dc8>
```



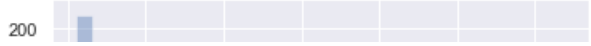
In [17]:

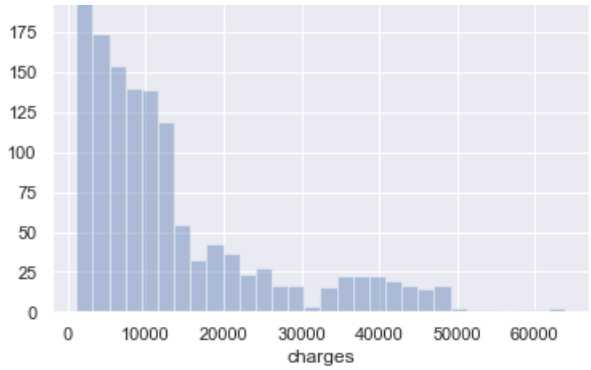
```
print("Skewness of charges = ",insurance['charges'].skew())
sns.distplot(insurance["charges"],kde=False)
```

```
Skewness of charges =  1.5158796580240388
```

Out[17]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4b8cf48>
```





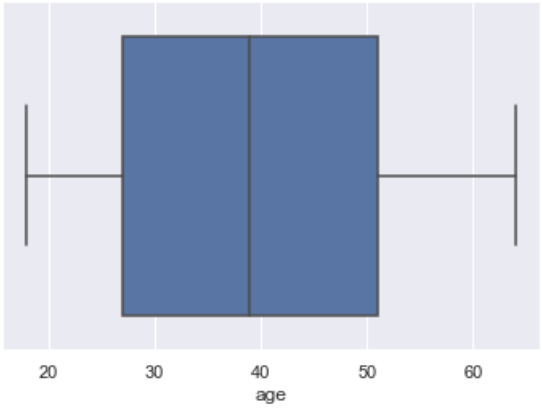
3g: Checking the presence of outliers in bmi, age and charges columns

In [18]:

```
sns.boxplot(insurance["age"])
```

Out[18]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b2e8c208>
```

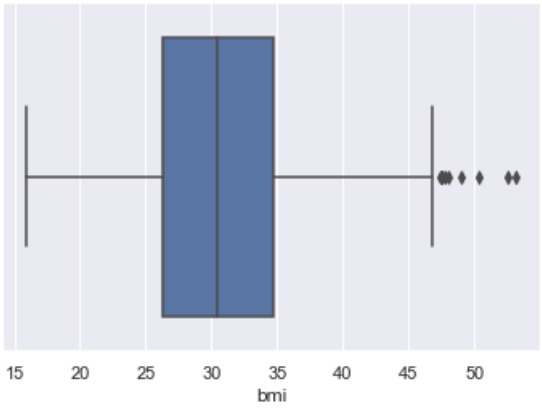


In [19]:

```
sns.boxplot(insurance["bmi"])
```

Out[19]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4d8ca08>
```

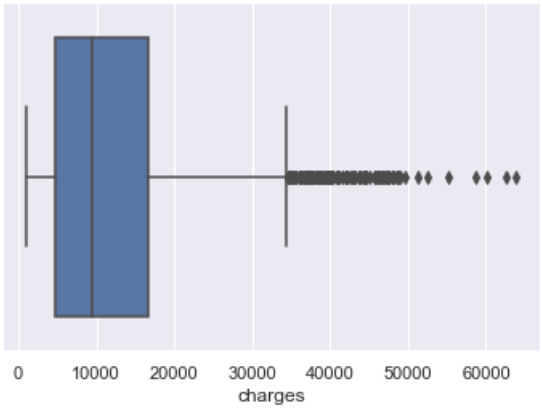


In [20]:

```
sns.boxplot(insurance["charges"])
```

Out[20]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4ddb308>
```



Thus presence of outliers is seen in bmi and charge. There is no outlier present in age.

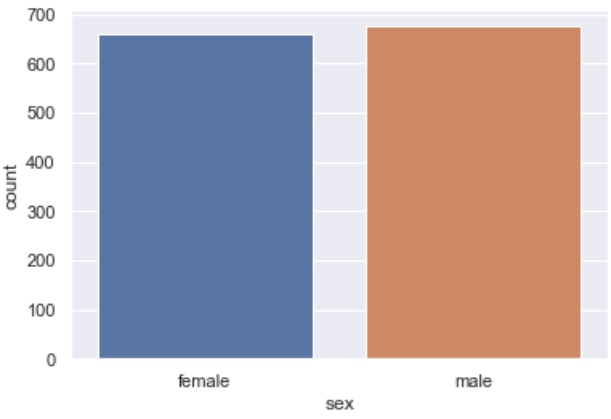
3h: Distribution of categorical columns

In [21]:

```
sns.countplot(insurance["sex"])
```

Out[21]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4e5a048>
```

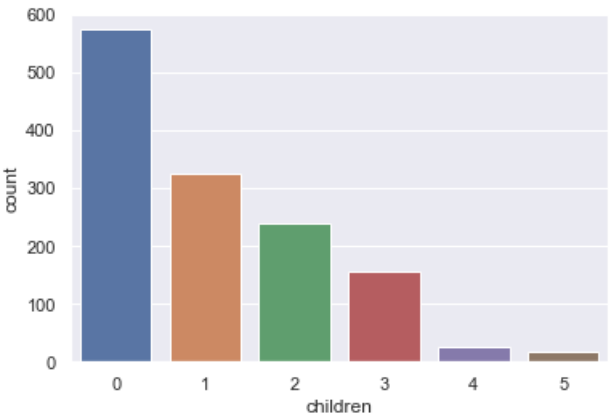


In [22]:

```
sns.countplot(insurance["children"])
```

Out[22]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4eadd48>
```

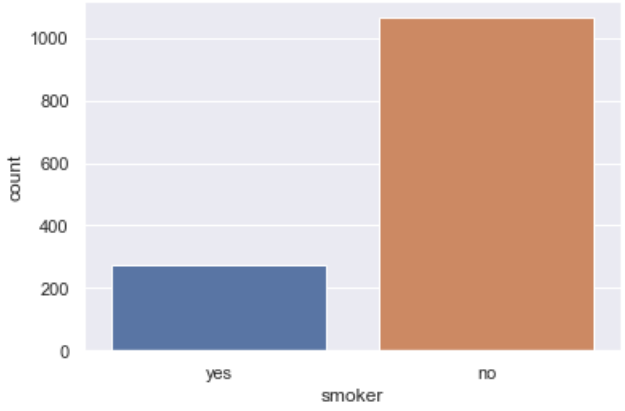


In [23]:

```
sns.countplot(insurance["smoker"])
```

Out[23]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4f02ec8>
```

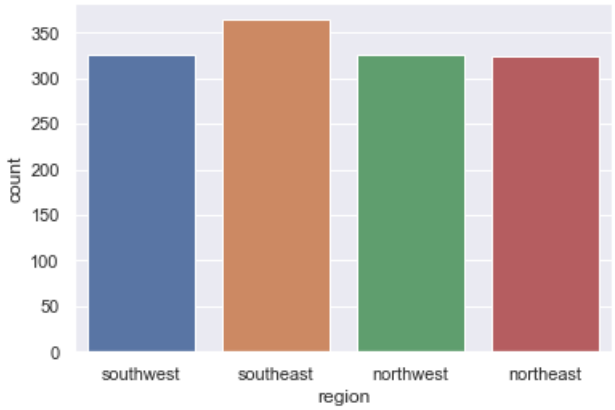


In [24]:

```
sns.countplot(insurance["region"])
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b4f614c8>
```



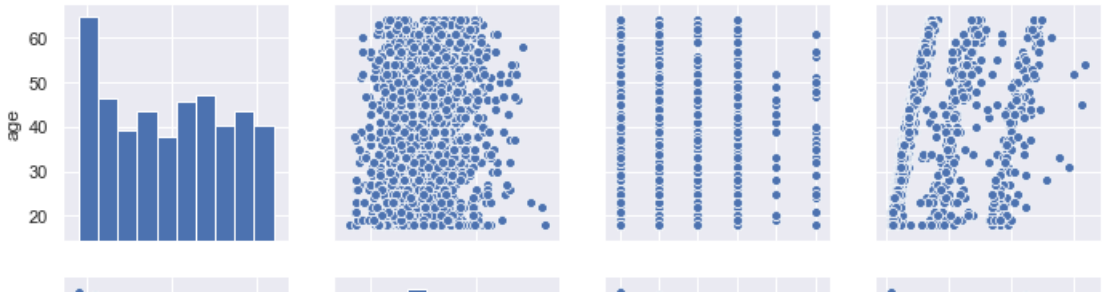
3i: Pair plot that includes all the columns of the data frame

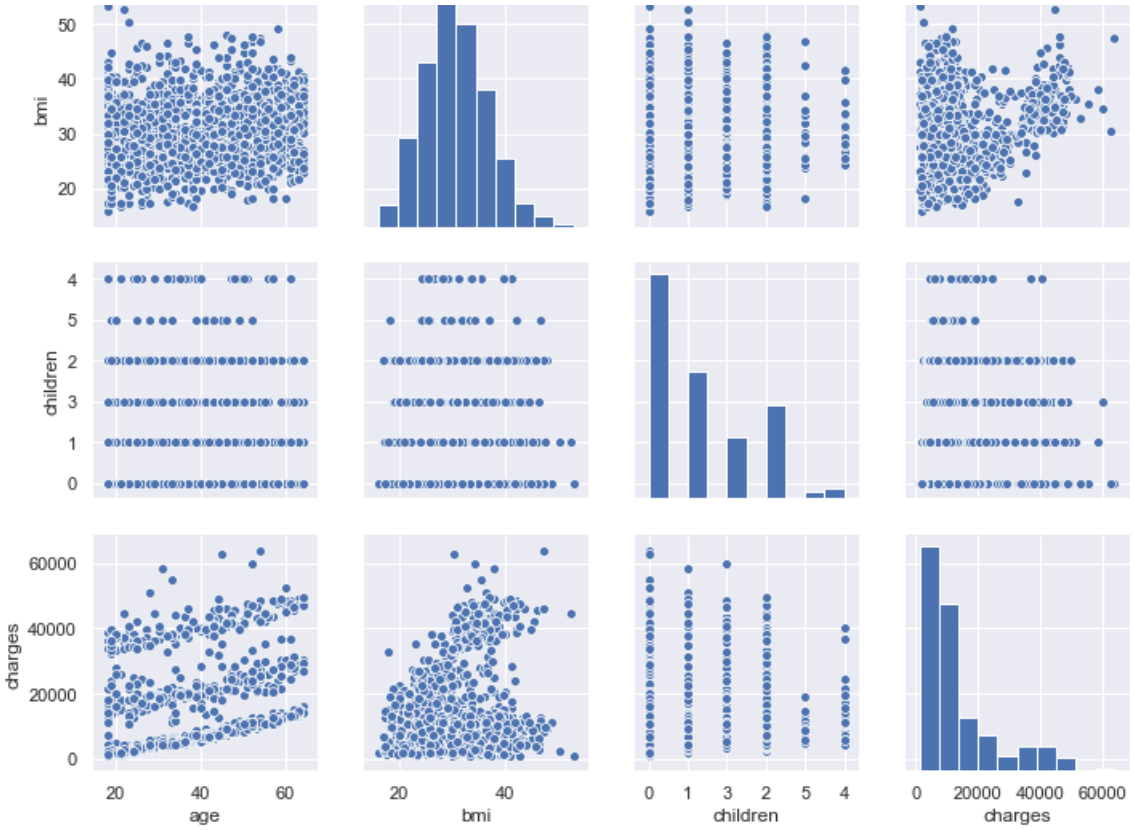
In [25]:

```
sns.pairplot(insurance[["age","sex","bmi","children","smoker","region","charges"]])
```

Out[25]:

```
<seaborn.axisgrid.PairGrid at 0x296b4ef7888>
```





4: Answering the following questions with statistical evidence

a. Do charges of people who smoke differ significantly from the people who don't?

b. Does bmi of males differ significantly from that of females?

c. Is the proportion of smokers significantly different in different genders?

d. Is the distribution of bmi across women with no children, one child and two children, the same?

Ans 4a: The following is the hypothesis.

```
    Null hypothessis = The mean charge of people who smoke is equal to people who don't
smoke.

    Alternate hypothesis = The mean charge of people who smoke is not equal to people who d
on't smoke.

    For the hypothesis testing we use paired t-test method.
```

In [26]:

```
a=insurance[insurance['smoker']=='yes'].charges
b=insurance[insurance['smoker']=='no'].charges
a_mean=np.mean(a)
b_mean=np.mean(b)
print("Average charge for smoker = ",a_mean," Average charge for non-smoker = ",b_mean)
```

```
Average charge for smoker =  32050.23183153285  Average charge for non-smoker =  8434.268297856199
```

In [27]:

```
t_statistic, p_value = ttest_ind(a , b)
print("t statistic = ",t_statistic," p value = ", p_value)
if p_value<0.05:
    print("The p value of",p_value,"is significant. Hence reject null hypothesis")
else:
    print("The p value of",p_value,"is not significant. Hence accept null hypothesis")
```

```
t statistic =  46.664921172723716  p value =  8.271435842177219e-283
The p value of 8.271435842177219e-283 is significant. Hence reject null hypothesis
```

4b: The following is the hyposthesis:

Null hypothesis = the mean bmi of male is equal to female.

Alternate hypothesis = the mean bmi of male is not euqal to female

For the hypothesis testing we use paired t-test method

In [28]:

```
male_group = insurance[insurance['sex']=='male'].bmi
female_group= insurance[insurance['sex']=='female'].bmi
t_statistic, p_value1 = ttest_ind(male_group,female_group)
print("t statistic =",t_statistic," p value =", p_value1)
if p_value1<0.05:
    print("The p value of",p_value1,"is significant. Hence reject null hypothesis")
else:
    print("The p value of",p_value1,"is not significant. Hence accept null hypothesis")
```

```
t statistic = 1.696752635752224  p value = 0.08997637178984932
The p value of 0.08997637178984932 is not significant. Hence accept null hypothesis
```

4c: The following is the hypothesis:

Null hypothesis = The proportion of male smokers is eqaul to the female smokers.

Alternate hypothesis = The proportion of male smokers is not equal to female smokers.

For the hypothesis testing we use z-test method

In [29]:

```
female_smokers = insurance[insurance['sex'] == 'female'].smoker.value_counts()[1]
male_smokers = insurance[insurance['sex'] == 'male'].smoker.value_counts()[1]
females = insurance.sex.value_counts()[1]
males = insurance.sex.value_counts()[0]
```

In [30]:

```
stat, pvalue_1 = proportions_ztest([female_smokers, male_smokers] , [females, males])
print("The stat value = ",stat," P value = ",pvalue_1)
if pvalue_1<0.05:
    print("The p value of",pvalue_1,"is significant. Hence reject null hypothesis")
else:
    print("The p value of",pvalue_1,"is not significant. Hence accept null hypothesis")
```

```
The stat value =  -2.7867402154855503  P value =  0.005324114164320532
The p value of 0.005324114164320532 is significant. Hence reject null hypothesis
```

4d: The following is the hypothesis:

Null hypothesis = The mean of bmi of females with different number of children are equal.

Alternate hypothesis = The mean of bmi of females with different number of children are not same.

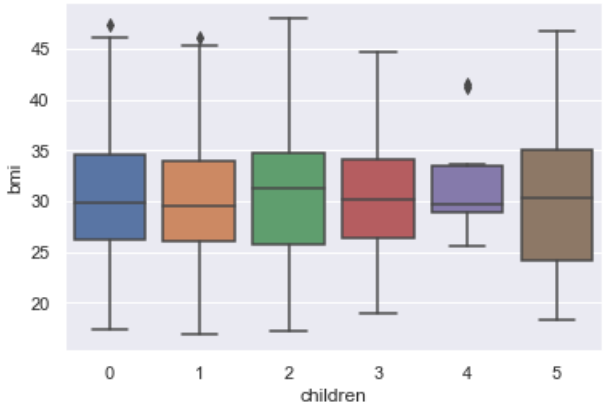
Since we have to analysis multiple means. We use ANOVA test.

In [31]:

```
data_2= insurance[insurance['sex']=='female'][['bmi','children']]
sns.boxplot(data_2["children"],data_2["bmi"])
```

Out[31]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x296b5877e08>
```



In [32]:

```
import statsmodels.api          as     sm
from   statsmodels.formula.api import ols

mod = ols('bmi ~ children', data = data_2).fit()
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```

```
                sum_sq     df         F    PR(>F)
children     53.274335    5.0  0.289914  0.918623
Residual  24109.180862  656.0       NaN       NaN
```

Here we see that the p value is greater than .05 at choosen level of significanceat 5%.

Thus statistically we accept the null hypothesis.

INFERENCES:

4a: It is seen that there is a significant difference for charge smoker and non smoker. The smokers paid more than non.smokers on average.

4b: It seen using statistical evidence that the difference in bmi of male from female is not significant.

4c: It seen that there is a significant difference in proportion of male to female smokers.

4d: It is seen that the distribution bmi of women was same irrespective of number of children.

In [ ]: