Summarizing Numerical Data

Measures of center

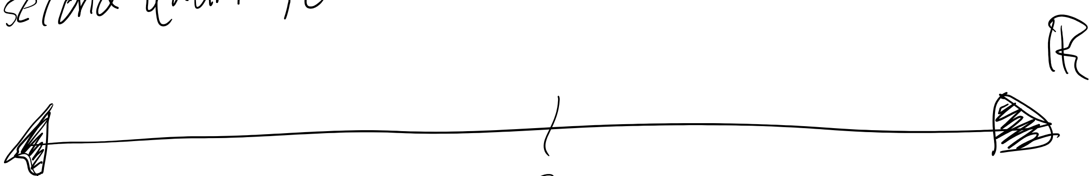
What is typical, standard, expected

Measures of spread

How much do things deviate from what is expected

5-number a summary

Given a dataset we divide it into quarters, the first quarter contain the 25% lowest values, 25-50% lowest values are in the second quarter, etc.



We define the points at the ends of these quarters as quartiles.

Q_0 (Minimum) lowest point

Q_1 divides the lowest 25% from highest 75% of our data

Q_2 (Median) divides the lowest 50% from the highest 50%

Q_3 divides lowest 75% from highest 25%

Q_4 (Maximum) highest point

Measure of centrality

Median (Q_2): The point at which 50% of observations are at or below.

Measures of spread

Range ($Q_4 - Q_0$): Range of the data from highest to lowest

Inner Quartile Range ($Q_3 - Q_1$): Range between the first and third quartiles (Range of the middle 50%)

Inner Quartile Range $(Q_3 - Q_1)$. Range between the [illegible]
(Range of the middle 50%)

Mean

The center of mass of the distribution of data.

Population: Mean is μ (mu, Greek letter)

Sample: Mean is $\overline{X}$

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Dataset: $X_1, X_2, X_3, \dots, X_n$ total of n observations

If data are binary (0 or 1) the proportion of 1s is the mean $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \left(\frac{\#\text{ones}}{n}\right)$

Why n-1 in denom?
Degrees of freedom:
If you give me all the data except one observation, and $\overline{X}$, I can calculate the last observation.

Standard Deviation or Variance

These are measures of the spread of the distribution.

Variance = (Standard Deviation)^2

(SD or Std Dev) = $\sqrt{\text{Variance}}$.

Variance is a little bit easier to calculate theoretically, Std Dev is more mathematically useful and human readable.

Standard Deviation

Population use σ (sigma, Greek letter)

Sample use s

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Root mean squared deviation

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2}$$