Lecture 2 – Measuring variability and association

Plan:

(1) Briefly review concept of variance / standard deviation

(2) Introduce concept of **correlation**

- Pearson correlation
- Phi coefficient
- Point-biserial correlation
- Spearman correlation

- Measuring variability

↳ How much, on average, does each number differ from the most typical value?

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	-4	16
3	-3	9
5	-1	1
8	2	4
12	6	36
$\bar{x} = 6$		

← find average

$$= \frac{16 + 9 + 1 + 4 + 36}{5}$$

$$= \frac{66}{5} = 13.2$$

Average "squared" deviation

↑ average "deviation" = 0

Not helpful!

Based on this, we define variance as the average of the squared deviations.

$$\text{variance} = \frac{\sum (x_i - \bar{x})^2}{N}$$

Definition: the standard deviation is the square root of the variance.

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

Ex: for the data set 2, 3, 5, 8, 12, the standard deviation is:

$$SD = \sqrt{\text{variance}} = \sqrt{13.2} = 3.63$$

Notation:

Standard deviation typically written as σ

↳ variance = σ^2

Measuring association

Consider two sets of scores — to what degree are they associated?

↳ i.e., how do they co-vary?

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
6	6	2	2	4
2	2	-2	-2	4
5	6	1	2	2
3	4	-1	0	0
4	2	0	-2	0
$\bar{X}=4$	$\bar{Y}=4$			$\sum = 10$
$\sigma_X = 1.41$	$\sigma_Y = 1.79$			

Define: covariance

$$\sigma_{XY} = \frac{1}{N}\sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$\sum = 10 \longrightarrow \sigma_{XY} = \frac{1}{5} \times 10 = 2$$

From covariance, we can define the Pearson correlation:

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

So for our example,

$$r = \frac{2}{(1.41)(1.79)} = 0.79$$

Facts:

* $-1 \leq r \leq 1$
* as $r \longrightarrow \pm 1$, degree of association increases

Alternative measures of association

The Pearson correlation assumes **both** variables X, Y are **continuous variables** (i.e., on interval or ratio scale)

What if this is NOT the case?

Three alternative "correlations"

* Phi coefficient (X, Y = dichotomous variables)
* Point-biserial correlation (X = dichotomous, Y = continuous)
* Spearman correlation (X, Y = ordinal variables)

Phi coefficient - for pairs of dichotomous (yes/no) variables.

- Consider two test items scored as correct (1) or incorrect (0)

		X = 0	X = 1
Y	0	8	12
	1	4	16

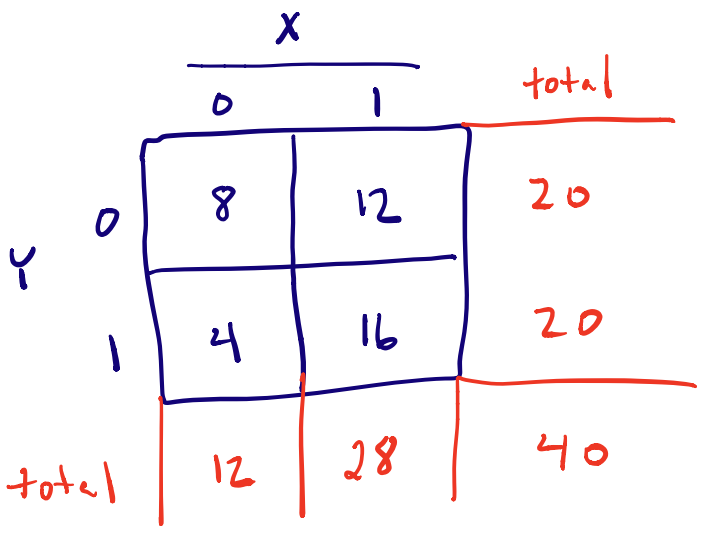
Phi coefficient:

$$\varphi_{XY} = \frac{P_{XY} - P_X P_Y}{\sqrt{P_X(1-P_X)\,P_Y(1-P_Y)}}$$

P_{XY} = proportion scoring 1 on both X and Y

P_X = proportion scoring 1 on X

P_Y = proportion scoring 1 on Y



$$\rightarrow \quad P_{XY} = \frac{16}{40} = 0.4$$

$$P_X = \frac{28}{40} = 0.7$$

$$P_Y = \frac{20}{40} = 0.5$$

So the phi coefficient is

$$\varphi = \frac{P_{XY} - P_X P_Y}{\sqrt{P_X(1-P_X)P_Y(1-P_Y)}} = \frac{0.4 - (0.7)(0.5)}{\sqrt{(0.7)(0.3)(0.5)(0.5)}}$$

$$= \frac{0.05}{0.229} = 0.22$$

Facts:

* Phi coefficient interpreted exactly same as Pearson correlation

* but, the range $[-1, 1]$ is reduced when $P_X \neq P_Y$

(i.e, when item difficulty is not equivalent)

Point-biserial correlation → X = dichotomous, Y = continuous

X	0	0	0	0	1	1	1	1	1	1
Y	1	4	0	5	7	4	5	7	9	8

Point-biserial formula

$$r_{pbis} = \left(\frac{\bar{Y}_1 - \bar{Y}}{\sigma_Y}\right)\sqrt{\frac{P_X}{1 - P_X}}$$

we can compute

$\bar{Y}_1 = 6.67$

$\bar{Y} = 5$

$\sigma_Y = 2.76$

$P_X = 0.6$

so $r_{pbis} = \left(\frac{6.67 - 5}{2.76}\right)\sqrt{\frac{0.6}{0.4}}$

$= (0.605)(1.225)$

$= 0.74$

Facts:

* range $[-1, 1]$ is reduced if $P_X \neq 0.5$
* value identical to Pearson correlation, but slightly easier to compute.

Spearman correlation → X, Y = ordinal variables.

(e.g., Likert-scaled measurements)

X	Y
7	2
1	3
3	6
5	7
4	5

Spearman formula:

$$r_s = 1 - \frac{6 \sum D_i^2}{N^3 - N}$$

D_i = difference in ranks for the i^{th} pair.

X	Y	Rank X	Rank Y	D_i	D_i^2
7	2	1	5	−4	16
1	3	5	4	1	1
3	6	4	2	2	4
5	7	2	1	1	1
4	5	3	3	0	0

→ $\sum D_i^2 = 22$

So

$$r_s = 1 - \frac{6 \sum D_i^2}{N^3 - N} = 1 - \frac{6 \cdot 22}{5^3 - 5} = 1 - \frac{132}{120}$$

$$= -0.10$$