26.10.2021

Seminar W5 - 832

(X,Y) 2-dimensional characteristic on an $m \times n$ dataset

$$\operatorname{cov}(X,Y) = \frac{1}{mn} \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\dots$	f_{mj}	$\dots$	f_{mn}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

$$\bar{\sigma}_X = \sqrt{\frac{1}{m} \sum_{i=1}^{m} f_{i.} (x_i - \bar{x})^2}$$

$$\bar{\sigma}_Y = \sqrt{\frac{1}{n} \sum_{j=1}^{n} f_{.j} (y_j - \bar{y})^2}$$

The correlation coefficient: $\bar{\rho} = \dfrac{\operatorname{cov}(X,Y)}{\bar{\sigma}_X \, \bar{\sigma}_Y}$

$$\bar{\rho} \in [-1, 1] \iff |\bar{\rho}| \leq 1$$

Exercise 3. The following table represents the annual consumption (between 2000 and 2009) of cheese in the U.S. (in lbs), along with the number of people who died by becoming tangled in their bedsheets, in the same time period.

Year (Y)	Cheese consumed (C) (in lbs)	Bedsheet tanglings (B) (in deaths)
2000	29.8	327
2001	30.1	456
2002	30.5	509
2003	30.6	497
2004	31.3	596
2005	31.7	573
2006	32.6	661
2007	33.1	741
2008	32.7	809
2009	32.8	717

Find the correlation coefficient of (B, C) and the lines of regression of C on B and of B on C.

Sol.:

$$\overline{c} = \frac{1}{10} \cdot \sum c_i = 31.52$$

$$\overline{b} = 588.6$$

$$\overline{\rho} = \frac{\operatorname{cov}(C, B)}{\overline{\sigma_c} \cdot \overline{\sigma_B}}$$

$$\operatorname{cov}(C, B) = \frac{1}{10} \sum_{i=1}^{10} (c_i - \overline{c}) \cdot (b_i - \overline{b}) =$$

$$= 154.268$$

$$\overline{\sigma_c} = \sqrt{\frac{1}{10} \sum (c_i - \overline{c})^2} = 1.167733$$

$$\overline{\sigma_B} = 139.4892$$

$$\overline{\rho} = \frac{154.268}{1.167733 \cdot 139.4892} = 0.94709 = 94\%$$

Lines of regression:

$$c - \overline{c} = \overline{\rho} \cdot \frac{\overline{\sigma_c}}{\overline{\sigma_B}} \cdot (b - \overline{b})$$

(regression of c on B)

$$b - \overline{b} = \overline{\rho} \cdot \frac{\overline{\sigma_B}}{\overline{\sigma_c}} \cdot (c - \overline{c})$$

(regression of B on c)

Exercise 1. According to the International Data Base of the U.S. Census Bureau, the population of the world grows according to the following table:

Year (Y)	Population(P) (million people)
1950	2558
1955	2782
1960	3043
1965	3350
1970	3712
1975	4089
1980	4451
1985	4855
1990	5287
1995	5700
2000	6090
2005	6474
2010	6864

Denote by Y and P the year and population, respectively.

(a) Find the standard deviations $\overline{\sigma}_Y$ and $\overline{\sigma}_P$;

(b) Find the centroid of the distribution of the characteristic (Y, P);

(c) Find the corelation coefficient of (Y, P);

(d) Find the regression lines of Y on P and of P on Y;

(e) The world population in 2015 was about 7378 million, while in 2020 it was about 7795 million. How well does the line of regression of P on Y predict these values?

(f) According to this line of regression, in what year did the world population reach the 7 billion milestone?

(a) $\bar{y} = 1980 \qquad \bar{p} = 4558.077$

$\overline{\sigma}_Y = 18.708 \qquad \overline{\sigma}_P = 1389.945$

(b) $(\bar{y}, \bar{p}) = (1980, 4558.077)$

(c) $\bar{\rho} = \dfrac{\operatorname{cov}(Y,P)}{\overline{\sigma}_Y \cdot \overline{\sigma}_P} = 0.997$

(d) regression line of P on Y

$$p - \bar{p} = \bar{\rho} \cdot \frac{\overline{\sigma}_P}{\overline{\sigma}_Y} \cdot (y - \bar{y})$$

$$\Rightarrow \quad P = 0.997 \cdot \frac{1389.94}{18.7} (y - 1980) + 4558.077$$

regression line of Y on P:

$$y - \bar{y} = \bar{\rho} \cdot \frac{\overline{\sigma}_Y}{\overline{\sigma}_P} \cdot (p - \bar{p})$$

$$\Rightarrow \quad y = 1980 + 0.997 \cdot \frac{18.7}{1389.94} \cdot (p - 4558.077)$$

(e)

$$P = \underbrace{0.997 \cdot \frac{1389.94}{18.7} (y - 1980) + 4558.077}_{f(y)}$$

$f(2015) = 7151.8 \qquad$ real value: 7378

$f(2020) = 7522.3$ real value: 7795

$$y = \underbrace{1980 + 0.997 \cdot \frac{18.7}{1389.54} \cdot (p - 4558.077)}_{g}$$

$g(7000) = 2012.8$ real value: 2011

Exercise 4. In a certain experiment, the vibrations of the high E string of a guitar have been tested. The following contingency table data shows the results of the measurements of the tension T, respectively the speed S:

Tension \ Speed	0.3 km/s	0.35 km/s	0.4 km/s	0.45 km/s	0.5 km/s
25 N	8	0	0	0	0
36 N	2	7	0	0	0
49 N	1	2	8	0	0
64 N	0	1	2	6	0
81 N	0	0	1	2	10

Do you suspect that there is a linear correlation between $\sqrt{T}$ and S? Find the conditional means of the two-dimensional characteristic $(\sqrt{T}, S)$, the coefficient of correlation and plot its curves of regression.

$\sqrt{T}$ \ S	0.3	0.35	0.4	0.45	0.5	
5	8	0	0	0	0	8
6	2	7	0	0	0	9
7	1	2	8	0	0	11
8	0	1	2	6	0	9
9	0	0	1	2	10	13
	11	10	11	8	10	50

(1) *The* ***conditional mean*** *of* Y, *given* $X = x_i$, *is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i.}} \sum_{j=1}^{n} f_{ij} y_j, \; i = \overline{1, m}.$$

(2) *The* ***conditional mean*** *of* X, *given* $Y = y_j$, *is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{.j}} \sum_{i=1}^{m} f_{ij} x_i, \; j = \overline{1, n}.$$

The conditional mean of Y, given $X = x_3$

$$\overline{y}_3 = \overline{y}(x_3) = \frac{1}{11} \cdot (1 \cdot 0.3 + 2 \cdot 0.35 + 8 \cdot 0.4) =$$

$$= \frac{1}{11}(0.3 + 0.7 + 3.2) = \frac{4.2}{11} = 0.38182$$

The conditional mean of X, given $Y = y_2$

$$\overline{x_2} = \overline{x}(y_2) = \frac{1}{10} \cdot (7 \cdot 6 + 2 \cdot 7 + 1 \cdot 8) =$$

$$= \frac{1}{10} \cdot (42 + 14 + 8) = \frac{64}{10} = 6.4$$

$\overline{y_3} = 0.38$

$\overline{y_1} = 0.3$

$\overline{y_2} = 0.33889$

$\overline{y_4} = 0.42778$

$\overline{y_5} = 0.48077$

$$\bar{\rho} = \frac{\operatorname{cov}(\underbrace{\sqrt{T}}_{= U}, S)}{\overline{\sigma}_{\sqrt{T}} \cdot \overline{\sigma}_S}$$

$$S - \bar{S} = \bar{\rho} \cdot \frac{\overline{\sigma_S}}{\overline{\sigma_{\sqrt{T}}}} (\mu - \bar{\mu})$$