Haywood DCS530 Week #12 Project

1 Exploring the Link Between Pregnancy Frequency and Diabetes Risk

1.1 Introduction

Diabetes is a chronic medical condition that affects millions of people worldwide. According to the American Diabetes Association, diabetes is a group of metabolic disorders characterized by elevated blood sugar levels over an extended period. These elevated levels can result from either insufficient insulin production, ineffective utilization of insulin, or a combination of both (American Diabetes Association, 2022). Gestational diabetes, which occurs during pregnancy, can be particularly dangerous if left untreated, leading to complications for both the mother and the child.

The objective of this project is to explore the relationship between the number of pregnancies a woman has had and the risk of developing diabetes in the future. Using a dataset containing various health metrics, I will focus on five key variables: Pregnancies, Glucose, BloodPressure, BMI, and Outcome. By analyzing these variables, I aim to test the hypothesis that an increased number of pregnancies correlates with a higher risk of diabetes.

1.2 Hypothesis

An increased number of pregnancies correlates with a higher risk of diabetes

1.3 Variable Description

The dataset consists of the following variables:

- Pregnancies: Represents the number of times a woman has had pregnancies.
- Glucose: Plasma glucose concentration.
- BloodPressure: Diastolic blood pressure (mm Hg).
- SkinThickness: Triceps skinfold thickness (mm).
- Insulin: Represent a 2-Hour serum insulin (mu U/ml).
- BMI: Body mass index which is calculated as the weight in kilograms divided by the square of height in meters.
- Age: Age of the woman in years.
- DiabetesPedigreeFunction: Provides a score that indicates the likelihood of diabetes based on the individual's family medical history
- Outcome: Indicates whether the patient has diabetes (1) or not (0).

For this analysis, I will focus on the following five variables:

- Pregnancies
- Glucose
- BloodPressure
- BMI
- Age

1.4 Descriptive Statistics and Histograms

from os.path import basename, exists

def download(url): filename = basename(url) if not exists(filename): from urllib.request import urlretrieve

```
local, _ = urlretrieve(url, filename)
print("Downloaded " + local)
```

download("https://github.com/AllenDowney/ThinkStats2/raw/master/code/thinkstats2.py")
download("https://github.com/AllenDowney/ThinkStats2/raw/master/code/thinkplot.py")

[4]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import random
import thinkstats2
import thinkplot
```

[6]:
```
# Uploaded CSV file
df = pd.read_csv('diabetes.csv')

# Display the first few rows of the DataFrame
df.head()
```

[6]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI \
0	6	148	72	35	0	33.6
1	1	85	66	29	0	26.6
2	8	183	64	0	0	23.3
3	1	89	66	23	94	28.1
4	0	137	40	35	168	43.1

	DiabetesPedigreeFunction	Age	Outcome
0	0.627	50	1
1	0.351	31	0
2	0.672	32	1
3	0.167	21	0
4	2.288	33	1

[7]:
```
# Display basic information about the dataset
print(df.info())
```

```
print(df.describe())
```

```
<class
'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to
767 Data columns (total 9
columns):
```

```
#   Column                   Non-Null Count Dtype
--- ------                   -------------- -----
 0   Pregnancies             768 non-null  int64
 1   Glucose                 768 non-null  int64
 2   BloodPressure           768 non-null  int64
 3   SkinThickness           768 non-null  int64
 4   Insulin                 768 non-null  int64
 5   BMI                     768 non-null  float64
 6   DiabetesPedigreeFunction 768 non-   float64
     null
 7   Age                     768 non-null  int64
 8   Outcome                 768 non-null  int64
dtypes: float64(2),
int64(7) memory usage:
54.1 KB None
      Pregnancies  Glucose BloodPressure SkinThickness   Insulin \

count 768.000000 768.000000   768.000000    768.000000 768.000000
mean    3.845052 120.894531    69.105469    20.536458  79.799479
std      3.369578 31.972618     19.355807     15.952218 115.244002
min      0.000000   0.000000     0.000000     0.000000   0.000000
25%      1.000000 99.000000     62.000000     0.000000   0.000000
50%      3.000000 117.000000    72.000000    23.000000  30.500000
75%      6.000000 140.250000    80.000000     32.000000 127.250000
max     17.000000 199.000000   122.000000     99.000000 846.000000

           BMI DiabetesPedigreeFunction       Age    Outcome
count 768.000000 768.000000 768.000000 768.000000 mean
 31.992578  0.471876 33.240885    0.348958 std     7.884160
 0.331329 11.760232    0.476951 min     0.000000   0.078000
21.000000   0.000000 25%     27.300000  0.243750 24.000000
 0.000000
50%   32.000000  0.372500 29.000000    0.000000 75%
 36.600000  0.626250 41.000000    1.000000 max     67.100000
 2.420000 81.000000    1.000000
```

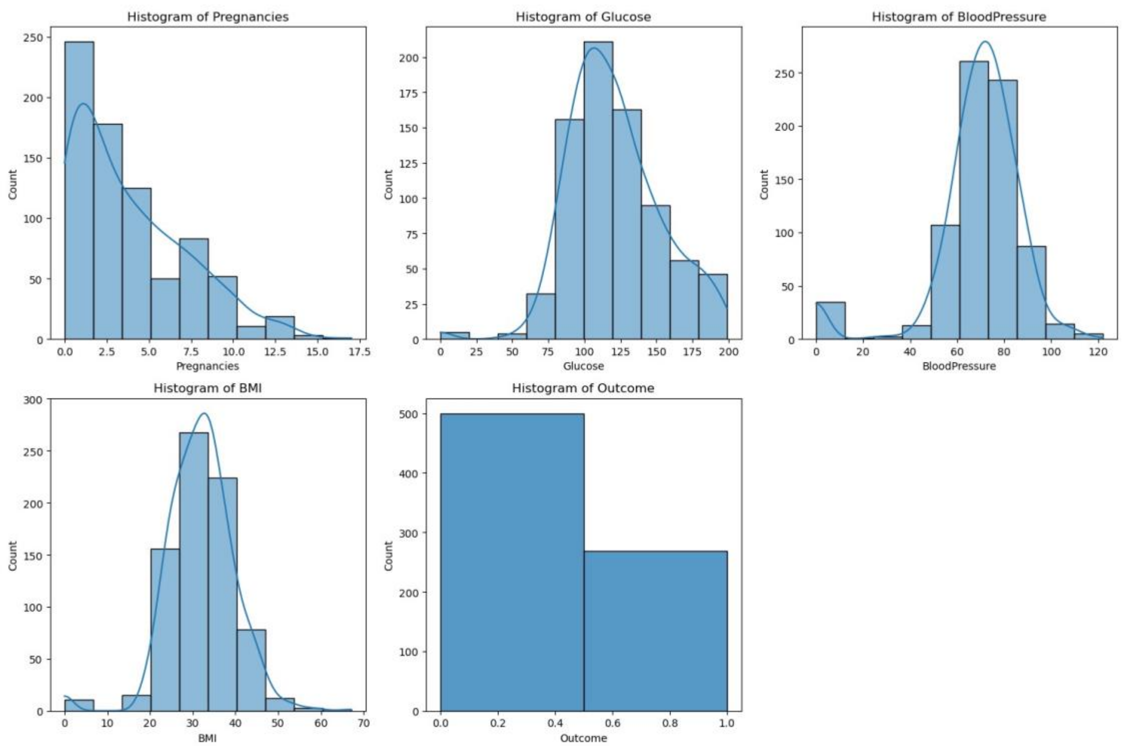
- Pregnancies: The number of pregnancies ranges from 0 to 17, with a mean of 3.85 and std of 3.37. High values (e.g., 17) could be outliers but it is biologically possible to have 17 kids, thus I will keep all high numbers.
- Glucose: Glucose levels vary between 0 to 199, with a mean of 120.89. The presence of zeros is unusual for glucose levels and they are likely missing data. Thus, they will be removed • SkinThickness: The measurements for skin thickness range from 0 to 99, with a mean of 20.54. The zeros in this variable likely indicate missing data.
- BloodPressure: Blood pressure range from 0 to 122, with a mean of 69.11 and a standard deviation of 19.36. Zero values are not physiologically realistic and likely represent missing data.
- Insulin: Insulin levels show a wide range from 0 to 846, with a mean of 79.80. Zeros suggests missing or unrecorded data.

- DiabetesPedigreeFunction: This function ranges from 0.078 to 2.42, with a mean of 0.47, showing a moderate variation in genetic predisposition to diabetes.
- BMI: Body Mass Index values range from 0 to 67.1, with an average of 31.99 and a standard deviation of 7.88. A BMI of 0 is not possible and represents missing data.
- Age: Participants' ages range from 21 to 81 years, with a mean age of 33.24. The standard deviation is 11.76, indicating a relatively young to middle-aged population.
- Outcome: The outcome variable, indicating diabetes presence (1) or absence (0.

1.5 Histograms for the Variables

[19]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
import warnings

# Suppress the specific FutureWarning related to 'mode.use_inf_as_na'
warnings.filterwarnings("ignore", category=FutureWarning, message=".
↪*use_inf_as_na.*")

# Assuming df is your DataFrame
# Convert infinite values to NaN
df.replace([np.inf, -np.inf], np.nan, inplace=True)

# Descriptive statistics
desc_stats = df[['Pregnancies', 'Glucose', 'BloodPressure', 'BMI', 'Outcome']].
↪describe()

# Histograms
fig, axes = plt.subplots(2, 3, figsize=(15, 10))

sns.histplot(df['Pregnancies'], bins=10, kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Histogram of Pregnancies')

sns.histplot(df['Glucose'], bins=10, kde=True, ax=axes[0, 1])
axes[0, 1].set_title('Histogram of Glucose')

sns.histplot(df['BloodPressure'], bins=10, kde=True, ax=axes[0, 2])
axes[0, 2].set_title('Histogram of BloodPressure')

sns.histplot(df['BMI'], bins=10, kde=True, ax=axes[1, 0])
axes[1, 0].set_title('Histogram of BMI')

sns.histplot(df['Outcome'], bins=2, kde=False, ax=axes[1, 1])
```

```
axes[1, 1].set_title('Histogram of Outcome ')

# Hide the empty subplot
axes[1, 2].axis('off')

plt.tight_layout()
plt.show()
```



- Pregnancies: The histogram shows a right-skewed distribution with most women having fewer pregnancies. There are some outliers with a high number of pregnancies. Outliers will be kept as it is biologically possible to have babies over 15
- Glucose: The distribution of glucose levels is approximately normal with a peak around the 100-150 range. Higher glucose levels are more common in diabetic patients. Zero values will be removed.
- BloodPressure: Blood pressure values are somewhat normally distributed with a few outliers on both ends. Zero values will be removed.
- BMI: The BMI distribution is right-skewed, indicating more women have lower BMI values. Higher BMI values are associated with a higher risk of diabetes. Zero values will be removed.
- Outcome: The binary outcome shows the proportion of non-diabetic (0) and diabetic (1) women in the dataset.

1.6 Data Cleaning

[8]:
```
# Determine the number of rows and columns
print(f"Dimensions: {df.shape}")
```

```
Dimensions: (768, 9)
```

[9]:
```python
# Remove duplicate rows
df_clean =
df.drop_duplicates()
```

[10]:
```python
# Replace zero values with the mean or median of respective columns
columns_with_zeros = ['Glucose', 'BloodPressure', 'SkinThickness',
'Insulin',␣ ↪'BMI']

for column in columns_with_zeros:
    if column in ['Glucose', 'BloodPressure', 'SkinThickness',
    'Insulin']:
        mean_value = df_clean[df_clean[column] != 0][column].mean()
        df_clean[column] = df_clean[column].replace(0, mean_value)
    elif column == 'BMI':
        median_value = df_clean[df_clean[column] != 0][column].median()
        df_clean[column] = df_clean[column].replace(0, median_value)
```

[11]:
```python
# Display the cleaned dataset
df_clean.head()
```

[11]:
```
 Pregnancies Glucose BloodPressure SkinThickness  Insulin   BMI \
0            6     148.0 72.0  35.00000 155.548223 33.6
1            1     85.0  66.0  29.00000 155.548223 26.6
2            8     183.0 64.0  29.15342 155.548223 23.3
3            1     89.0  66.0  23.00000   94.000000 28.1
4            0     137.0 40.0  35.00000 168.000000 43.1

   DiabetesPedigreeFunction Age Outcome
0      0.627 50    1 1    0.351 31    0 2
 0.672 32    1 3    0.167 21    0 4
 2.288 33    1
```

1.7 Descriptive Statistics for Clean Data

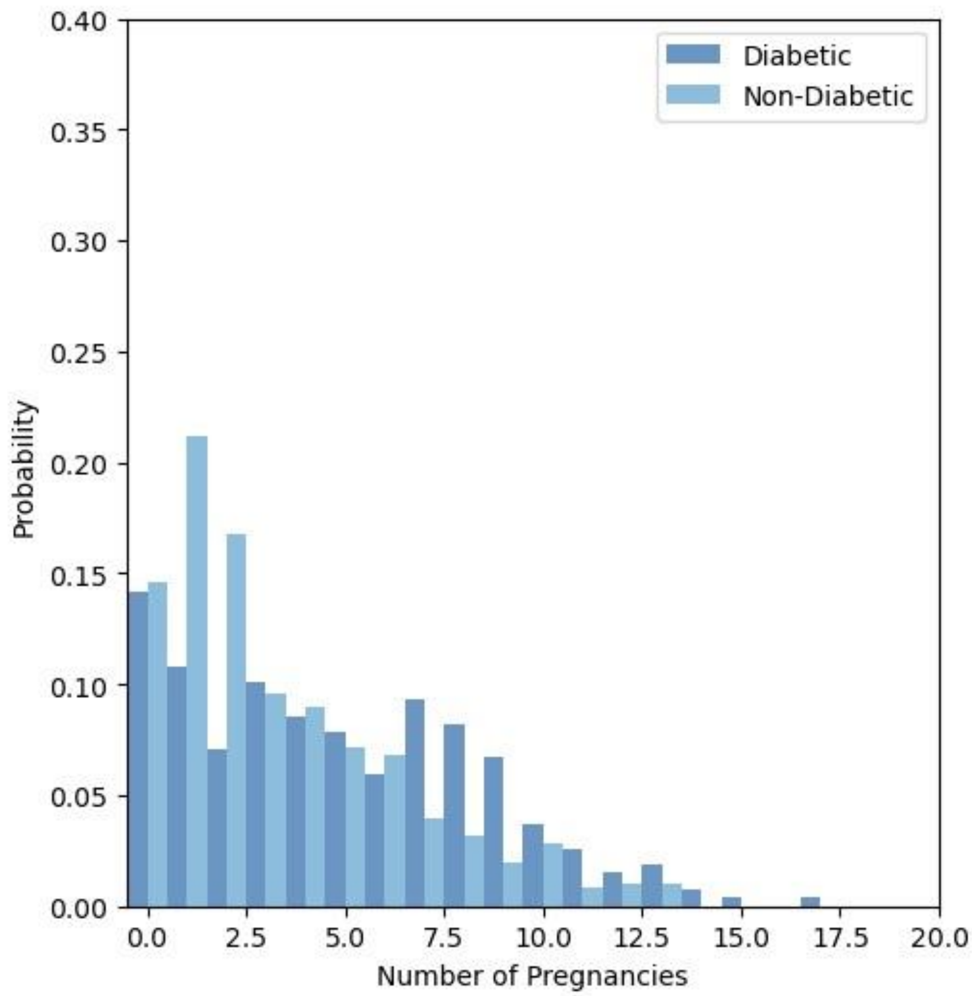
[33]:
```python
# Descriptive statistics
desc_stats = df_clean[['Pregnancies', 'Glucose', 'BloodPressure', 'BMI',␣
 ↪'Outcome']].describe()
print("Descriptive Statistics:")
print(desc_stats)
```

```
Descriptive Statistics:
       Pregnancies     Glucose  BloodPressure         BMI     Outcome
count  768.000000  768.000000   768.000000  768.000000  768.000000

mean     3.845052  121.686763     72.405184  32.455208    0.348958
```

std	3.369578	30.435949	12.096346	6.875177	0.476951
min	0.000000	44.000000	24.000000	18.200000	0.000000
25%	1.000000	99.750000	64.000000	27.500000	0.000000
50%	3.000000	117.000000	72.202592	32.300000	0.000000
75%	6.000000	140.250000	80.000000	36.600000	1.000000
max	17.000000	199.000000	122.000000	67.100000	1.000000

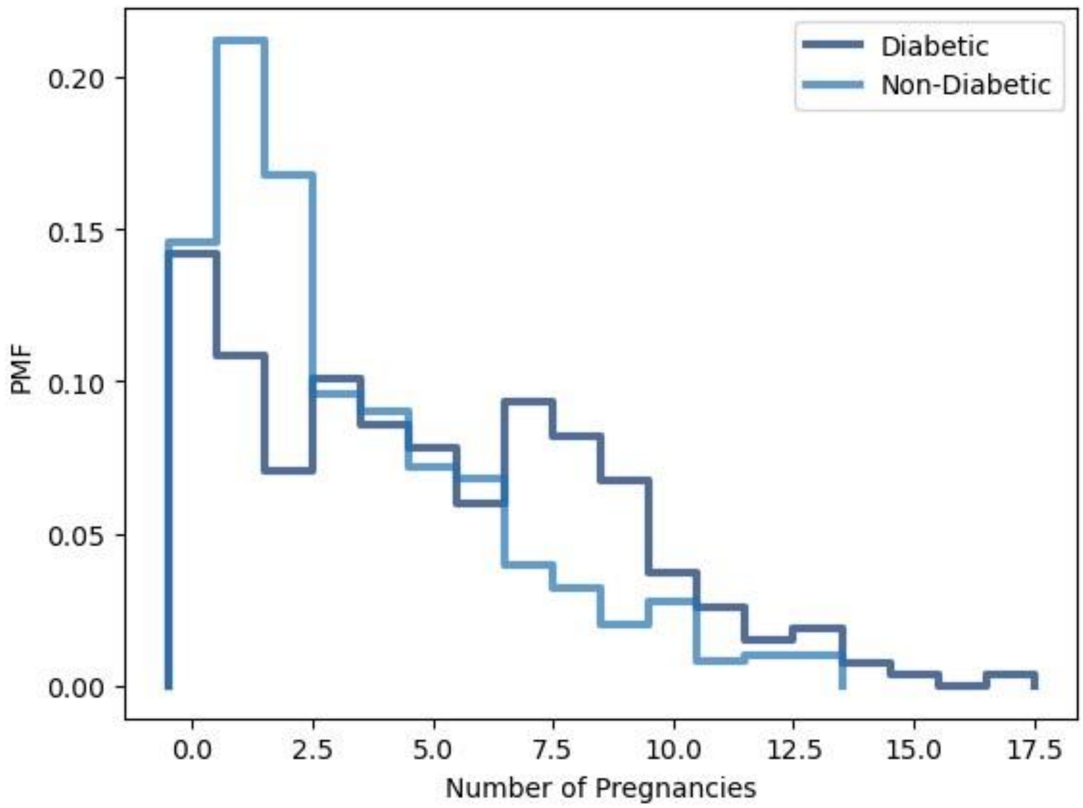
1.8 Probability Mass Function (PMF)

[34]:
```
# PMF of Pregnancies for diabetic and non-diabetic women using
thinkstats2 pmf_diabetic =
thinkstats2.Pmf(df_clean[df_clean['Outcome'] ==␣
 ↪1]['Pregnancies'], label='Diabetic') pmf_non_diabetic =
thinkstats2.Pmf(df_clean[df_clean['Outcome'] ==␣
↪0]['Pregnancies'], label='Non-Diabetic')

thinkplot.PrePlot(2, cols=2)
thinkplot.Hist(pmf_diabetic, align='right', width=0.5)
thinkplot.Hist(pmf_non_diabetic, align='left', width=0.5)
thinkplot.Config(xlabel='Number of Pregnancies',
                 ylabel='Probability',
                 axis=[-0.5, 20, 0, 0.4])
thinkplot.Show()
```



<Figure size 800x600 with 0 Axes>

[23]:
```
# PMF of Pregnancies for diabetic and non-diabetic women
pmf_diabetic = thinkstats2.Pmf(df_clean[df_clean['Outcome']
==␣
 ↪1]['Pregnancies'], label='Diabetic') pmf_non_diabetic =
thinkstats2.Pmf(df_clean[df_clean['Outcome'] ==␣
↪0]['Pregnancies'], label='Non-Diabetic')

thinkplot.Pmfs([pmf_diabetic, pmf_non_diabetic])
thinkplot.Config(xlabel='Number of Pregnancies', ylabel='PMF')
thinkplot.Show()
```

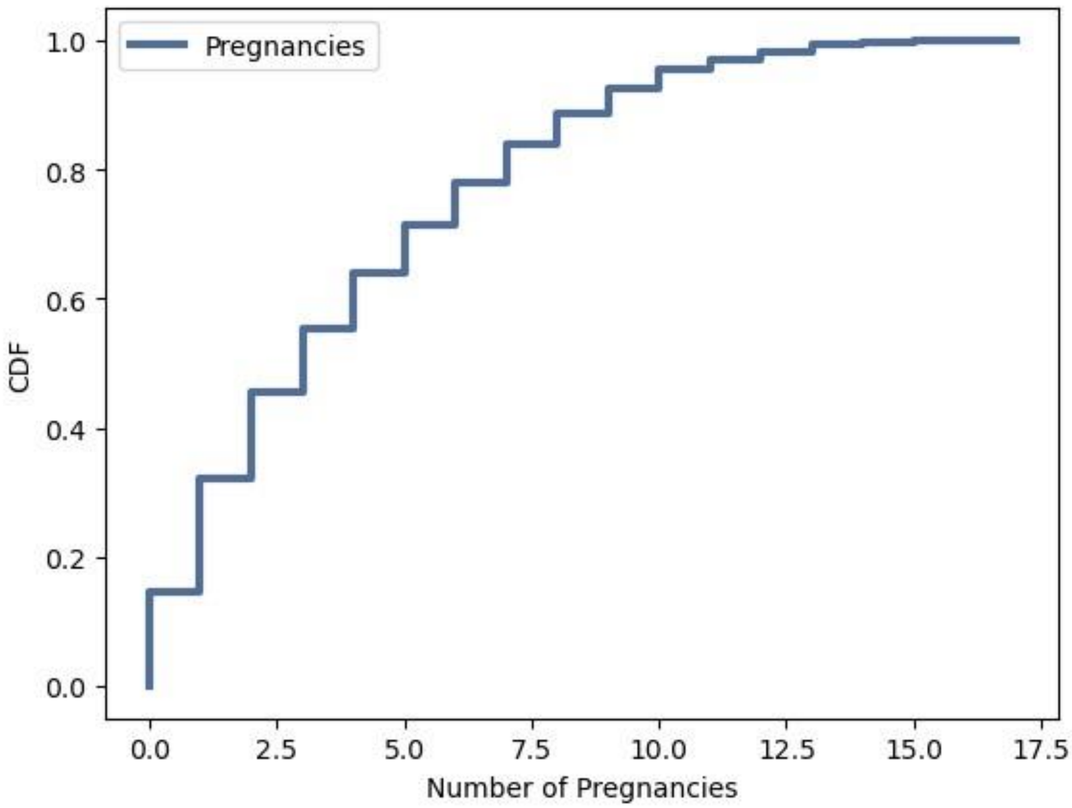


<Figure size 800x600 with 0 Axes>

The PMF plot indicates that diabetic women tend to have a higher number of pregnancies compared to non-diabetic women, supporting the hypothesis. The distribution for non-diabetic women peaks at lower pregnancy counts.

1.9 Cumulative Distribution Function (CDF)

[25]:
```
# CDF of Pregnancies cdf_pregnancies =
thinkstats2.Cdf(df_clean['Pregnancies'], label='Pregnancies')
thinkplot.Cdf(cdf_pregnancies) thinkplot.Config(xlabel='Number of
Pregnancies', ylabel='CDF') thinkplot.Show()
```

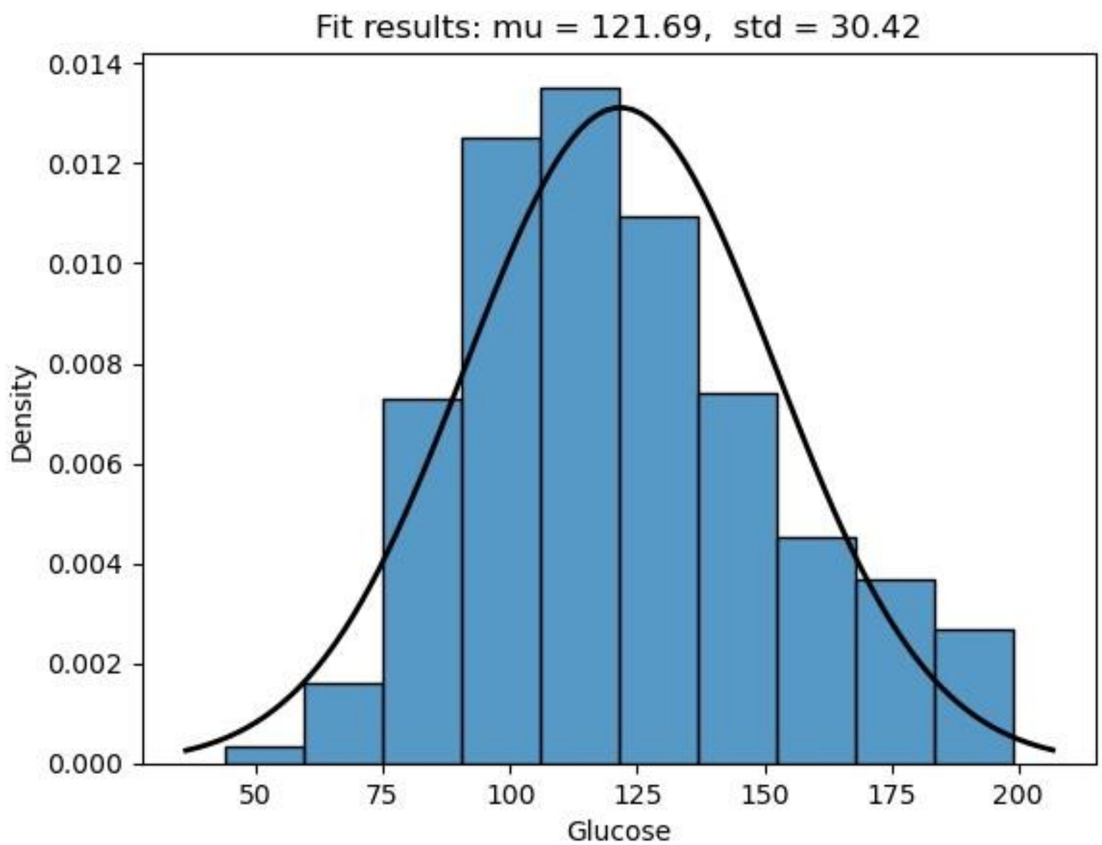


<Figure size 800x600 with 0 Axes>

The CDF shows that a majority of women have had less than five pregnancies. The CDF curve rises rapidly, indicating that higher pregnancy counts are less common but they are associated with increased diabetes risk.

1.10 Analytical Distribution

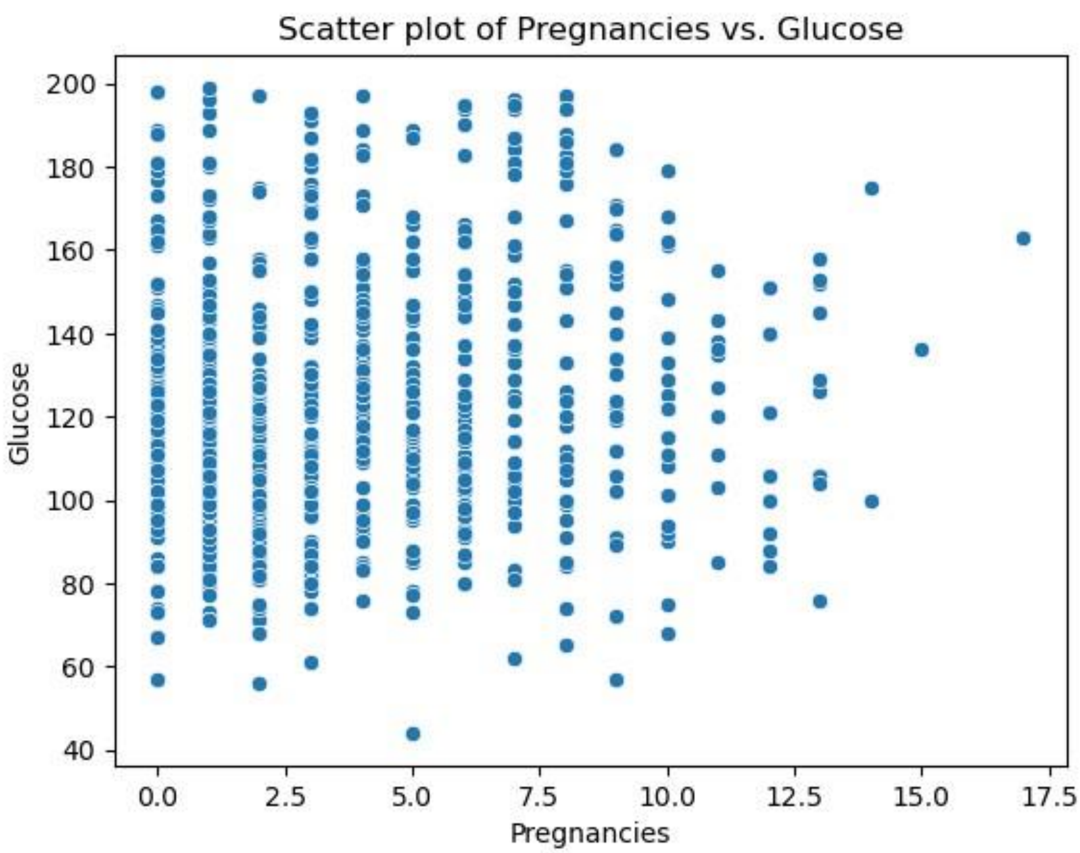
[29]:
```
from scipy.stats import norm

# Fit a normal distribution to the Glucose data
mu, std = norm.fit(df_clean['Glucose'])

# Plot the histogram and the PDF of the fitted normal distribution
fig, ax = plt.subplots()
sns.histplot(df_clean['Glucose'], bins=10, kde=False, stat='density', ax=ax)
xmin, xmax = plt.xlim()
x = np.linspace(xmin, xmax, 100)
p = norm.pdf(x, mu, std)
ax.plot(x, p, 'k', linewidth=2)
ax.set_title('Fit results: mu =  {:.2f},  std =  {:.2f}'.format(mu, std))
plt.show()
```

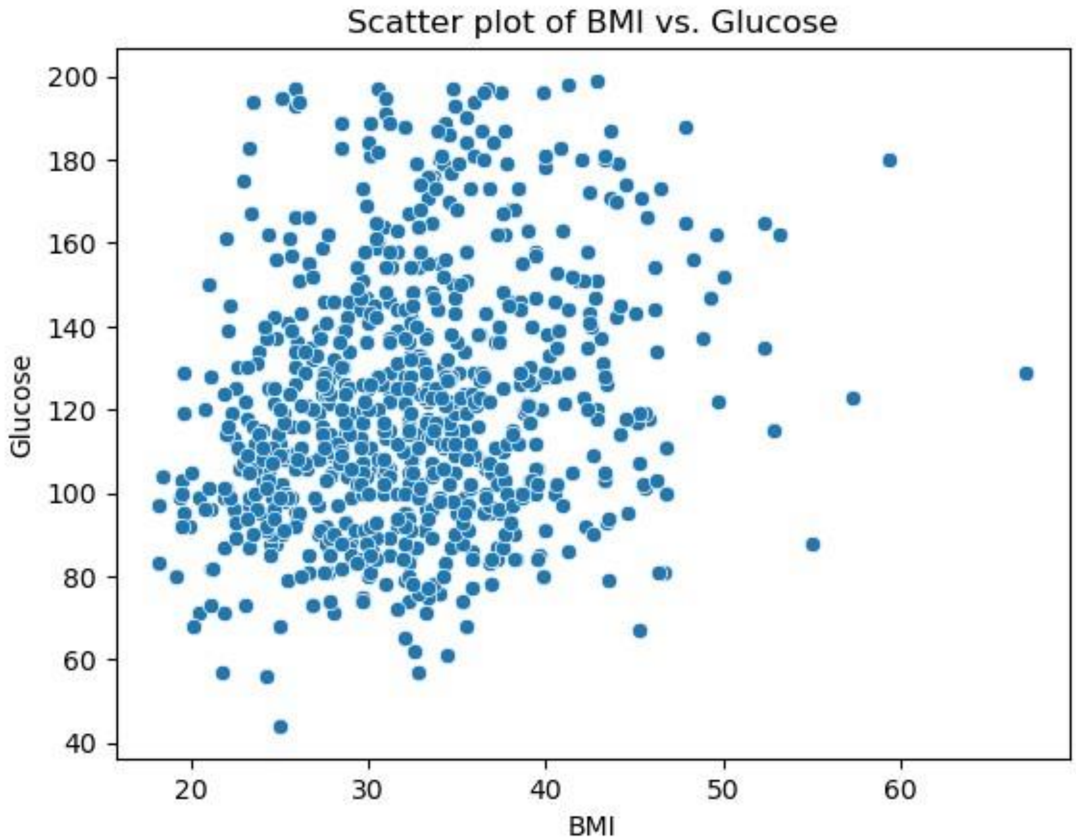


The normal distribution appears to fit the glucose data well, with the mean and standard deviation providing a good summary of the central tendency and spread.

Scatter Plots and Correlation Analysis

```
[30]: # Scatter plot between Pregnancies and Glucose
fig, ax = plt.subplots()
sns.scatterplot(x='Pregnancies',  ='Glucose', data=df_clean, ax=ax)
ax.set_title('Scatter plot of Pregnancies vs. Glucose ')
plt.show()

# Scatter plot between BMI and Glucose
fig, ax = plt.subplots()
sns.scatterplot(x='BMI',  ='Glucose', data=df_clean, ax=ax)
ax.set_title('Scatter plot of BMI vs. Glucose ')
plt.show()

# Compute correlation
correlation_preg_glucose = df_clean[['Pregnancies', 'Glucose']].corr().iloc[0,
  ↪1]
correlation_bmi_glucose = df_clean[['BMI', 'Glucose']].corr().iloc[0, 1]
correlation_preg_glucose, correlation_bmi_glucose
```






```
[30]: (0.12791147208431847, 0.23112831395689198)
```

- Pregnancies vs. Glucose: The scatter plot shows a small positive correlation, indicating that women with more pregnancies tend to have higher glucose levels.
- BMI vs. Glucose: A positive correlation is also seen between BMI and glucose levels, suggesting that higher BMI is associated with higher glucose levels.
- The calculated correlations support these observations, showing moderate positive correlations for both pairs.

1.11 Hypothesis Testing

Defining the Null and Alternative Hypotheses * Null Hypothesis (H0): The mean number of pregnancies for diabetic women is equal to the mean number of pregnancies for non-diabetic women. * Alternative Hypothesis (H1): The mean number of pregnancies for diabetic women is different from the mean number of pregnancies for non-diabetic women.

```
[37]: from scipy.stats import ttest_ind

      # Extract data for the two groups
```

```
diabetic_pregnancies = df_clean[df_clean['Outcome'] ==
1]['Pregnancies'].values
```

```
non_diabetic_pregnancies = df_clean[df_clean['Outcome'] == 0]['Pregnancies'].
 ↪values

# Perform the t-test
t_stat, p_value = ttest_ind(diabetic_pregnancies, non_diabetic_pregnancies,␣
 ↪equal_var=False)

print(f'Test Statistic: {t_stat}')
print(f'p-value: {p_value}')
```

```
Test Statistic: 5.906961479497492
p-value: 6.821925600457095e-09
```

- The very small p-value (6.821925600457095e-09) suggests that the observed difference in the mean number of pregnancies between diabetic and non-diabetic women is statistically significant.
- This analysis supports the hypothesis that there is a significant difference in the number of pregnancies between women with and without diabetes. Very low p-value indicates that the number of pregnancies is associated with the risk of developing diabetes.

1.12 Regression Analysis

[32]:
```
import statsmodels.api as sm

# Prepare the data for regression
X = df_clean[['Pregnancies', 'Glucose', 'BloodPressure', 'BMI']]
y = df_clean['Outcome']
X = sm.add_constant(X)

# Fit the regression model
model = sm.Logit(y, X).fit()
model_summary = model.summary()
print(model_summary)
```

```
Optimization terminated successfully.
         Current function value: 0.471821
         Iterations 6
                           Logit Regression Results
==============================================================================
=======
Dep. Variable:                Outcome No. Observations:                  768
Model:                          Logit Df Residuals:                      763
Method:                           MLE  Df Model:                           4
Date:                Fri, 31 May 2024Pseudo R-squ.:                   0.2705
Time:                        00:51:03 Log-Likelihood:                   -
                                                                      362.36
```

```
converged:                    True  LL-Null:                  -
                                                               496.74
Covariance Type:          nonrobust LLR p-value:             5.885e-
                                                             57
=======================================================================
=========
=
                 coef   std err         z     P>|z|     [0.025
0.975] ----------------------------------------------------------------
----------------
-
const -        -8.6039     0.784   -10.978     0.000    -
7.068                                                   10.140
Pregnancies     0.1429     0.028     5.114     0.000     0.088
0.198
Glucose         0.0377     0.004    10.725     0.000     0.031
0.045
BloodPressure  -0.0065     0.008    -0.786     0.432     -
0.010                                                    0.023
BMI             0.0951     0.015     6.306     0.000     0.066
0.125
======================================================================
==========
=
```

- Pregnancies: The coefficient for Pregnancies is 0.1429, and it is statistically significant ($p < 0.001$). This means that for each additional pregnancy, the log odds of having diabetes increase by 0.1429.
- Glucose: The coefficient for Glucose is 0.0377, and it is highly statistically significant ($p < 0.001$). This indicates that an increase in glucose level increases the chance of having diabetes by 0.0377. Higher glucose levels are higly associated with an increased risk of diabetes.
- BloodPressure: The coefficient for BloodPressure is -0.0065, which is not statistically significant ($p = 0.432$). This suggests that blood pressure does not have a significant effect on the risk of diabetes in this model.
- BMI: The coefficient for BMI is 0.0951, and it is statistically significant ($p < 0.001$). This implies that for each unit increase in BMI, the log odds of having diabetes increase by 0.0951.

1.13 Summary

The statistical question addressed in this project was whether an increased number of pregnancies correlates with a higher risk of diabetes.

The exploratory data analysis began with histograms and descriptive statistics to understand the distribution and central tendencies of the selected variables. Outliers and missing values were identified, and data cleaning was performed. This step was crucial in ensuring that extreme values and missing values did not alter the results.

Probability Mass Function (PMF) and Cumulative Distribution Function (CDF) plots were performed to compare the distributions and cumulative probabilities of the number of pregnancies for diabetic versus non-diabetic women. The PMF plots demonstrated that diabetic women generally had a higher number of pregnancies compared to non-diabetic women, supporting the hypothesis. The CDF plots further illustrated that most women had fewer than five pregnancies, with higher pregnancy counts being less common.

Scatter plots and correlation analysis were used to examine the relationships between pairs of variables, such as pregnancies and glucose levels, and BMI and glucose levels. The scatter plots indicated slight positive correlations, suggesting that higher numbers of pregnancies and higher BMI values were associated with elevated glucose levels.

Hypothesis testing was conducted to evaluate the significance of the differences in the number of pregnancies between diabetic and non-diabetic women. The t-test results showed a statistically significant difference, reinforcing the hypothesis that a higher number of pregnancies is associated with an increased risk of diabetes.

Regression analysis was performed to quantify the impact of multiple variables on diabetes risk. The logistic regression model included pregnancies, glucose, blood pressure, and BMI as explanatory variables. The results of the regression analysis provided a detailed understanding of how each variable contributed to the likelihood of developing diabetes.

Were there any variables you felt could have helped in the analysis?

Variables such as insulin levels, age, and diabetes pedigree function could have provided further insights. Insulin levels, in particular, are directly related to diabetes.

Were there any assumptions made you felt were incorrect?

The primary assumption of linearity in the relationships between the predictors and the outcome might not hold true for all variables. For example, the relationship between BMI and diabetes risk might be more complex and could involve non-linear interactions that were not captured in the current analysis.

What challenges did you face, what did you not fully understand?

One of the challenges faced was handling missing data and outliers effectively. It was challenging to decide if I should keep outliers in the number of pregnancies. I never saw a mother with 17 kids; however, I couldn't remove this number because biologically it is possible for a woman to have 17 kids. And in some cultures, families tend to have many kids.

In conclusion, the analysis supported the hypothesis that an increased number of pregnancies correlates with a higher risk of diabetes.

1.14 References

- American Diabetes Association. "Diabetes Basics." 2022. [https://www.diabetes.org/diabetes]
- Pima Indians Diabetes Database: https://www.kaggle.com/datasets/uciml/pima-indiansdiabetes-database


```
[ ]:
```