Statistics Unit

Andy Yan

September 2023

1 Data Analysis Grouped Data

Raw Data:

35.6, 39.3, 39.8, 40.8, 43.9, 45.7, 45.9, 47.5, 48.6, 49.2, 52.6, 55.4, 56.4, 57.4, 58.1, 58.8, 60.0, 62.2, 63.7, 64.2, 64.5, 64.9, 66.9, 68.3, 68.8, 70.1, 70.7, 73.3

Basic Terms:
Raw Data: The unprocessed information collected for a study
Continuous Variable: Can have any value within the range (Ex: Volume, Weight)
Discrete Variable: Can have only separate values, mostly integers (# of people)

Grouped Data
Grouped Data is organized with intervals and the frequency within the intervals.

Cumulative Relative Frequency Table				
Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
35.6 - 41.1	4	4	0.1429	0.1429
42.1 - 47.6	4	8	0.1429	0.2857
48.6 - 54.1	3	11	0.1070	0.3929
55.1 - 60.6	6	17	0.2143	0.6071
61.6 - 67.1	6	23	0.2143	0.8217
68.1 - 73.6	5	28	0.1786	1

Number of Values **n**
# of class intervals $\mathbf{c} = \lceil 1 + 3.222\log(n) \rceil$
Interval Size $\mathbf{i} = \lceil \frac{\textbf{Max} - \textbf{Min}}{c} \rceil$

Frequency (F) : # of occurrences for a variable
Cumulative (C) : Totaling # of
Cumulative Frequency (CF) : $CF_k = F_k + CF_{k-1}$
Relative Frequency (RF) : $RF = \frac{F}{n}$
Cumulative Relative Frequency (CRF) : $CRF_k = RF_k + CRF_{k-1}$

2 Measures of Spread

Standard Deviation for Population

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = \sqrt{\frac{\sum(x_i - \mu)^2}{\sum f_i}}$$

Standard Deviation for Sample

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}$$

Z-score

$$Z_i = \frac{x_i - \mu}{\sigma}$$

3 Percentile

P_{PR}

PR is the percentage of numbers P_{PR} is bigger than
Percentile Rank Ungrouped Data
Central Tendency: Mean Median Mode

$$PR = \frac{b + \frac{1}{2}e}{n} \cdot 100\% \qquad \begin{array}{l}\text{b: how many values below}\\ \text{e: how many equal values}\end{array}$$

$$k = PR \cdot (total + 1) = \textbf{a number greater than PR\% of the data}$$

$$P_{PR} = x_{\lfloor k \rfloor} + (\lceil k \rceil - k) \cdot (x_{\lceil k \rceil} - x_{\lfloor k \rfloor})$$

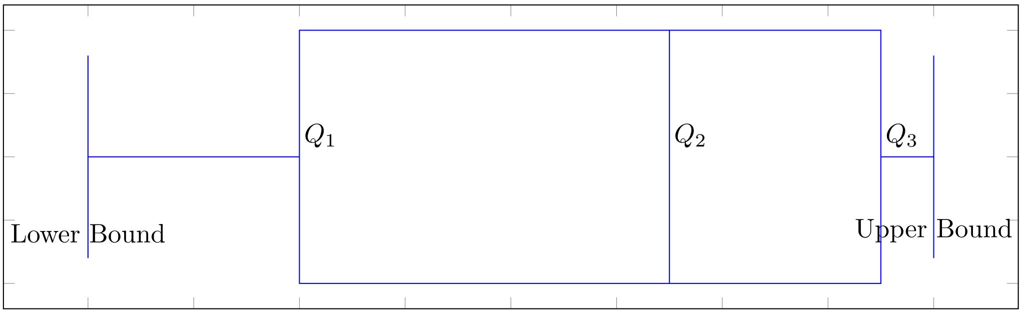
Percentile Rank Grouped Data
$\bar{x}_i$: middle value for the i^{th} interval
f_i: frequency value for the i^{th} interval
Central Tendency:

1. Mode: value of $\bar{x}_i$ in the interval with greatest frequency
2. Median $= l + (\frac{\frac{n}{2} - cf}{f}) \cdot h$
   To determine the median class, locate which class' cumulative frequency is closest to $\frac{n}{2}$.
   - l is the lower limit of the median class
   - n is the number of observations
   - f is the frequency of the median class
   - h is the class size
   - cf is the cumulative frequency of the class PRECEDING the median class
3. Mean $= \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{N} f_i \bar{x}_i}{N}$

$$P_{PR} = l + h \cdot \frac{PR \cdot n - cf}{f}$$

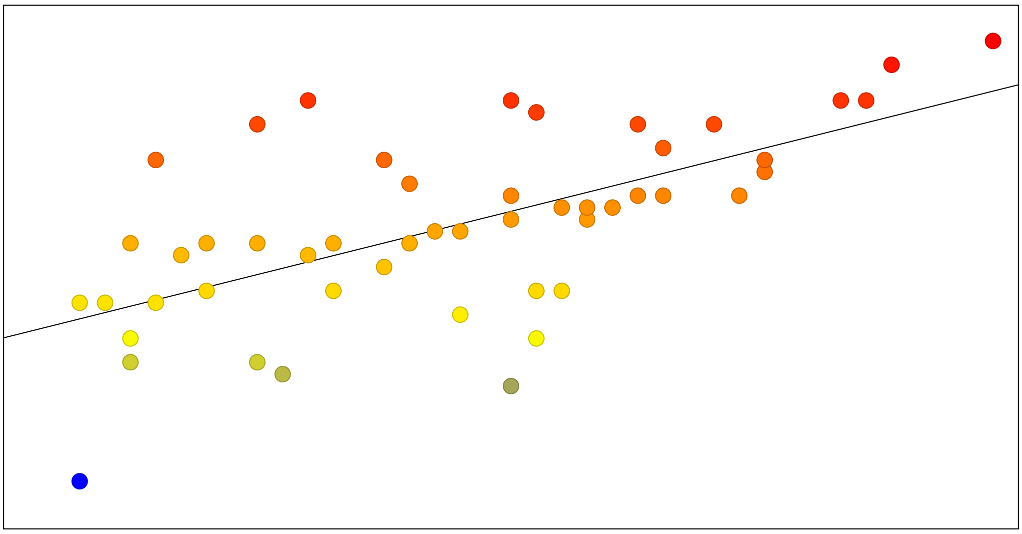
4 Box and Whisker Plot

Three Quartiles: $P_{25} = Q_1, P_{50} = Q_2, P_{75} = Q_3$



Lower Bound $= Q_1 - 1.5(Q_3 - Q_1)$
Upper Bound $= Q_3 + 1.5(Q_3 - Q_1)$
If a value isn't within these boundaries, it's classified as an outlier.

5 Two Variable Statistics



linear correlation (Pearson's R): $r = \frac{n\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{\sqrt{[(n\sum_{i=1}^{N} x_i^2) - (\sum_{i=1}^{N} x_i)^2][(n\sum_{i=1}^{N} y_i^2) - (\sum_{i=1}^{N} y_i)^2]}}$

Pearson's R value	Strength	Direction
$r > \frac{2}{3}$	Strong	Positive
$\frac{1}{3} < r < \frac{2}{3}$	Moderate	Positive
$0 < r < \frac{1}{3}$	Weak	Positive
0	None	None
$-\frac{1}{3} < r < 0$	Weak	Negative
$-\frac{2}{3} < r < \frac{1}{3}$	Moderate	Negative
$r < -\frac{2}{3}$	Strong	Negative

Line of Best Fit $y = ax + b$: $a = \frac{n\sum_{i=1}^{N} x_i y_i - (\sum_{i=1}^{N} x_i)(\sum_{i=1}^{N} y_i)}{n(\sum_{i=1}^{N} x_i^2) - (\sum_{i=1}^{N} x_i)^2}$,$b = \bar{y} - a\bar{x}$, $\bar{y} = \frac{\Sigma y}{n}$,$\bar{x} = \frac{\Sigma x}{n}$

6 Terms

- Cause-and-Effect Relationship: Change in X produces change in Y
- Common Cause Factor: External variables cause two variables to change in some way
- Accidental Relationship: Correlation exists without any causual relationship
- Presumed Relationship: Correlation seems logical with no causual relationships