26.10.2021

Seminar W5 - 831

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?

(X,Y) two-dim characteristic

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

$X \setminus Y$	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}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

$$\overline{x} = \frac{1}{mn}\cdot\sum_{i=1}^{m} f_{i.}\cdot x_i$$

$$\overline{y} = \frac{1}{mn}\cdot\sum_{j=1}^{n} f_{.j}\cdot y_j$$

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

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

$$\overline{\mu}_{k_1k_2} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\,(x_i-\overline{x})^{k_1}\cdot(y_j-\overline{y})^{k_2}$$

$$\overline{\sigma}_X = \sqrt{\overline{\mu}_{2,0}} \qquad \overline{\sigma}_Y = \sqrt{\overline{\mu}_{0,2}}$$

Corelation coef: $\overline{\rho} = \dfrac{\operatorname{cov}(X,Y)}{\overline{\sigma}_X\cdot\overline{\sigma}_Y} \in [-1, 1]$

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) $\bar{y} =$

$$= \frac{1}{13}\left(1950 + \dots + 2010\right) =$$

$$= 1980$$

$$\bar{p} = 4558.07$$

$$\overline{\sigma}_Y = \sqrt{\frac{1}{13}\sum (y_i - 1980)^2} = 18.7082$$

$$\overline{\sigma}_P = 1446.7$$

$$Cov(Y,P) = \frac{1}{13}\sum_{i=1}^{13} (y_i - 1980)(p_i - 4558.07)$$

$$= 28704.16$$

(b) centroid: $(\bar{y}, \bar{p}) = (1980, 4558.07)$

(c) $$\bar{\rho} = \frac{Cov(Y,P)}{\overline{\sigma}_Y \cdot \overline{\sigma}_P} = \frac{28704.16}{18.7082 \cdot 1446.7} = 0.99765$$

(d) The regression lines: $\underbrace{\qquad}_{}$ $y - \bar{y} = \frac{\overline{\sigma}_Y}{\overline{\sigma}_P} \cdot \frac{1}{\bar{\rho}} \cdot (p - \bar{p})$

- of P on Y: $p - \bar{p} = \bar{\rho} \cdot \frac{\overline{\sigma}_P}{\overline{\sigma}_Y} \cdot (y - \bar{y})$

- of Y on P: $y - \bar{y} = \bar{\rho} \cdot \frac{\overline{\sigma}_Y}{\overline{\sigma}_P} \cdot (p - \bar{p})$

These are two different lines!

In our case, the lines are:

- of P on Y : $p - 4558.07 = 0.997 \cdot \frac{1446.7}{18.7}(y - 1980)$
- of Y on P : $y - 1980 = 0.997 \cdot \frac{18.7}{1446.7} \cdot (p - 4558.07)$

e) Regression line of P on Y :

$$p = f(y) = 4558.07 + 0.997 \cdot \frac{1446.7}{18.7}(y - 1980)$$

$f(2015) = 7257.8$ Reality: $p = 7378$

$f(2020) = 7522.3$ Reality: $p = 7795$

f) $y = g(p) = 1980 + 0.997 \cdot \frac{18.7}{1446.7}(7000 - 4558.07) =$

$= 2011.5$

$g(7000) = 2011.5$ Reality $y = 2011$

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.

$\bar{c} = 31.520$

$\bar{b} = 588.60$

$\bar{\sigma}_c = 1.2309$

$\bar{\sigma}_B = 147.03$

$cov(C, B) = 171.47$

$$\bar{\rho} = \frac{171.47}{1.2309 \cdot 147.03} = 0.94713$$

line of regression of B on C

$$b - \bar{b} = \bar{\rho} \cdot \frac{\bar{\sigma}_B}{\bar{\sigma}_c} \cdot (c - \bar{c})$$

$$b - 588.60 = 0.94713 \cdot \frac{147.03}{1.2309} (c - 31.520)$$

line of regression of C on B

$$c - 31.520 = 0.94713 \cdot \frac{1.2309}{147.03} (b - 588.60)$$

Exercise 2. The time it takes to transmit a file always depends on the file size. Suppose you transmitted 30 files, with the average size of 126 Kbytes and the standard deviation of 35 Kbytes. The average transmittance time was 0.04 seconds with the standard deviation of 0.01 seconds. The correlation coefficient between the time and the size was 0.86. Based on this data, fit a linear regression model and predict the time it will take to transmit a 400 Kbyte file.

size, time

(S, T)

$\bar{s} = 126 \qquad \bar{\sigma}_s = 35$

$\bar{t} = 0.04 \qquad \bar{\sigma}_T = 0.01$

$\bar{\rho} = 0.86$

The line of regression of T on S

$$t - \bar{t} = \bar{\rho} \cdot \frac{\bar{\sigma}_T}{\bar{\sigma}_S} \cdot (s - \bar{s})$$

$$t - 0.04 = 0.86 \cdot \frac{0.01}{35} \cdot (s - 126)$$

$$\Rightarrow t = 0.04 + 0.86 \cdot \frac{0.01}{35} \cdot (400 - 126) \doteq 0.10733$$