Discriptive Statistics

Dr. Siddiqua Mazhar

Mid Michigan College

Sep 16, 2021

Sample Standard Deviation for a Frequency Distribution

$$s = \sqrt{\frac{\sum (x - \overline{x})^2 f}{n - 1}}$$

where

x is the midpoint of a class.

Question

The following table shows the grouped data, in classes, for the heights of 50 people.

height in cm	f	x	fx	$x - \overline{x}$	$(x - \overline{x})^2)$	$(x - \overline{x})^2).f$
120-129.	2					
130-139	5					
140-149	25					
150-159	10					
160-169	8					

Calculate the standard deviation of the height of 50 people.

Solution

Variance

- Sample Variance: S^2
- Population Variance: σ^2

Note:

- Closely grouped data will have a small standard deviation.
- Spread-out data will have a large standard deviation.

1. Given the following frequency distribution, find the following things.

Class	frequency
0-4	2
5-9	3
10-14	6
15-19	9
20-24	7
25-29	2
30-24	1

- Mean
- Median
- Mode
- Standard Deviation
- Variance

Chebyshev's Theorem

For any set of data (either population or sample) and for any constant k greater than 1, the proportion of the data that must lie within k standard deviations on either side of the mean is at least $1 - \frac{1}{k^2}$.

Explanation

Results of Chebyshev's Theorem

For any set of data:

- at least 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.
- at least 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.
- at least 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

Applications

Idea behind Within k standard deviations of the mean
Example: $\mu = 10$ and $\sigma = 3$

- Within one standrd deviation means between 7 and 13
- Within two standrd deviation means between 4 and 16
- Within three standrd deviation means between 1 and 19

Chebyshev's Theorem: Example 1

- Average score on a test is 80 points with standard deviation of 5 points. What percentage of people must have scores between 70 and 90?

- Average score on a test is 80 points with standard deviation of 5 points. What percentage of people must have scores between 55 and 105?

Empirical Rule

Empirical Rule: If the distribution is normal (bell shaped), then

- Approx 68% of data falls within one standard deviation.
- Approx 95% of data falls within 2 standard deviation.
- Approx 99.7 of data falls within 3 standard deviation.

Empirical Rule: Example 1

- Assume we have a Normal Distribution. Average score on a test is 80 points with standard deviation of 5 points.
- What percentage of people must have scores between 70 and 90?

Empirical Rule: Example 2

- Assume we have a Normal Distribution. Average score on a test is 80 points with standard deviation of 5 points.
- What percentage of people must have scores between 75 and 85?

Empirical Rule: Example 3

- Assume we have a Normal Distribution. Average score on a test is 80 points with standard deviation of 5 points.
- What percentage of people must have scores between 80 and 95?

Empirical Rule: Example 4

- Assume we have a Normal Distribution. Average score on a test is 80 points with standard deviation of 5 points.
- What percentage of people must have scores below 70?

Comparing measures between or within data sets.

Z-Score is the number of standard deviation a data value (x) is away from the mean.

Z-Score for a Sample: $\frac{X-\overline{X}}{s}$

Z-Score for Population: $\frac{X-\mu}{\sigma}$

Note: Allows comparison of the variation in two different samples/populations.

Example

Lyndon B. Johnson Height $= 76''$
Mean for Pres. height of USA$= 71.5''$
Std.Deviation $=2.1''$

z-score$= \frac{76-71.5}{2.1} = \frac{4.5}{2.1} = 2.14$

Example

Shaq's Height $= 86''$
Mean for HEAT$= 80.0''$
Std.Deviation $=3.3''$

z-score$= \frac{86-80.0}{3.3} = \frac{6}{3.3} = 1.818$
Who is relatively taller? (Use Z- score for populations)

Usual and Unsual Z-Score

- A Z-Score between -2 aqnd 2 is considered to be 'Usual'.
- A Z-Score less than -2 and greater than 2 is considered as 'Unusual'.

Note: The larger the Z- Score in terms of absolute value, the rarer the data value is.

Percentile

For whole numbers P (where $1 \leq \mathrm{P} \leq 99$), the Pth **percentile** of a distribution is a value such that P % of the data fall at or below it and (100 - P)% of the data fall at or above it.

Note: It breaks the data into 100 parts.
In short,

$$Percentile\ of\ X = \frac{No.\ of\ Values\ less\ than\ X}{Total\ No.\ of\ Values} \times 100$$

Example

You scored 87 out of 100 in a Test exam. 39 people score lower than you. There are total 54 people in the class. What will be your percentile score?

Percentile of 87$=\frac{39}{54} \times 100 = 72.$

Quartiles

- Key percentiles: 25%, 50 %, 75 %.
- First Quartile: (Q_1) Bottom 25% of sorted data.
- 2nd Quartile: (Q_2)Bottom 50 % of sorted data i.e. Median.
- Third Quartile: (Q_3)Bottom 75% of sorted data.

Note: We don't need Q_4 because it is everything in the data.

The inter quartile range (IQR) is the middle half of the data. It is teh difference between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

Outlier

Any data value less than $Q_1 - 1.5(IQR)$ and greater than $Q_3 + 1.5(IQR)$ is the outlier.

Quartiles for Even number of sample space

Find the quartiles for the following data:
1,6,3,15,10,28,36,21.

Step 1: Write in order. 1,3,6,10,15,21,28,36.

Step 2: Find Median: $Q_2 = \frac{10+15}{2} = \frac{25}{2} = 12.5$

Step 3: $Q_1 = \frac{3+6}{2} = 4.5$

Step4: $Q_3 = \frac{21+28}{2} = 24.5$

Quartiles for Odd number of sample space

Find the quartiles for the following data:
1,6,3,15,10,28,36,21,39.

Five - Number Summary
Lowest value, Q_1, *median* Q_2, Q_3, highest value.

Graphical sketch of *five-number summary* of the data is called a box plot.

Box Plot: Explanation