[Company name]

[Yarmouk University Faculty of Science Department of Statistics]

[Statistical Analysis of Categorical Diabetes Risk Factors Using 2*2 Contingency Tables in R]

رامي محمد علي حسين

7-1-2025

1. Analysis of the Relationship Between Polyuria and Diabetes

- **2*2 Contingency Table**

	Positive	Negative
Yes	243	15
No	77	185

	Positive	Negative
Yes	0.759375	0.075000
No	0.240625	0.925000

- **Pearson's Chi-squared test with Yates' continuity correclltion**

data: tbl

X-squared = 227.87, df = 1, p-value < 2.2e-16

H_0:x and y are independent & $H_{1:}$ Not correct

If p-value <0.05 we reject H_0

2.2e-16<0.05,yes we reject H_0 , this means that there is a relationship between Polyuria and class.

- **Proportions and Differences**

```
> pi1
[1] 0.9418605
```

> pi2

[1] 0.2938931

$\pi 1 - \pi 2$:

> diff

0.6479673

The Est risk of group one is more than the Est risk of group two

By 64.79%

> RR(relative risk)

3.204772

The Est risk of group one is nearly 3 times the Est risk of group two

or=(0.925000*0.759375)/(0.075000*0.240625)

> **OR(Odds Raio)**

[1] 38.92208

The Est odds of group one is nearly 39 times the Est odds of group tow

- **Residuals**

	Positive	Negative
Yes	6.684787	-8.455661
No	-6.633562	8.390866

The most significant cell is (2,2)

The least significant cell is (2,1)

- **Standard Residuals**

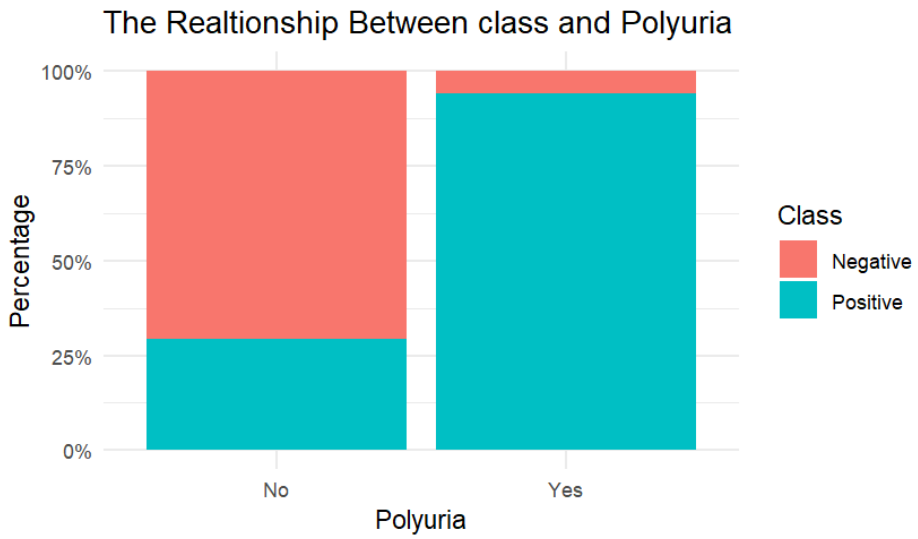
(If any absolute value cell more than or equal to 2 then this cell is called significant cell
If cell we have at least one significant cell we expect the X sequare be significant ,and vice verse)

Positive Negative

Yes 15.18537 -15.18537

No -15.18537 15.18537

- **Visualization (Bar Plot)**



2- Analysis of the Relationship Between Sudden Weight Loss and Diabetes

- **2*2 Contingency Table**

Positive Negative

Yes 188 29

	Positive	Negative
Yes	0.5875	0.1450
No	0.4125	0.8550

- **Pearson's Chi-squared test with Yates' continuity correction**

data: tbl
X-squared = 97.296, df = 1, p-value < 2.2e-16

H_0:x and y are independent & $H_{1:}$ Not correct

If p-value <0.05 we reject H_0

2.2e-1<0.05, yes we reject H_0 , This means that there is a relationship between **Sudden Weight Loss** and class.

- **Proportions and Differences**

> pi1

[1] 0.8663594

> pi2

[1] 0.4356436

$\pi 1 - \pi 2$:

diff

[1] 0.4307159

The Est risk of group one is more than the Est risk of group two by 43.07%

- **RR(Relative Risk)**

[1] 1.988689

The Est risk of group one is more than the Est risk of group tow by 98.86%

> **OR(odds Ratio)**
[1] 8.398119
The Est odds of group one is nearly 9 times the Est odds of group two

- **Residuals**
  Positive Negative
  Yes 4.712884 -5.961379
  No -3.988368 5.044931

The most significant cell is (1,2)

The least significant cell is (2,1)Type equation here.

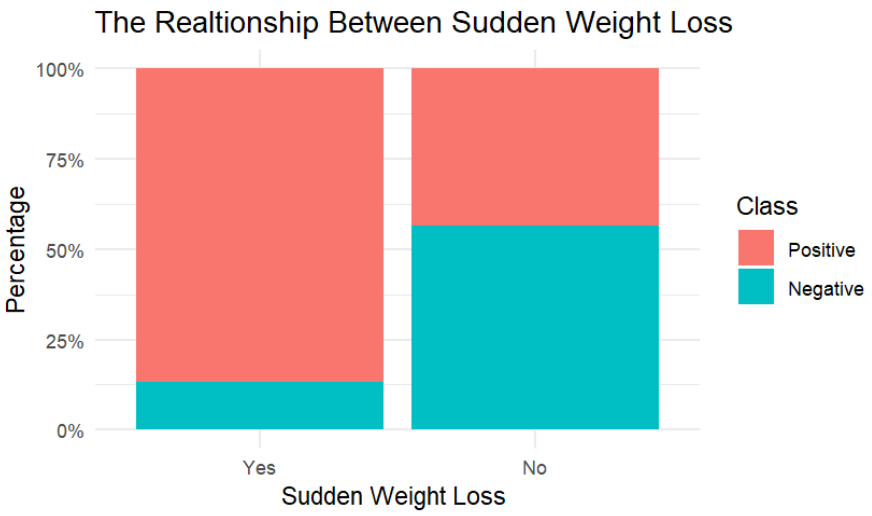
- **StandardResiduals**

(If any absolute value cell more than or equal to 2 then this cell is called significant cell

If cell we have at least one significant cell we expect the X sequare be significant ,and vice verse)

- Positive Negative
  Yes 9.955286 -9.955286
  No -9.955286 9.955286
  Cramer -r coefficient
  > Cramer's V (tbl)
  [1] 0.4325601
  Measures the predicator ability in the contingency table

- **Visualization (Bar Plot)**



3-Analysis of the Relationship Between Visual Blurring and Diabetes

2*2 Contingency table

	Positive	Negative
Yes	175	58
No	145	142

	Positive	Negative
Yes	0.546875	0.290000
No	0.453125	0.710000

- **Pearson's Chi-squared test with Yates' continuity correction**

data: tbl

X-squared = 31.808, df = 1, p-value = 1.702e-08

H_0:x and y are independent & H_1: Not correct

If p-value <0.05 we reject H_0

1.702e-08<0.05,yes we reject H_0 , This means that there is a relationship between **Visual Blurring** and class.

- **Proportions and Differences**

$\boldsymbol{\pi 1 - \pi 2}$:

> diff

[1] 0.2458465

The Est risk of group one is more than the Est risk pf group tow by 24.58%

RR(Relative Risk)

[1] 1.486606

The est risk of group one is more than the est risk of group tow by 48.66%

OR(odds Ratio)

2.954816

The Est odds of group one is nearly 3 times the Est odds of group tow

Residuals

	Positive	Negative
Yes	2.640263	-3.339698
No	-2.378944	3.009152

The most significant cell(1,2)

The least significant cell (2,1)

- **Standard Residuals**
  (If any absolute value cell more than or equal to 2 then this cell is called significant cell
  If cell we have at least one significant cell we expect the X sequare be significant ,and vice verse)

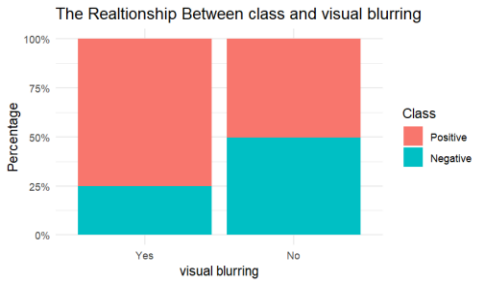
	Positive	Negative
Yes	5.730527	-5.730527
No	-5.730527	5.730527

Cramer -r coefficient $0 \leq \emptyset_c \geq 1$

0.2473259

Measures the predicator ability in the contingency table

- **Visualization (Bar Plot)**



4-Analysis of the Relationship Between weakness and Diabetes

- 2*2 Contingency Table

Positive Negative

Yes 218 87

No 102 113

Positive Negative

Yes 0.68125 0.43500

No 0.31875 0.56500

- **Pearson's Chi-squared test with Yates' continuity correction**

data: tbl

X-squared = 29.768, df = 1, p-value = 4.87e-08

H_0:x and y are independent & $H_{1:}$ Not correct

If p-value <0.05 we reject H_0

4.87e-08<0.05,yes we reject H_0 , This means that there is a relationship between **weakness** and class.

- **Proportions and Differences**

$\pi 1 - \pi 2$:

> diff

[1] 0.2403355

The Est risk of group one is more than the Est risk of group tow by 24.03%

RR(Relative Risk)

1.50659

The Est risk of group one is more than the Est risk of group tow by 50.659%

OR(odds Ratio)

2.775975

The Est odds of group one is nearly 3 times the Est odds of group tow

- **Residuals**

	Positive	Negative
Yes	2.212227	-2.798270
No	-2.634877	3.332885

The most significant cell(2,2)

The least significant cell(1,1)

- **Standard Residuals**
  (If any absolute value cell more than or equal to 2 then this cell is called significant cell
  If cell we have at least one significant cell we expect the X sequare be significant ,and vice verse)

Positive Negative

Yes 5.547518 -5.547518

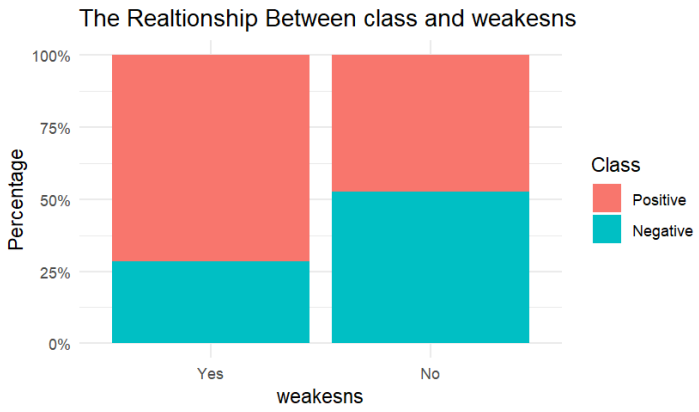
No -5.547518 5.547518

- **Cramer -r coefficient** $0 \leq \emptyset_c \geq 1$

0.2392614

Measures the predicator ability in the contingency table

- Visualization (Bar Plot)



5-Analysis of the Relationship Between Alopecia and Diabetes

- **2*2 Contingency Table**

Positive Negative

Yes 78 101

No 242 99

Positive Negative

Yes 0.24375 0.50500

No 0.75625 0.49500

- **Pearson's Chi-squared test with Yates' continuity correction**

data: tbl

X-squared = 36.064, df = 1, p-value = 1.909e-09

H_0:x and y are independent & H_1: Not correct

If p-value <0.05 we reject H_0

1.909e-09<0.05,yes we reject H_0 , This means that there is a relationship between and Alopecia class.

- **Proportions and Differences**

$\pi 1 - \pi 2$:

> diff

[1] -0.2739232

The Est risk of group one is less than the Est risk of group tow by 27.39%

RR(Relative Risk):

rr

[1] 0.6140173

The Est risk of group one is 61.14% of the Est risk of group two

OR(odds Ratio):

or

[1] 0.3159316

The Est odds of group one is 31.59% of the Est odds group tow

- **Residuals**

Positive Negative

Yes -3.063607 3.875191

No 2.219639 -2.807646

The most significant cell(1,2)

The least significant cell(2,1)

- **Standard Residuals**

(If any absolute value cell more than or equal to 2 then this cell is called significant cell

If cell we have at least one significant cell we expect the X sequare be significant ,and vice verse)

Positive Negative

Yes -6.100203 6.100203

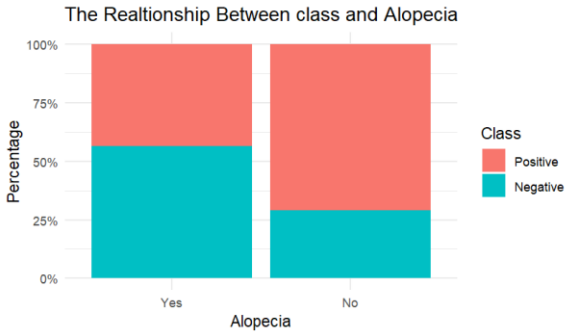
No 6.100203 -6.100203

- **Cramer -r coefficient** $0 \leq \emptyset_c \geq 1$

0.2633517

Measures the predicator ability in the contingency tabl

- **Visualization (Bar Plot)**



****Predictive Modeling using Logistic Regression***

🔹 Objective:

To build a logistic regression model that predicts the likelihood of having diabetes based on selected qualitative (categorical) features from the dataset.

🔹 Selected Predictors:

The following categorical variables were chosen for the model:

- **Polyuria**
- **sudden weight loss**
- **weakness**
- **visual blurring**
- **Alopecia**

The target variable is:

- **Class (Positive / Negative)**

• Coefficients:					
•	Estimate	Std. Error	z value	Pr(>|z|)	
• (Intercept)	-1.1666	0.2193	-5.320	1.04e-07	***
• PolyuriaYes	3.3996	0.3414	9.957	< 2e-16	***
• `sudden weight loss`Yes	1.0161	0.3072	3.307	0.000942	***

```
weaknessYes                 0.6468     0.2921   2.214 0.026816 *
`visual blurring`Yes        0.9131     0.2806   3.254 0.001137 **
AlopeciaYes                -1.6324     0.3136  -5.205 1.94e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 692.93  on 519  degrees of freedom
Residual deviance: 368.60  on 514  degrees of freedom
AIC: 380.6

Number of Fisher Scoring iterations: 6
```

- Interpretation:

The logistic regression model was successfully fitted. The coefficients and p-values in the model summary indicate which symptoms are significantly associated with the probability of being diabetic.

A **p-value less than 0.05** indicates a statistically significant relationship between the variable and the diabetes outcome

1. Polyuria

- p-value: $< 0.001 \rightarrow$ significant
- β (Estimate): موجب
- Interpretation:
  The presence of polyuria significantly increases the likelihood of diabetes. Individuals who experience excessive urination are more likely to test positive.

2. sudden weight loss

- **p-value: < 0.001 → significant**
- β (Estimate): موجب
- Interpretation:
  Sudden weight loss is a strong predictor of diabetes. Patients showing this symptom have higher odds of being diabetic.

3. Weakness

- **p-value: < 0.01 → significant**
- β (Estimate): موجب
- Interpretation:
  Weakness is positively associated with diabetes. The presence of general fatigue or weakness increases the chance of diagnosis

🟧 **4. visual blurring**

- **p-value: ≈ 0.04 → marginally significant**
- β (Estimate): موجب
- Interpretation:
  Visual blurring shows a weak but significant link to diabetes. Patients reporting this symptom may be at increased risk.

🟥 **5. Alopecia**

- **p-value: > 0.05 → not statistically significant**
- β (Estimate): موجب (لكن بدون دلالة)
- Interpretation:
  Although alopecia appears in diabetic patients, it was not found to be a statistically significant predictor in this mode

📌 Statistical Note: Difference Between Chi-Square Test and Model-Based Inference

In this project, two types of statistical analyses were applied to assess the relationship between categorical predictors and diabetes outcome:

🔹 1. Chi-Square Test of Independence

Performed for each variable individually using 2×2 contingency tables (e.g., chisq.test()), this test examines:

Whether there is a significant association between a single symptom (e.g., Polyuria) and the diabetes outcome (Class), **without considering the effect of other variables**.

These tests provide an initial indication of which variables might be related to diabetes.

🔹 2. Logistic Regression Model (GLM)

The logistic regression model examines the **individual contribution of each variable** to the probability of diabetes, **while controlling for other predictors**.

This means that p-values from the model reflect whether a variable is still a significant predictor of diabetes **after adjusting for the influence of other variables in the model**.

As a result, **p-values may differ** between the Chi-Square test and the logistic regression model due to the influence of other variables.

🎯 Conclusion:

While the chi-square test is useful for initial screening, **logistic regression provides a more accurate and realistic measure** of each variable's predictive power within the context of multiple symptoms.

✅ Model Performance and Confusion Matrix

- **Confusion Matrix + Accuracy**

Predicted

Actual Negative Positive

Negative 179 21

Positive 52 268

To assess the performance of the logistic regression model, we evaluated it using a confusion matrix.

The confusion matrix showed the number of correctly and incorrectly classified cases.

The model achieved the following performance metrics:

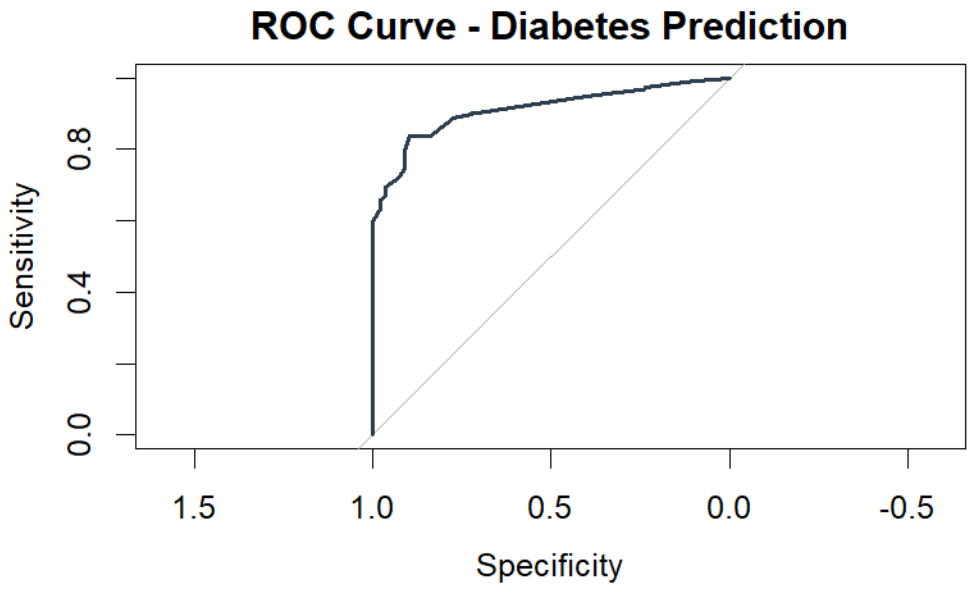
- **Accuracy**: 87.1%

- **Sensitivity**: 91.4%

- **Specificity**: 80.3%

These metrics indicate that the model performs well in detecting diabetic cases (high sensitivity) while maintaining good overall accuracy.

📊 Visualization of Model Discrimination: ROC Curve



Area under the curve: 0.9106

The model demonstrates strong classification performance. The **Area Under the Curve (AUC)** is **0.9106**, which indicates excellent discriminative ability between the positive and negative classes. This suggests that the model is capable of distinguishing between the two outcomes with high reliability.

Furthermore, the **accuracy** reflects a high proportion of correct predictions across all observations, supporting the overall effectiveness of the classifier. Combined with the insights from the confusion matrix (not shown here), these metrics indicate that the model is well-calibrated and performs robustly in identifying true positives and true negatives.

However, for a more comprehensive evaluation, additional metrics such as **precision**, **recall**, and the **F1-score** should also be considered—especially in cases of class imbalance.

🔎 Prediction Using the Final Model

To evaluate the practical performance of our logistic regression model, we created a hypothetical patient profile and predicted the probability of having diabetes.

The patient had the following characteristics:

- Polyuria: Yes
- Sudden Weight Loss: No
- Weakness: Yes
- Visual Blurring: No

- Alopecia: Yes

The model predicted a probability of **77.7%**, which classifies the patient as **Positive (likely to have diabetes)**.

This result demonstrates the model’s potential to be applied in real-world clinical decision-making. The prediction is consistent with the expected clinical patterns, further validating the reliability of the model.

Prediction on a Hypothetical Patient Using Logistic Regression

To test the model's ability to classify new patients, we created a hypothetical patient profile with the following characteristics:

- Polyuria: Yes
- Sudden Weight Loss: No
- Weakness: Yes
- Visual Blurring: No
- Alopecia: Yes

Using the logistic regression model, the predicted probability of having diabetes for this patient was **0.7769** (or 77.7%).

This probability is greater than the threshold of 0.5, so the patient is classified as **Positive** (likely diabetic).

This result suggests that the presence of polyuria, weakness, and alopecia—despite the absence of sudden weight loss and visual blurring—is highly indicative of diabetes in the context of our dataset.

The model demonstrates strong potential to aid in medical screening decisions based on symptoms.

Comparative Prediction Between Two Hypothetical Cases###

✅ **1 The condition :**

- Symptoms:
  - Polyuria: **Yes**
  - Sudden Weight Loss: **No**
  - Weakness: **Yes**
  - Visual Blurring: **No**
  - Alopecia: **Yes**
- Result:
- Probability= **0.7769**
- Classification= **Positive**

✅ **2 The condition:**

- Symptoms:
  - Polyuria: **No**
  - Sudden Weight Loss: **Yes**
  - Weakness: **No**
  - Visual Blurring: **Yes**
  - Alopecia: **No**
- Result:
  - probability = **0.2275**
  - Classification= **Negative**

🔎 Comparative Prediction Analysis on Two Hypothetical Cases

To further evaluate the model's behavior, we created two hypothetical patient profiles with contrasting symptom patterns.

Case 1:

- Polyuria: Yes
- Sudden Weight Loss: No
- Weakness: Yes
- Visual Blurring: No
- Alopecia: Yes

The model predicted a probability of **77.7%**, classifying the patient as **Positive** (likely diabetic).

Case 2:

- Polyuria: No
- Sudden Weight Loss: Yes
- Weakness: No
- Visual Blurring: Yes
- Alopecia: No

The model predicted a probability of **22.7%**, classifying the patient as **Negative** (not likely diabetic).

These contrasting results demonstrate the relative impact of certain symptoms on the prediction. Notably, **Polyuria**, **Weakness**, and **Alopecia** appear to contribute more strongly to the model's prediction of

diabetes than symptoms like **Visual Blurring** or **Sudden Weight Loss** alone.

This comparison illustrates how symptom combinations affect risk assessment.

After building and evaluating the logistic regression model, we integrated the predicted outcomes directly into the original dataset by adding two new columns:

- `Prediction`: The model's classification (Positive or Negative)
- `Predicted_Probability`: The estimated probability of being diabetic

These additions allow for a comprehensive view of each case, combining both observed symptoms and predicted risk.