Measures of Variability

Recall: Measures of location
↓
"where" the data might be located. → also want to know "how" the data might be spread about this measure of location
↓
measures of variability.
↓
Note: choice of measure of variability depends on the measure of location under consideration.

(I) Standard deviation from mean

Suppose $\{x_1, x_2, \ldots, x_n\}$ sample data. → X
↓
$\bar{x}$ = sample mean = $\frac{\sum_{i=1}^{n} x_i}{n}$
↓
Want: Calculate the deviation of data points from $\bar{x}$.

Naive Approach:

Step 1: Calculate how each data point varies from $\bar{x}$. → $x_1-\bar{x}$, $x_2-\bar{x}$, $x_3-\bar{x}$, ⋮, $x_n-\bar{x}$

Step 2) Add all these individual variations.

ie: total variation $= (x_1-\bar{x}) + (x_2-\bar{x}) + (x_3-\bar{x}) + \cdots + (x_n-\bar{x})$.

Problems:

① Want the measure of spread to positive.

② positive $(x_i-\bar{x})$ will cancel contribution of negative $(x_i-\bar{x})$.
↓
can give the impression of small variation if x_i's are symmetrically distributed about $\bar{x}$.

Solution:

Step 2 corrected) Add the squares of individual variations.

ie.

$$\text{Variation} = (x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + \cdots + (x_n-\bar{x})^2$$

Still not a good measure!

If $|X|$ = size of data set is large → can lead to larger value for Variation(X).

Solution: Divide by the size of the data set.

$$\text{Sample Variance} = s^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Note ① $(n-1)$ instead of n
↓
to make s^2 unbiased estimator for population variance (denoted by σ^2)

② units of s^2 are squared of units of x_i's.
↓ take square root

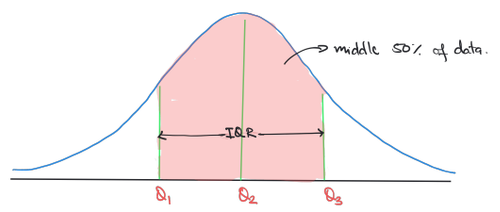
$$\text{Sample std deviation} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$$

③ s^2 is a good measure of variability if data is
a) symmetric
b) unimodal
c) does not have outliers

(II) Interquartile range for Median

Recall: Q1 → 25th percentile
Q2 → 50th "
Q3 → 75th percentile.

Interquartile range := Q3 − Q1 = the range of the middle 50% of data.



Note: Like the median, IQR is robust to minor changes in data and to outliers.

(III) Summary Statistics

Location
① Sample mean
② Sample median
③ Quartiles: Q1, Q2, Q3

Variation/Spread
① sample variance / std dev
② IQR
③ range = max(X) − min(X)

(IV) Outliers

Define:

upper fence = largest data value less than Q3 + 1.5 × IQR.

lower fence = smallest data value larger than Q1 − 1.5 × IQR.

An outlier → any data value outside of the fences.
↓
ie larger than upper fence or smaller than lower fence.

(V) Box plots

↓
pictorial representation of summary statistics.