MODULE 1 – CATEGORICAL DATA

Prof. Sarala Mary

Positive Correlation, Negative Correlation and No Correlation

- Refer Notes

Exploring two or more variables

- Contingency tables

  A tally of counts between two or more categorical variables.

- Hexagonal Binning

  A plot of two numeric variables with the records binned into hexagons.

Contingency Table

- A useful way to summarize two categorical variables is a contingency table

Gender	Smoker	Non-Smoker	Total
Male	72	44	116
Female	34	53	87
Total	106	97	203

Example: A table showing total number of smoker and non – smoker in an organization

Contingency Table

Relative Frequency Contingency Table

$$\text{Percentage value for cell X} = \frac{\text{Count value in cell X}}{\text{Total Number Surveyed}} \times 100$$

Cell 1: $(72/203) \times 100 = 35.47\%$
Cell 2: $(44/203) \times 100 = 21.67\%$
Cell 3: $(34/203) \times 100 = 16.75\%$
Cell 4: $(53/203) \times 100 = 26.11\%$

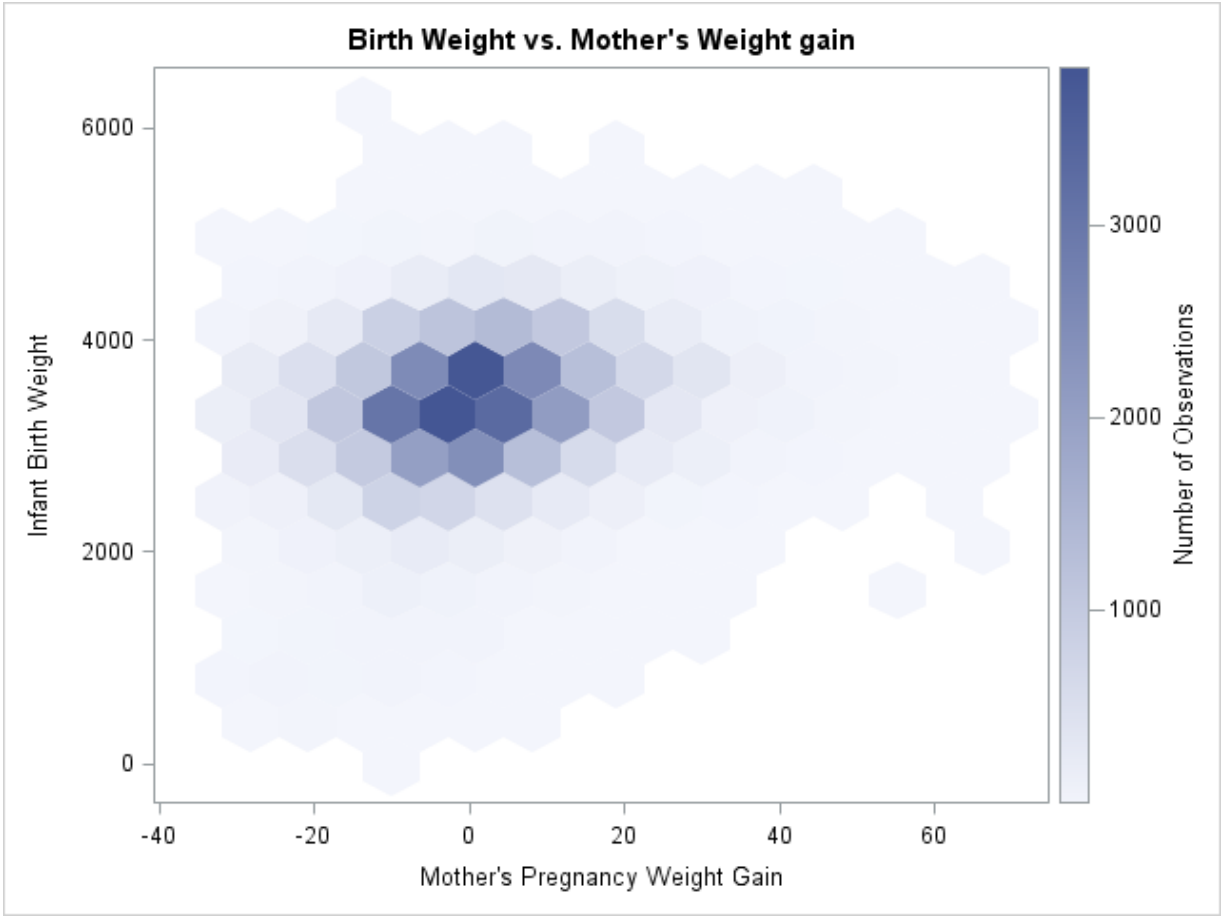
Gender	Smoker	Non-Smoker	Total
Male	72	44	116
Female	34	53	87
Total	106	97	203

Gender	Smoker	Non-Smoker	Total
Male	35.47%	21.67%	57.14%
Female	16.75%	26.11%	42.86%
Total	52.22%	47.78%	100%

Hexagonal Binning

- For data sets with hundreds of thousands or millions of records, a scatterplot will be too dense, so we need a different way to visualize the data.

Example:


Birth Weight vs. Mother's Weight gain
Infant Birth Weight
6000
4000
2000
0
-40
-20
0
20
40
60
Mother's Pregnancy Weight Gain
3000
2000
1000
Number of Observations