Homework of WEEK1 - Birthweight dataset

Let's start importing the dataset:

In [2]:

```
# %%
import os
import pandas as pd
path = os.path.join(os.getcwd(), 'datasets','Birthweight.csv')
dataset = pd.read_csv(path, sep=',', decimal='.')

dataset.head()
```

Out[2]:

	ID	Length	Birthweight	Headcirc	Gestation	smoker	mage	mnocig	mheight	mppwt	fage
0	1360	56	4.55	34	44	0	20	0	162	57	23
1	1016	53	4.32	36	40	0	19	0	171	62	19
2	462	58	4.10	39	41	0	35	0	172	58	31
3	1187	53	4.07	38	44	0	20	0	174	68	26
4	553	54	3.94	37	42	0	24	0	175	66	30

Q1. What is the mean birth weight for babies of non-smoking mothers?

In [175…

```
m = float(dataset['Birthweight'].mean())

print(f"The mean birth weight for babies of non-smoking mothers is {m:.3f} Kg.")
```

```
The mean birth weight for babies of non-smoking mothers is 3.313 Kg.
```

Q2. What is the mean birth weight for babies of smoking mothers?

In [176…

```
smoking_mothers = dataset[dataset['smoker']==1]

m = float(smoking_mothers['Birthweight'].mean())

print(f"The mean birth weight for babies of smoking mothers is {m:.3f} Kg.")
```

```
The mean birth weight for babies of smoking mothers is 3.134 Kg.
```

Q3. What is the mean head circumference for babies of non-smoking mothers?

In [177…

```
non_smokers = dataset[dataset['smoker']==0]
m = float(non_smokers['Headcirc'].mean())

print(f"The mean head circumference for babies of non-smoking mothers {m:.3f} cm")
```

```
The mean head circumference for babies of non-smoking mothers 35.050 cm
```

Q4. What is the mean gestational age at birth for babies of smoking mothers?

In [178…

```python
smoking_mothers = dataset[dataset['smoker']==1]
m = float(smoking_mothers['Gestation'].mean())

print(f"The mean gestational age at birth for babies of smoking mothers {m:.3f} weeks.")
```

```
The mean gestational age at birth for babies of smoking mothers 38.955 weeks.
```

Q5. What is the maximum head circumference for babies of non-smoking mothers?

In [179…

```python
non_smokers = dataset[dataset['smoker']==0]
m = float(non_smokers['Headcirc'].max())

print(f"The maximum head circumference for babies of non-smoking mothers {m:.3f} cm.")
```

```
The the maximum head circumference for babies of non-smoking mothers 39.000 cm.
```

Q6. What is the minimum gestational age at birth for babies of smoking mothers?

In [180…

```python
smoking_mothers = dataset[dataset['smoker']==1]
m = float(smoking_mothers['Gestation'].min())

print(f"The minimum gestational age at birth for babies of smoking mothers {m:.3f} months.")
```

```
The minimum gestational age at birth for babies of smoking mothers 33.000 months.
```

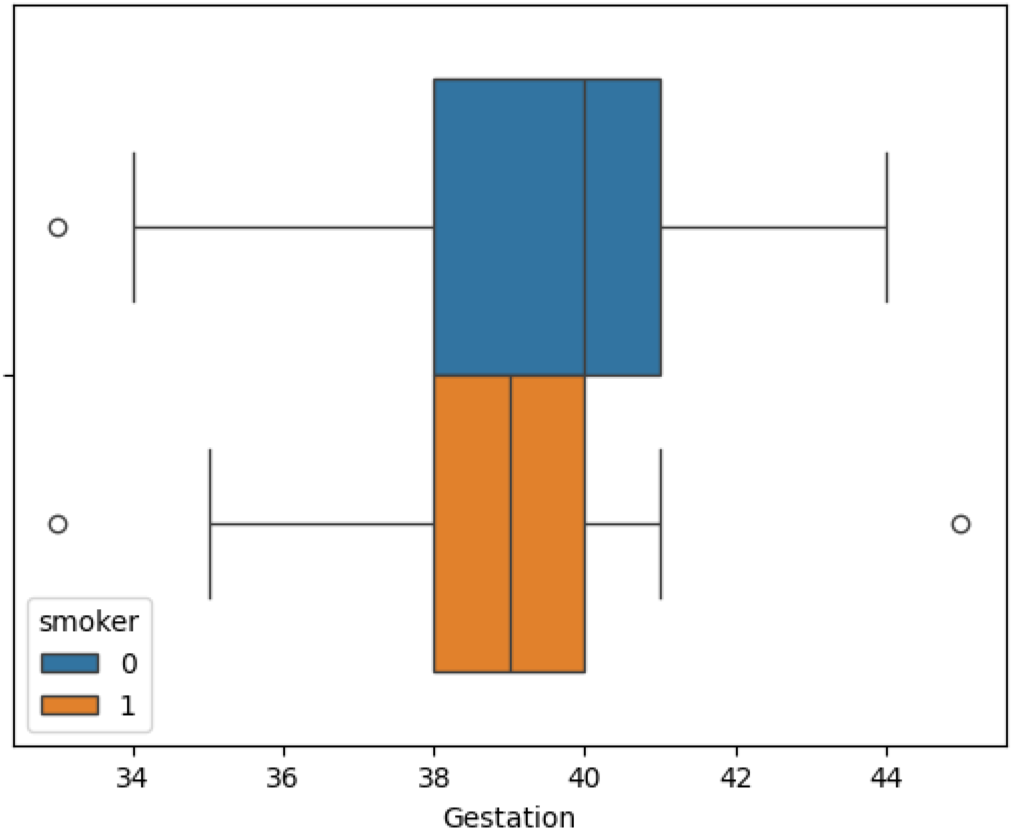
Q7. Based on the dataset you have, out of the two, which one would be a better bet:

```
+ Pregnancy period in smoking mothers is shorter
+ Pregnancy period in non-smoking mothers is shorter
```

In [18]:

```python
import seaborn as sns
from scipy.stats import ttest_ind
sns.boxplot(data=dataset, x='Gestation', hue='smoker')

smoking_mothers = dataset[dataset['smoker']==1]['Gestation']
non_smokers_mothers = dataset[dataset['smoker']==0]['Gestation']
print(ttest_ind(smoking_mothers, non_smokers_mothers))

print("Based on the below distribution, pregnancy period in smoking mothers is shorter.")
```

```
TtestResult(statistic=np.float64(-0.601934634696611), pvalue=np.float64(0.5506145436931532), d
f=np.float64(40.0))
Based on the below distribution, pregnancy period in smoking mothers is shorter.
```



Q8. Justify the above choice in a few words.

The graph above shows the distribution of pregnancy duration, separated by smoking and non-smoking mothers. Clearly, non-smoking mothers have a greater median pregnancy duration. However, this difference is not statistically significant, based on a t-test comparing the means.

Q9. What is the baby birth weight range for babies of smoking mothers?

In [19]:

```
smoking_mothers = dataset[dataset['smoker']==1]['Gestation']
print(f"The range is {smoking_mothers.min():.2f}-{smoking_mothers.max():.2f}")
```

```
The range is 33.00-45.00
```

Q10. In your own words describe what the value of the above range for baby's birthweight tells us about smoking versus non-smoking mothers?

The range above is not sufficient to draw any conclusions about the influence of cigarette use. Although the range is visibly shorter (excluding outliers), it is not statistically significant. There may be hidden variables (such as genetics and other health habits) influencing the results.

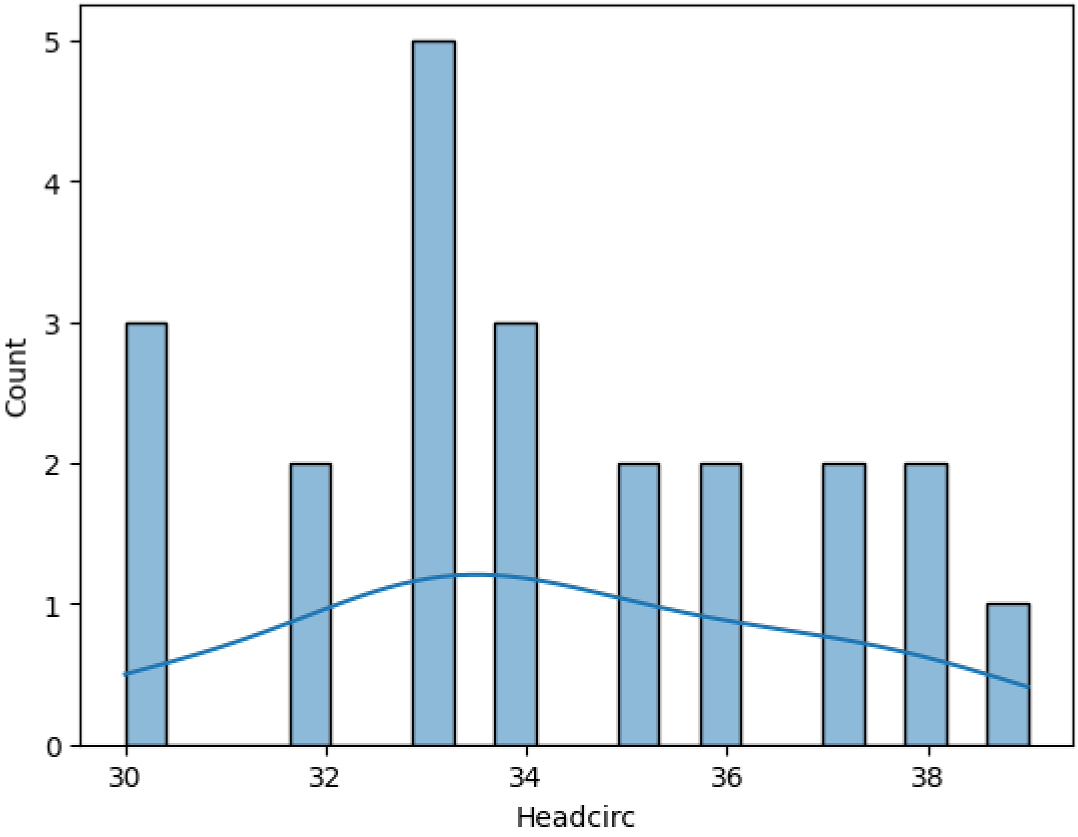
Q11. Are head circumference data for babies of smoking mothers normally distributed?

In [181…

```
import seaborn as sns

smoking_mothers_head_circ = dataset[dataset['smoker']==1]['Headcirc']
```

```python
sns.histplot(data=smoking_mothers_head_circ, bins=len(smoking_mothers_head_circ), kde=True)

print("At first, it does not seems to be normally distributed.")
```

```
At first, it does not seems to be normally distributed.
```



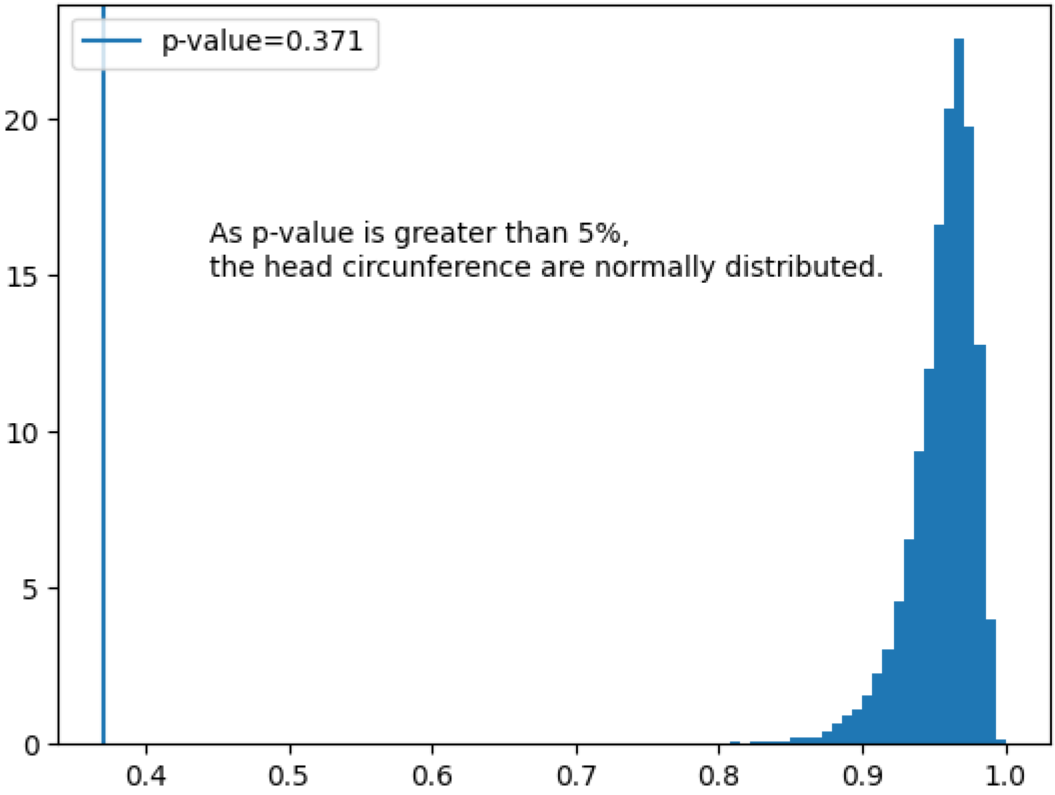
Q12. What is the significance value for the above on the Shapiro-Wilk test?

In [94]:

```python
from scipy.stats import shapiro, monte_carlo_test
import scipy.stats as st
import numpy as np
import matplotlib.pyplot as plt

def shapiro_stats(x):
    return shapiro(x).statistic

smoking_mothers_head_circ = dataset[dataset['smoker']==1]['Headcirc']
ML_data = monte_carlo_test(smoking_mothers_head_circ, rvs=st.norm.rvs,
                           statistic=shapiro_stats, alternative='less')
# shapiro(smoking_mothers_birth_wgt)
fig, ax = plt.subplots()
bins = np.linspace(0.65, 1, 50)
ax.hist(ML_data.null_distribution, density=True, bins=bins)
ax.axvline(ML_data.pvalue, label=f'p-value={ML_data.pvalue:.3f}')
ax.legend(loc='upper left')
ax.annotate('As p-value is greater than 5%, \nthe head circunference are normally distributed
```

Out[94]:

```
MonteCarloTestResult(statistic=np.float64(0.953652491270578), pvalue=np.float64(0.3707), null
_distribution=array([0.9676    , 0.97367155, 0.96610278, ..., 0.97362656, 0.90155096,
       0.95475988]))
```



As p-value is greater than 5%,
the head circumference are normally distributed.

Q13. What is the standard score (Z-score) for head circumference of 35.05 (X=35.05) in non-smoking mothers?

In [44]:

```
smoking_mothers_head_circ = dataset[dataset['smoker']==0]['Headcirc']

def z_score(x, mean, desvpad):
    if len(x)==1:
        return (x[0]-mean)/desvpad
    else:
        return [(x_i-mean)/desvpad for x_i in x]


z = z_score([35.05], smoking_mothers_head_circ.mean(),smoking_mothers_head_circ.std())
print(f"Z-score is {z:.2f}")
```

```
Z-score is 0.00
```

Q14. How are birth weight data of non-smoking mothers skewed?

In [47]:

```
from scipy.stats import skew

smoking_mothers_birth_weight = dataset[dataset['smoker']==0]['Birthweight']

s = skew(smoking_mothers_birth_weight)

print(f"The skewness of birth weight data of non-smoking mothers is {s:.3f}.")
```

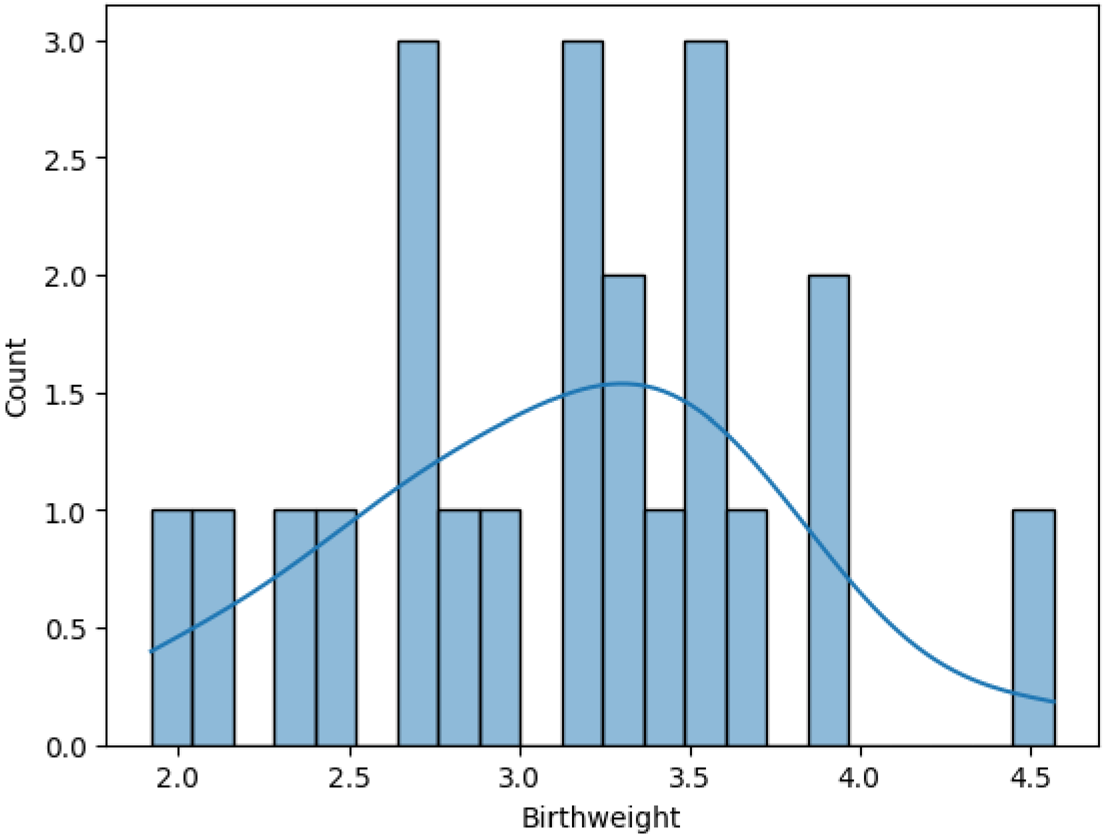
```
The skewness of birth weight data of non-smoking mothers is 0.333.
```

Q15. Are birth weight data for babies of smoking mothers normally distributed?

In [50]:

```
import seaborn as sns

smoking_mothers_birth_wgt = dataset[dataset['smoker']==1]['Birthweight']
sns.histplot(data=smoking_mothers_birth_wgt, bins=len(smoking_mothers_birth_wgt), kde=True)

print("It seems to be a normal distribution.")
```

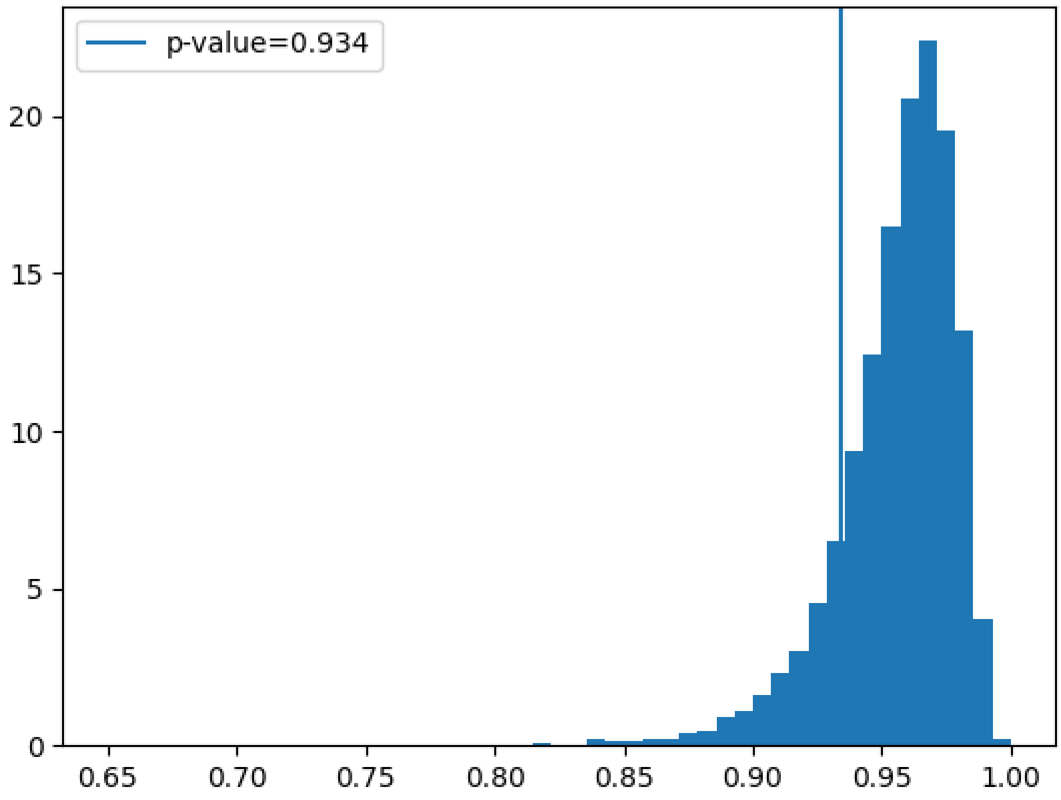
```
It seems to be a normal distribution.
```



Q16. What is the significance value for the above on the Shapiro-Wilk test?

In [95]:

```
from scipy.stats import shapiro, monte_carlo_test
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as st

def shapiro_stats(x):
    return shapiro(x).statistic

smoking_mothers_birth_wgt = dataset[dataset['smoker']==1]['Birthweight']
ML_data = monte_carlo_test(smoking_mothers_birth_wgt, rvs=st.norm.rvs,
                           statistic=shapiro_stats, alternative='less')
# shapiro(smoking_mothers_birth_wgt)
fig, ax = plt.subplots()
bins = np.linspace(0.65, 1, 50)
ax.hist(ML_data.null_distribution, density=True, bins=bins)
ax.axvline(ML_data.pvalue, label=f'p-value={ML_data.pvalue:.3f}')
ax.legend(loc='upper left')
ax.annotate('As p-value is greater than 5%, \nthe weights are normally distributed.', xy=(ML_
Out[95]:  Text(1.1208, 15, 'As p-value is greater than 5%, \nthe weights are normally distributed.')
```



Q17. Based on the dataset you have, how confident can you be in saying that a baby's birth weight will be +/- 1 standard deviation from the mean?

Based on Shapiro test, I would be pretty confident about make conclusions using a normal distribution.

Q18. Based on the dataset you have, what is the probability that the birth weight for a baby of a smoking mother will be less than 4.2 kg?

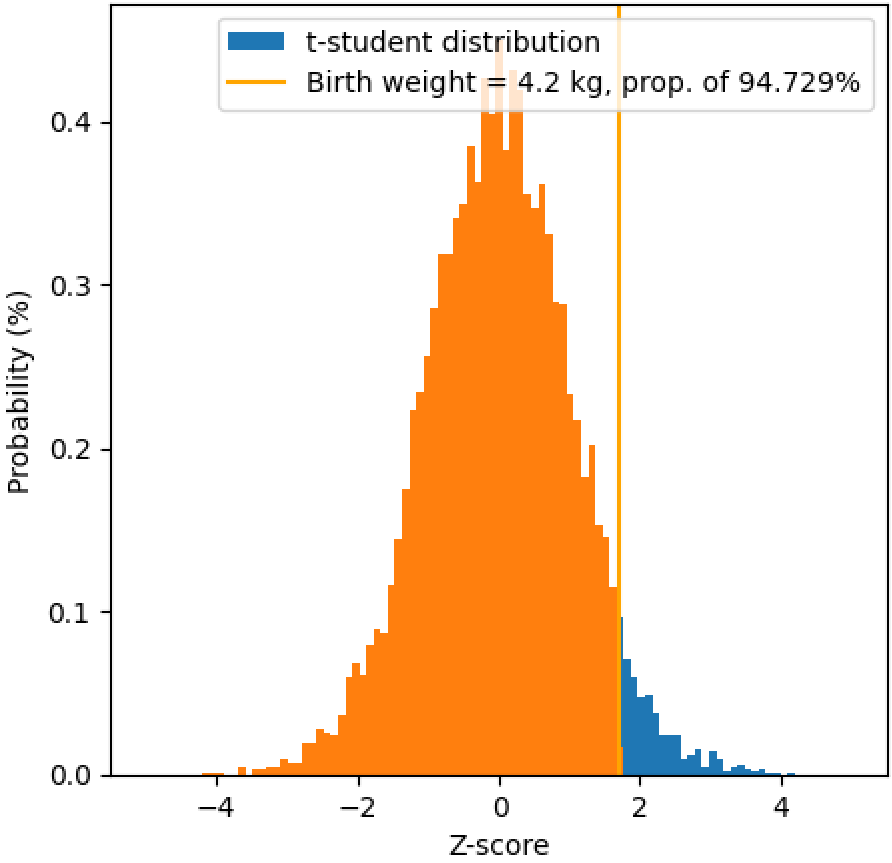
In [172…

```
from scipy.stats import t
import matplotlib.pyplot as plt
smoking_mothers_birth_wgt = dataset[dataset['smoker']==1]['Birthweight']

ddof = len(smoking_mothers_birth_wgt)

r = t.rvs(df= ddof, size=10000)

fig, ax = plt.subplots(figsize=(5,5))

bins = np.linspace(-5,5,100)
ax.hist(r, bins=bins, density=True,label='t-student distribution')

def z_score(x, mean, desvpad):
    if len(x)==1:
        return (x[0]-mean)/desvpad
    else:
        return [(x_i-mean)/desvpad for x_i in x]

X = 4.2
Z = z_score([X], smoking_mothers_birth_wgt.mean(), desvpad=smoking_mothers_birth_wgt.std())
```

```
prop = t.cdf(Z, ddof)
ax.axvline(Z, label=f'Birth weight = 4.2 kg, prop. of {prop:.3%}', color='orange')
ax.hist(r[r<Z], bins=bins, density=True)
ax.set_xlabel("Z-score")
ax.set_ylabel("Probability (%)")
ax.legend()
```

Out[172…

```
<matplotlib.legend.Legend at 0x1d381e09430>
```

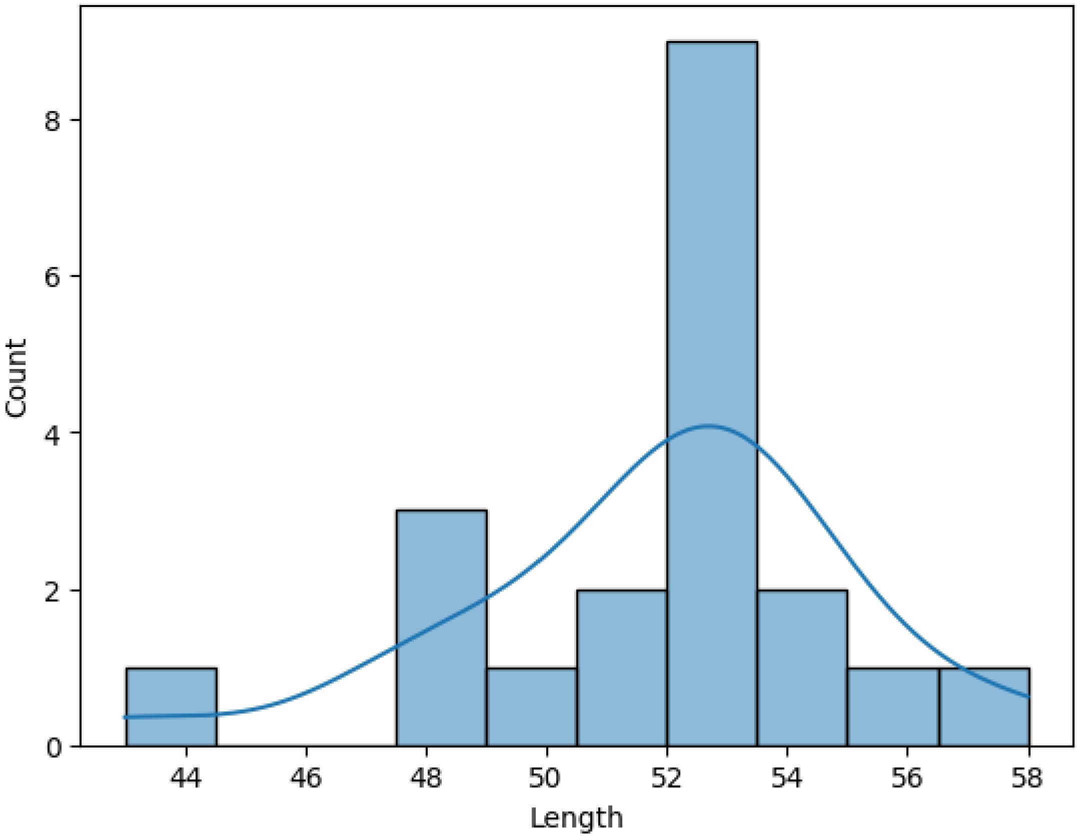


Q19. Are data for length of baby of non-smoking mothers normally distributed?

In [130…

```
import seaborn as sns

baby_length_non_smokers = dataset[dataset['smoker']==0]['Length']

sns.histplot(baby_length_non_smokers, kde=True)

print("Based on the below plot, probably it is not.")
```

```
Based on the below plot, probably it is not.
```



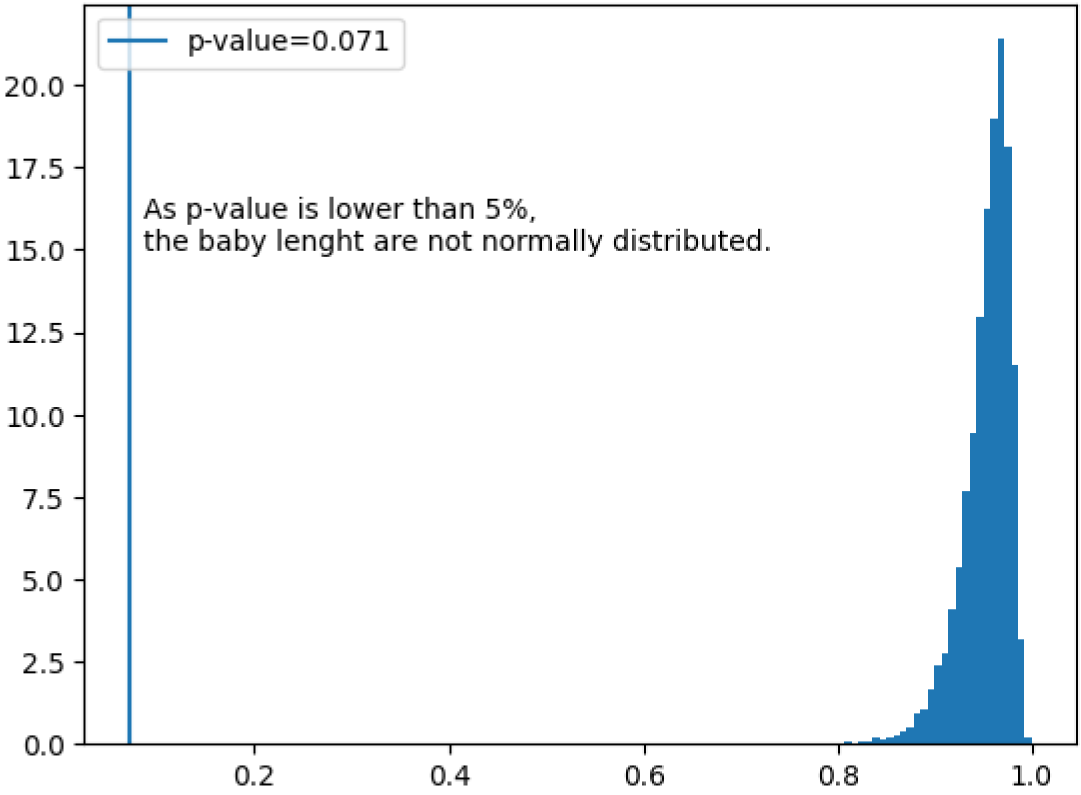
Q20. What is the significance value for the above on the Shapiro-Wilk test?

In [131…

```
from scipy.stats import shapiro, monte_carlo_test
import scipy.stats as st
import numpy as np
import matplotlib.pyplot as plt

def shapiro_stats(x):
    return shapiro(x).statistic

baby_length_non_smokers = dataset[dataset['smoker']==0]['Length']
ML_data = monte_carlo_test(baby_length_non_smokers, rvs=st.norm.rvs,
                           statistic=shapiro_stats, alternative='less')
fig, ax = plt.subplots()
bins = np.linspace(0.65, 1, 50)
ax.hist(ML_data.null_distribution, density=True, bins=bins)
ax.axvline(ML_data.pvalue, label=f'p-value={ML_data.pvalue:.3f}')
ax.legend(loc='upper left')
ax.annotate('As p-value is lower than 5%, \nthe baby lenght are not normally distributed.', x
```

Out[131…

```
Text(0.08568, 15, 'As p-value is lower than 5%, \nthe baby lenght are not normally distribute
d.')
```



Q21. What is the standard score for the length of a baby of 48.5cm for non-smoking mothers?

In [134…

```
baby_length_non_smokers = dataset[dataset['smoker']==0]['Length']

mean = baby_length_non_smokers.mean()
std = baby_length_non_smokers.std()


def z_score(x, mean, desvpad):
    if len(x)==1:
        return (x[0]-mean)/desvpad
    else:
        return [(x_i-mean)/desvpad for x_i in x]

print(f"The standard score for the length of a baby of 48.5cm is {z_score([48.5], mean, std):
```

```
The standard score for the length of a baby of 48.5cm is -1.014
```

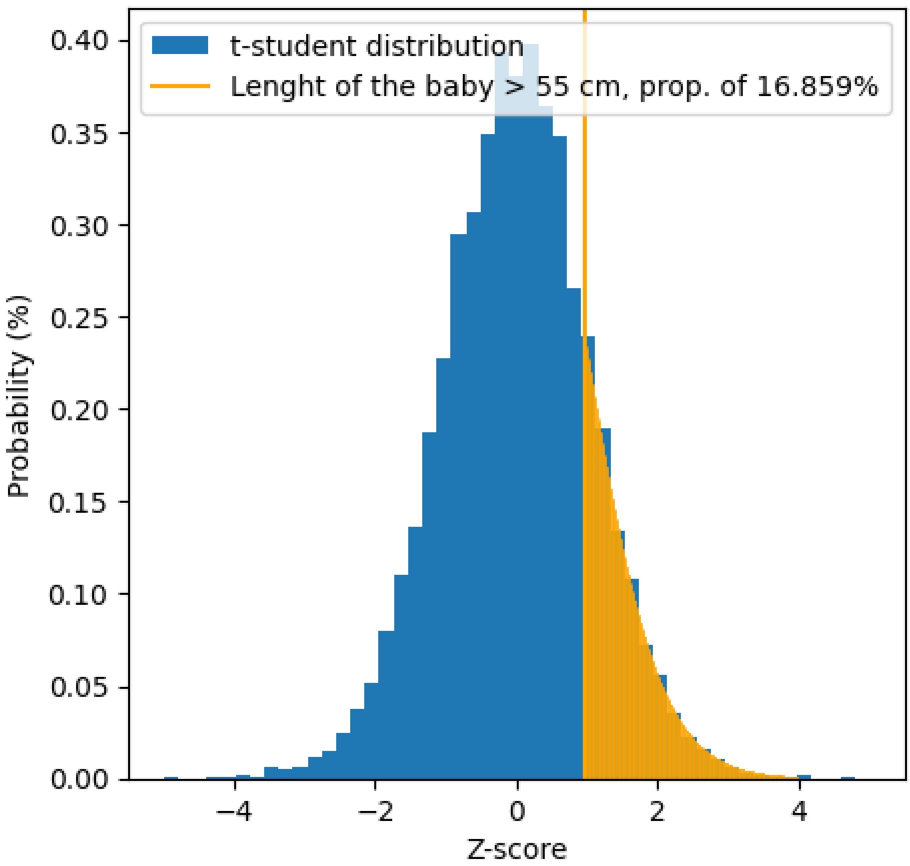
Q22. Based on the dataset you have, what is the probability that the length of baby for non-smoking mothers will be more than 55 cm?

In [171…

```
from scipy.stats import t
import matplotlib.pyplot as plt

baby_length_non_smokers = dataset[dataset['smoker']==0]['Length']
mean = baby_length_non_smokers.mean()
std = baby_length_non_smokers.std()


ddof = len(baby_length_non_smokers)

r = t.rvs(df= ddof, size=10000)
```

```
fig, ax = plt.subplots(figsize=(5,5))

bins = np.linspace(-5,5)

ax.hist(r, bins=bins, density=True,label='t-student distribution')

def z_score(x, mean, desvpad):
    if len(x)==1:
        return (x[0]-mean)/desvpad
    else:
        return [(x_i-mean)/desvpad for x_i in x]

X = 55
Z = z_score([X], mean, desvpad=std)

x_dist = np.linspace(Z,4,1000)
t_dist = t.pdf(x_dist, ddof)
prop = t.cdf(Z, ddof)
ax.axvline(Z, label=f'Lenght of the baby > 55 cm, prop. of {1-prop:.3%}', color='orange')
ax.set_xlabel("Z-score")
ax.set_ylabel("Probability (%)")
ax.bar(x_dist,t_dist, color='orange', width=0.01, alpha=0.5)
ax.legend()
```

```
Out[171…  <matplotlib.legend.Legend at 0x1d3841a39b0>
```


t-student distribution
Lenght of the baby > 55 cm, prop. of 16.859%
Probability (%)
Z-score