Srednja vrijednost i standardno odstupanje

Prilikom ponavljanja *n* mjerenja, uz iste (koliko je to moguće) uvjete (isti mjeritelj, mjerni instrumenti, mjerna metoda i okolni uvjeti), neke stalne fizikalne veličine *x*, dobit će se *n* izmjerenih vrijednosti (neka su označene sa $x_1, x_2, \ldots, x_{n-1}, x_n$). Kako su svi pojedinačni rezultati dobiveni uz jednake uvjete, ni jedan od njih nema prednost pred drugima, pa će najvjerojatnija vrijednost mjerene veličine biti aritmetička sredina $\bar{x}$ pojedinačnih rezultata:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_{n-1} + x_n}{n}$$

Pojedinačni rezultati mjerenja to se međusobno manje razlikuju što je mjerni postupak precizniji. Brojčanu procjenu preciznosti mjernog postupka daje standardno odstupanje (standardna devijacija):

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n}}$$

Umjesto apsolutnog iznosa standardnog odstupanja često se koristi njegov relativni iznos (koeficijent varijacije). On se dobije dijeljenjem apsolutnog iznosa s aritmetičkom sredinom:

$$V = \frac{\sigma}{\bar{x}} \cdot 100 [\%]$$

Što je standardno odstupanje manje, to je preciznost mjerenja veća. Preciznost odnosno nepreciznost ne smije se zamijeniti pojmom netočnost, koji pokazuje bliskost slaganja mjernog rezultata s pravom vrijednosti mjerene veličine. Poboljšanje preciznosti nužan je uvjet za smanjene netočnosti.

Za uočavanje razlike tih dvaju pojmova neka posluži sljedeći primjer. Strijelci A, B i C gađaju svoje mete. Nakon ispucanih 10 metaka stanje na metama prikazano je na slici 1. Strijelac A pokazao je dobru preciznost jer se pogodci malo rasipaju, a također i dobru točnost jer se svi pogodci nalaze u središtu ili u njegovoj blizini. Strijelac B pokazao je također dobru preciznost jer se i njegovi pogodci malo rasipaju, ali mu je točnost manja. Njegovi su se pogodci grupirali dalje od središta mete. Strijelac C pokazao je malu preciznost i malu točnost jer mu je rasipanje pogodaka veliko, a pogodci su grupirani daleko od središta mete.

Slika 1. Mete strijelaca; a. precizan i točan, b. precizan, ali netočan i c. neprecizan i netočan

Varijanca

Često se umjesto standardnog odstupanja kao procjena preciznosti mjernog postupka koristi varijanca:

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

Kovarijanca

Kovarijanca pokazuje koliko se dvije varijable mijenjaju zajedno. To je različito od varijance, koja pokazuje koliko se jedna varijabla mijenja. Kovarijanca postaje više pozitivnom za svaki par vrijednosti koji se razlikuje od njihovih srednjih vrijednosti u istom smjeru, te postaje više negativna za svaki par vrijednosti koji se razlikuje od njihovih srednjih vrijednosti u suprotnim smjerovima. Ako postoje tri skupa vrijednosti $(x_1, x_2, \ldots, x_n)$, $(y_1, y_2, \ldots, y_n)$ i $(z_1, z_2, \ldots, z_n)$ tada su njihove kovarijance:

$$\sigma_{xy}^2 = \sigma_{yx}^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n};\ \sigma_{xz}^2 = \sigma_{zx}^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(z_i - \bar{z}\right)}{n};\ \sigma_{yz}^2 = \sigma_{zy}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(z_i - \bar{z}\right)}{n}$$

Ili u matričnom obliku:

$$C = \begin{bmatrix} \sigma_{xx}^2 & \sigma_{xy}^2 & \sigma_{xz}^2 \\ \sigma_{yx}^2 & \sigma_{yy}^2 & \sigma_{yz}^2 \\ \sigma_{zx}^2 & \sigma_{zy}^2 & \sigma_{zz}^2 \end{bmatrix}$$

Važno je uočiti da vrijedi $\sigma_{xy}^2 = \sigma_{yx}^2$, $\sigma_{xz}^2 = \sigma_{zx}^2$ i $\sigma_{yz}^2 = \sigma_{zy}^2$. Općenito ako postoji *m* skupova od *n* mjerenja kovarijanca se može napisati kao matrica *m* x *n* (matrica kovarijanci je simetrična):

$$C = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2n}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{m1}^2 & \sigma_{m2}^2 & \dots & \sigma_{mn}^2 \end{bmatrix}$$

Koeficijent korelacije

Koeficijent korelacije između dvije slučajne varijable *x* i *y* sa srednjim vrijednostima $\overline{x}$ i $\overline{y}$ te standardnim devijacijama σ_x i σ_y definiran je izrazom:

$$R_{xy} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_{xx}^2 \sigma_{yy}^2}} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}}$$

Napomena: $\sigma_{xx}^2 = \dfrac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(x_i - \overline{x}\right)}{n} = \dfrac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}{n} = \sigma_x^2$

Koeficijent korelacije pokazuje stupanj linearne zavisnosti između varijabli. Što je koeficijent korelacije bliže 1 ili -1, veća je korelacija između varijabli. Ako su varijable nezavisne koeficijent korelacije je 0, ali obratno ne vrijedi (odnosno ako su dvije varijable zavisne njihov koeficijent korelacije može biti 0) zbog toga što koeficijent korelacije ustanovljava jedino linearnu zavisnost između varijabli. Pretpostavke kod računanja koeficijenta korelacije su:

- linearna zavisnost između dviju varijabli *x* i *y*,
- kontinuirane slučajne varijable,
- obje varijable moraju imati normalnu razdiobu,

- varijable *x* i *y* moraju biti nezavisne jedna o drugoj.

Ako postoje tri skupa vrijednosti $(x_1, x_2, \dots, x_n)$, $(y_1, y_2, \dots, y_n)$ i $(z_1, z_2, \dots, z_n)$ tada su njihovi koeficijenti korelacije:

$$R_{xy} = R_{yx} = \frac{\sigma_{xy}^2}{\sqrt{\sigma_{xx}^2 \sigma_{yy}^2}}; \; R_{xz} = R_{zx} = \frac{\sigma_{xz}^2}{\sqrt{\sigma_{xx}^2 \sigma_{zz}^2}}; \; R_{yz} = R_{zy} = \frac{\sigma_{yz}^2}{\sqrt{\sigma_{yy}^2 \sigma_{zz}^2}}$$

Ili u matričnom obliku:

$$R = \begin{bmatrix} R_{xx} & R_{xy} & R_{xz} \\ R_{yx} & R_{yy} & R_{yz} \\ R_{zx} & R_{zy} & R_{zz} \end{bmatrix}$$

Važno je uočiti da vrijedi $R_{xy} = R_{yx}$, $R_{xz} = R_{zx}$ i $R_{yz} = R_{zy}$. Općenito ako postoji *m* skupova od *n* mjerenja koeficijent korelacije se može napisati kao matrica *m* x *n* (matrica koeficijenta korelacije je simetrična):

$$R = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1n} \\ R_{21} & R_{22} & \dots & R_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ R_{m1} & R_{m2} & \dots & R_{mn} \end{bmatrix}$$

Nekoliko autora dalo je smjernice za tumačenje koeficijenta korelacije. Cohen (1988) je predložio tumačenje koeficijenta korelacije u psihološkim istraživanjima [1]:

Korelacija	Negativna	Pozitivna
Mala	-0,29 do -0,10	0,10 do 0,29
Srednja	-0,49 do -0,30	0,30 do 0,49
Velika	-0,50 do -1,00	0,50 do 1,00

Svi takvi kriteriji su proizvoljni i ne treba ih se striktno pridržavati. To je iz razloga što tumačenje koeficijenta korelacije ovisi o kontekstu i namjeni. Koeficijent korelacije od 0,90 može biti nedovoljan ako se provjerava fizikalni zakon pomoću vrlo preciznih i točnih instrumenata, ali se može smatrati i vrlo velikim u npr. istraživanjima u socijalnim znanostima.

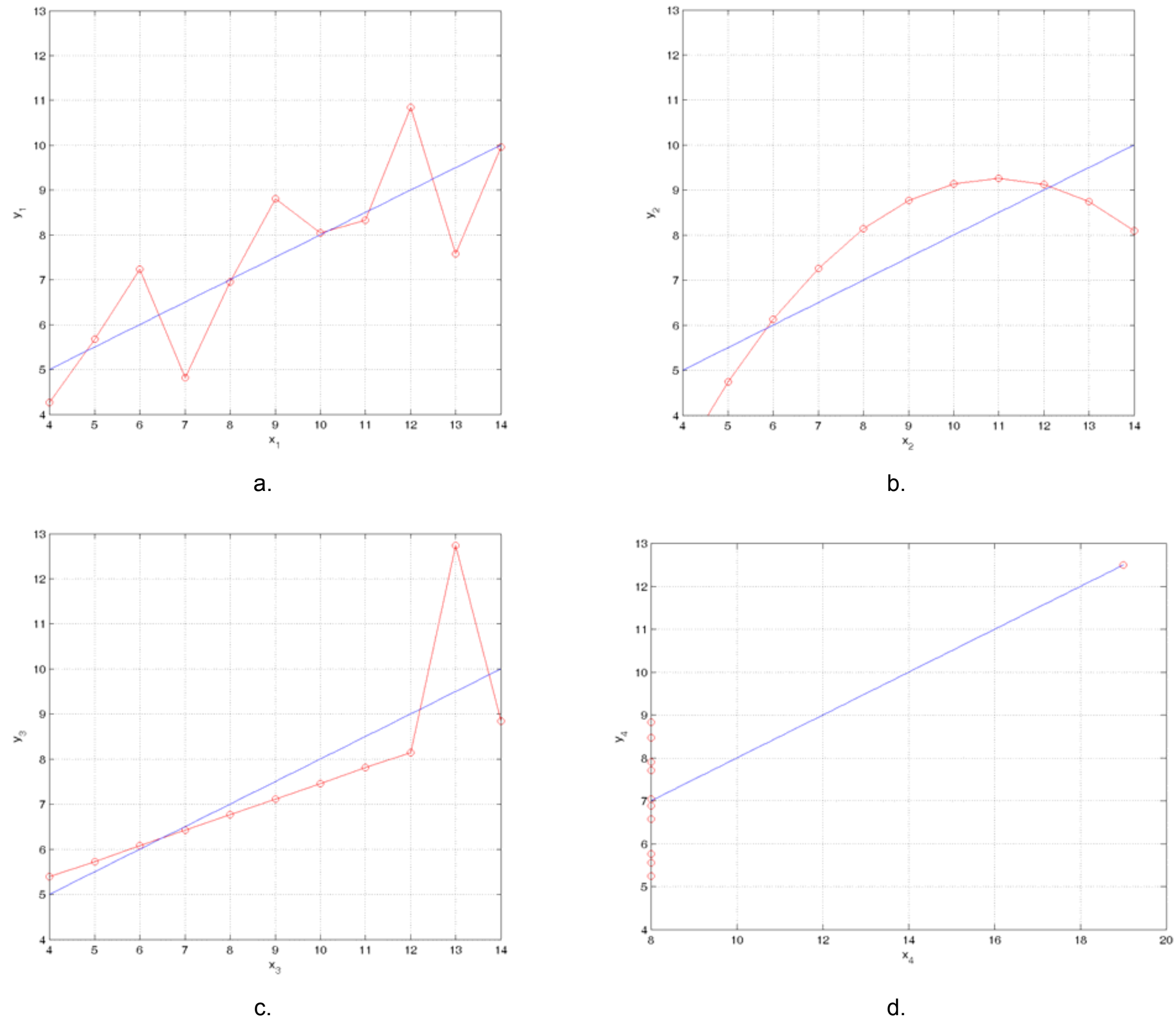
Korelacija i linearnost

Ovdje će biti prikazan primjer 4 skupa od dvije varijable i njihovo ponašanje koje je prvi puta opisao Francis Anscombe [2]. Uspoređivani su skupovi varijabli x_1 i y_1, x_2 i y_2, x_3 i y_3 te x_4 i y_4. Sve četiri varijable y_1, y_2, y_3 i y_4 imaju srednju vrijednost 7,5, standardno odstupanje 1,93, varijancu 3,75, kovarijancu 5,00, koeficijent korelacije 0,81 i pravac linearne regresije $y = 0{,}5x + 3$ (tablica 1., vidjeti i primjer u Excelu). Razdioba te četiri varijable je jako različita. Prva varijabla y_1 ima normalnu razdiobu i odgovara onom što se očekuje kada se koreliraju dvije varijable. Druga varijabla y_2 nema normalnu razdiobu. Vidljivo je da postoji zavisnost između dvije varijable, ali ona nije linearna. U trećem primjeru (varijabla y_3) linearna zavisnost je velika osim jednog podatka koji odskače. Četvrti skup podataka (varijabla y_4) pokazuje kako je jedan podatak (koji jako odskače) dovoljan da doprinese velikom koeficijentu korelacije iako odnos između dvije varijable nije linearan. Slika 2. prikazuje 4 skupa podataka od dvije varijable i pravac linearne regresije.

Tablica 1. Primjer 4 skupa od dvije varijable Francisa Anscombea

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
4,00	4,26	4,00	3,10	4,00	5,39	8,00	5,25
5,00	5,68	5,00	4,74	5,00	5,73	8,00	5,56
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,76
7,00	4,82	7,00	7,26	7,00	6,42	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	6,89
9,00	8,81	9,00	8,77	9,00	7,11	8,00	7,04
10,00	8,04	10,00	9,14	10,00	7,46	8,00	7,71
11,00	8,33	11,00	9,26	11,00	7,81	8,00	7,91
12,00	10,84	12,00	9,13	12,00	8,15	8,00	8,47
13,00	7,58	13,00	8,74	13,00	12,74	8,00	8,84
14,00	9,96	14,00	8,10	14,00	8,84	19,00	12,50

Na ovom primjeru je važno uočiti da statistički pokazatelji iako jednaki ne znače da se skupovi podataka ponašaju jednako. Važno je ispitati i nacrtati grafički podatke koje se žele analizirati.



Slika 2. Anscombova 4 skupa od dvije varijable, a. prvi skup, b. drugi skup, c. treći skup i d. četvrti skup

1. Cohen, J., *Statistical power analysis for the behavioral sciences*, (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
2. Anscombe, Francis J., *Graphs in statistical analysis*, American statistician, 27, 17-21, 1973.