# Mathematical Statistics

## Seminar Exercises: Week 5

**Definition.** Consider $(X, Y)$ to be a two-dimensional characteristic whose distribution is given by the following **contingency table**:

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

We then define:

- The **centroid** of $(X, Y)$:

$$(\overline{x}, \overline{y}) = \left( \frac{1}{m} \sum_{i=1}^{m} f_{i.} \cdot x_i, \frac{1}{n} \sum_{j=1}^{n} f_{.j} \cdot y_j \right)$$

- The **covariance** of $(X, Y)$:

$$\text{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \cdot (x_i - \overline{x})(y_j - \overline{y}) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$$

- The **correlation coefficient**:

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}} \sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}$$

- The **coefficient of regression of $Y$ on $X$**:

$$\overline{a}_{Y|X} = \overline{\rho}\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}$$

**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?

**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.

**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 $C$.

*Conclusion.* Does the result surprise you? Does it uncover a nefarious effect of cheese in the lives (and deaths) of U.S. citizens who like to lie in bed? Of course not! It is merely an illustration of the phrase:

**Correlation does not imply causation.**

More examples in the same vein can be found at:

`http://tylervigen.com/spurious-correlations`

**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.