(*) General amalgamation from Sabano's notes for this week. Enjoy!

Measures of Dispersion →

△ Dispersion: It quantifies how far the data points are from their measure of central tendency.

There are 3 main measures of dispersion →

1. Range
2. Mean Absolute deviation (MAD)
3. Standard deviation.

Let $X = \{x_1, \ldots, x_n\}$ be an ordered dataset.

Here,

1. Range $= x_n - x_1$
2. $MAD_{\mu} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \mu|$ (measure of CT used)
3. $\sigma_{\mu} = \left[\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n}\right]^{1/2}$ (measure of CT used)

If we have the central tendency used for MAD as the mean, it is minimized.

$$MAD_{\bar{x}} = \frac{1}{n}\sum_{i} |x_i - \bar{x}|$$

$$= \frac{1}{n}\sum_{x_i < \bar{x}} |x_i - \bar{x}| + \frac{1}{n}\sum_{x_i \geq \bar{x}} |x_i - \bar{x}|$$

$$= \frac{1}{n}\sum_{x_i < \bar{x}} (\bar{x} - x_i) + \frac{1}{n}\sum_{x_i \geq \bar{x}} (x_i - \bar{x})$$

We use the fact that, $\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

$$\Rightarrow \sum_{x_i < \bar{x}} (x_i - \bar{x}) = -\sum_{x_i \geq \bar{x}} (x_i - \bar{x})$$

Putting this in,

$$MAD_{\bar{x}} = \frac{2}{n}\sum_{x_i < \bar{x}} (\bar{x} - x_i) \quad \underline{OR} \quad MAD_{\bar{x}} = \frac{2}{n}\sum_{x_i \geq \bar{x}} (x_i - \bar{x})$$

Thus, we only have to compute for half the dataset if we know the mean.

Now, for standard deviation →

$$RMS_\mu = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right]^{1/2}$$ → Root mean square deviation about μ.

When $\mu = \bar{x}$ (AM), RMS about $\bar{x}$ is called **Standard deviation**

$$S.D = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2} = \left[\sum_{i=1}^{n}\left(\frac{x_i^2}{n}\right) - \bar{x}^2\right]^{1/2}$$

Similar to, $\langle x-\bar{x}\rangle^2 = \langle x^2\rangle - \langle x\rangle^2$

For weighed data,

$$S.D = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 f_i\right]^{1/2}, \quad n = \sum_{i=1}^{n} f_i$$

$$= \left[\frac{1}{n}\sum_{i=1}^{n}(x_i^2 f_i) - \bar{x}^2\right]^{1/2}$$

① Properties of standard deviation →

(1) $x_i = a \ \forall i \Rightarrow \bar{x} = \sum_{i=1}^{n}\frac{x_i}{n} = a \Rightarrow \sigma_{\bar{x}} = 0$

↳ Constant has 0 S.D

(2) Let $y = a + bx \Rightarrow \sigma_y = |b|\,\sigma_{\bar{x}}$ → Scaling of S.D.

(3) Let there be 2 sets of x,

$x_{11}, \dots, x_{1n_1}$ → n_1 elements, $\bar{x_1}$, $\sigma_{\bar{x}_1}$

$x_{21}, \dots, x_{2n_2}$ → n_2 elements, $\bar{x_2}$, $\sigma_{\bar{x}_2}$

$$\Rightarrow \sigma_{\bar{x}}^2 = \frac{n_1\sigma_{\bar{x}_1}^2 + n_2\sigma_{\bar{x}_2}^2}{n_1+n_2} + \frac{n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2}{n_1+n_2}$$

$\bar{x}$: Mean of both dist considered to be one.

This, in general → (Pooled Variance)

$$\sigma_{\bar{x}}^2 = \frac{\sum_{i=1}^{t} n_i\sigma_{\bar{x}_i}^2}{\sum_{i=1}^{t} n_i} + \frac{\sum_{i=1}^{t} n_i(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{t} n_i}$$

$\bar{x}$: Mean of entire set

(*) CT soon!!

Proof (1) is trivial,

Proof of (2) →

$$\sigma_{\bar{x}} = \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}$$

Changing variables,

$$\sigma_{\bar{y}} = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2} = \left[\frac{1}{n} \sum_{i=1}^{n} (6x_i - 6\bar{x})^2 \right]^{1/2}$$

$$= |6| \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^{1/2}$$

$$= |6| \, \sigma_{\bar{x}}$$

□ Try Proof of (3).

Moments →

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \quad \bar{x} = \frac{1}{n} \sum_i x_i f_i \text{ , where } n = \sum_i f_i$$

Let us consider the moments, defined as,

$$m_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - A)^n \quad , \quad m_n' = \frac{1}{n} \sum_i (x_i - A)^n f_i$$

Some measure of Central Tendency

(*) When there is grouped data, $x_i \rightarrow$ midpoint of a class.

When $A = \bar{x}$, it is called central moment.

$$\therefore \; m_n = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^n$$

$$\therefore \; m_n' = \frac{1}{n} \sum_i (x_i - \bar{x})^n f_i$$

Some values,

① $m_0 = m_0' = 1$

② $m_1 = 1$, $m_1' = \bar{x} - A$

Note, $(x_i - \bar{x})^n = \left[(x_i - A) - (\bar{x} - A) \right]^n$

Using Binomial theorem,

$$\Rightarrow (x_i - \bar{x})^n = \sum_j \binom{n}{j} (x_i - A)^{n-j} (\bar{x} - A)^j (-1)^j$$

$$\Rightarrow m_n = \frac{1}{n}\sum_i \sum_j \binom{n}{j}(x_i - A)^{n-j}(\bar{x} - A)^j (-1)^j$$

↙ DCheck!

$$= \frac{1}{n}\sum_j \binom{n}{j} m'_{n-j}\, m_1'^{\,j} (-1)^j$$

Now,

$$m_1 = m_1' - m_1' = 0$$

$$m_2 = m_2' - m_1'^2$$

$$m_3 = m_3' - 3m_2'm_1' + 3m_1'^3 - m_1'^3$$

$$= m_3' - 3m_2'm_1' + 2m_1'^3$$

$$m_4' = m_4' - 4m_3'm_1' + 6m_2'm_1'^2 - 3m_1'^4$$

Note, $m_1' = d$, $m_2' = m_2 + d$, $m_3' = m_3 + 3m_2 d + d^3$

$$m_4' = m_4 + 4m_3 d + 6m_2 d^2 + d^4$$

Sheppard's Correction → (for general moments)

c : Length of class interval.

$$m_1'(c) = m_1'$$

$$m_2'(c) = m_2' - \frac{c^2}{12}$$

$$m_3'(c) = m_3' - \frac{c^2 m_1}{4}$$

$$m_4'(c) = m_4' - \frac{c^2}{2}m_2' + \frac{7}{240}c^4$$

$$\Rightarrow m_1(c) = m_1$$

$$m_2(c) = m_2 - \frac{c^2}{12}$$

$$m_3(c) = m_3$$

$$m_4(c) = m_4 - \frac{c^2}{2}m_2 + \frac{7}{240}c^4$$

(*) Note how there are no correction for $m_n(c)$, $\forall n$ = odd.

When are these corrections valid?

① Observations from continuous data

② Freq curve of dist is tapered towards end (??)

③ Total freq should be large

④ Class length should not be too small.

◎ Relations b/w moments →

⊛ Cauchy-Schwarz Inequality for Expectations →

$$|E(f(x)\,g(x))| \le \sqrt{E(f(x)^2)} \cdot \sqrt{E(g(x)^2)}$$

Assumption: $E(f(x)^2) < \infty$, $E(g(x)^2) < \infty$

$$E(f(x) \cdot g(x)) < \infty$$

⊛ $E(x) = \sum_i x_i P(x = x_i) = \bar{x}$

↓ uniform dist

Here, $\frac{1}{n} \left| \sum_i f(x_i)\, g(x_i) \right| \le \sqrt{\frac{1}{n}\sum_i f^2(x_i)} \cdot \sqrt{\frac{1}{n}\sum_i g^2(x_i)}$

Let $f(x) = (x - \bar{x})^n$, $g = (x - \bar{x})$

$$\Rightarrow \frac{|m_{n+1}|}{\sigma} \le \sqrt{m_n}$$

, σ : Standard deviation

Let, $f(x) = (x - \bar{x}_0)^{n+1}$, $g(x) = (x - \bar{x})^n$

$$\Rightarrow |m_{2n+1}| \le \sqrt{m_{2n+2}} \cdot \sqrt{m_{2n}}$$

◎ Jensen's Inequality →

If ϕ is a convex function, and $E(|\phi(x)|) < \infty$, then,

$$E(|\phi(x)|) \ge \phi(E(x))$$

In the discrete case,

$$\frac{1}{n}\sum_i \phi(x_i) \ge \phi\left(\sum_i \frac{x_i}{n}\right)$$

Say, $\phi(x) = \frac{1}{x}$

Now,

$$\frac{1}{n}\sum_i \frac{1}{x_i} \geq \frac{n}{\sum_i x_i}$$

$$\Rightarrow \bar{x} \geq \frac{1}{\frac{1}{n}\sum_i \frac{1}{x_i}} \longrightarrow \underline{\text{Note}}: \text{This is AM-HM inequality.}$$

Now, let $\phi(x) = -\log(x)$

$$\therefore \frac{1}{n}\sum_i \log(x_i) \leq \log\left(\frac{1}{n}\sum_i x_i\right)$$

$$\Rightarrow \log\left\{\left(\prod_i x_i\right)^{1/n}\right\} \leq \log\left(\frac{1}{n}\sum_i x_i\right)$$

⊛ log is a monotonic function, so we may remove log to get AM-GM inequality!

∴ In general, ⟶ To get this, take AM ≥ GM and then set $x_i \to \frac{1}{x_i}$

$$QM \geq AM \geq GM \geq HM$$

(?)

Surprise addition !

Grouped data →

$$\Delta_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_i f_j |x_i - x_j|$$

So, here the data is all possible absolute differences b/w the data points

Why $f_i f_j$? f_i ways to select x_i, f_j ways to select x_j for same value of abs diff.

$$\Delta_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (x_i - x_j)^2 \qquad (??)$$

$$\Delta_2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{k} f_i f_j (x_i - x_j)^2$$

Now,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - \bar{x} + \bar{x} - x_j)^2$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[(x_i - \bar{x})^2 + (x_j - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x})\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left[(x_i - \bar{x})^2 + (x_j - \bar{x})^2\right] \quad (-2(x_i - \bar{x})(x_j - \bar{x}) \to 0)$$

$$= 2n^2\sigma_{\bar{x}}^2$$

⑪ Quartile deviation →

Inter Quartile Range → IQR

Reminder: for any R.V. X, the median m is defined as value for which,

$$P(X \geq m) \geq \frac{1}{2}, \quad P(X \leq m) \geq \frac{1}{2}$$

Generalizing this,

we define p^{th} quartile of a R.V. X, denoted as z_p, as,

$$P(X \leq z_p) \geq p, \quad P(X \geq z_p) \geq 1 - p$$

i.e. if $p = \frac{1}{4} \Rightarrow z_{1/4} \to Q_1$ (1st quartile)

$p = 1/2 \Rightarrow z_{1/2} \to Q_2$ (median)

$p = \frac{3}{4} \Rightarrow z_{3/4} \to Q_3$ (3rd Quartile)

Δ IQR $= Q_3 - Q_1$

Classes	f_i	C.F ($<c$)
$[a_0, a_1]$	f_1	$F_1 = f_1$
$(a_1, a_2]$	f_2	$F_2 = f_1 + f_2$
⋮	⋮	⋮
$(a_{k-1}, a_k]$	f_k	$F_k = n$
	$\sum_i f_i = n$	

$\Rightarrow$ for $F_{i+1} \geq \frac{n}{2}$ and $F_i < \frac{n}{4}$

$\Rightarrow (a_i, a_{i+1}] \to$ 1st quartile

and $F_{i+1} \geq \frac{3n}{4}$ and $F_i < \frac{3n}{4}$

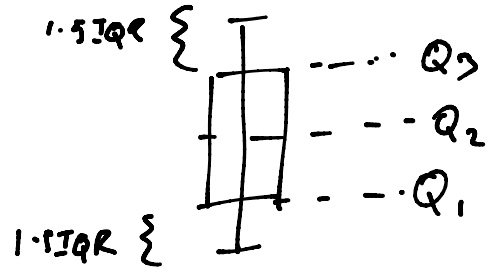
$\Rightarrow (a_i, a_{i+1}] \to$ 3rd Quartile

We use standard linearizing technique to determine value of the quartiles.

$$\tilde{Q}_1 = q_i + \frac{h}{f_{i+1}}\left(\frac{n}{4} - F_i\right)$$

$$\tilde{Q}_3 = q_j + \frac{h}{f_{j+1}}\left(\frac{3n}{4} - F_j\right)$$

⊛ Box and whisker plot (?) →



I assume this is how one data point, with spread, is represented in a plot.

Something about "coefficient of variation"?

△ $v = 100 \cdot \frac{\sigma_{\bar{x}}}{\bar{x}}$, given $\bar{x} \neq 0$

Some sort of confidence measure? Who knows.