LECTURE 7: COVARIANCE AND CORRELATIONS

AP STATS

Abstract. We review 2-way tables and relative frequencies. We define covariance of two variables and then definee the correlation coefficient using covariance.

1. Two-way Tables

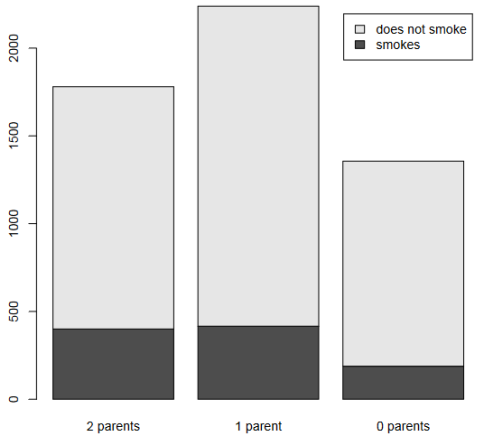
Relationships between two categorical variables can be shown through a two-way table:

Example 1. The following example is Example 1.9 in Here.

In 1964, Surgeon General Dr. Luther Leonidas Terry published a landmark report saying that smoking may be hazardous to health. This led to many influential reports on the topic, including the study of the smoking habits of 5375 high school children in Tucson in 1967. Here is a two-way table summarizing some of the results:

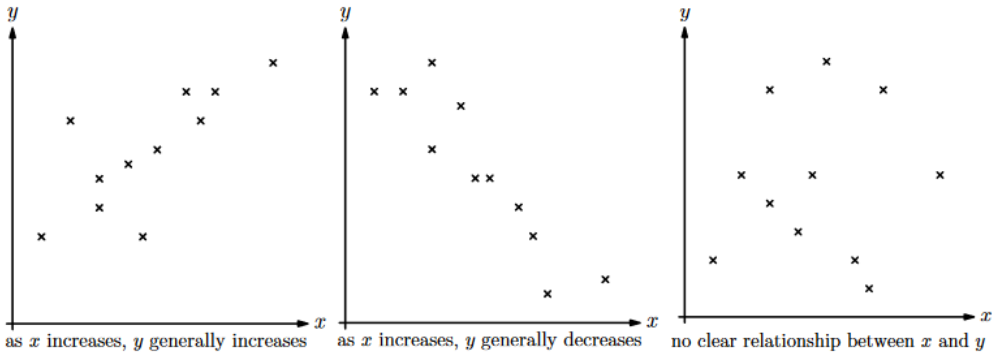
	students smoke	students do not smoke	Total
2 parents smoke	400	1380	1780
1 parent smokes	416	1823	2239
0 parents smoke	188	1168	1356
Total	1004	4371	5375

- The **relative frequencies** can easily be computed
- You can create a segmented bar graph for each category such as the one below:



2. Covariance

To study the relationship between two variables, we oftentimes create a *scatterplot* of the data.



We use two tools to study the relationships:

(1) Correlation
- studies the relationship in a symmetric manner
- correlation $\neq$ causation!!
- does not establish cause and effect

(2) Regression
- relationship between **response variable** and **explanatory variable(s)**
- response variables are typically denoted Y while explanatory variables are typically denoted X_i

Definition. The **covariance** measures the linear relationship between a pair of quantitative variables $X = \{x_1, x_2, \ldots, x_n\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$. The **covariance** is computed by

$$\operatorname{cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y), \quad \mu_X = \text{mean of } X,\ \mu_Y = \text{mean of } Y$$

- If $\operatorname{cov}(X, Y) > 0$, then the terms $(x_i - \mu_X)(y_i - \mu_Y)$ in the sum are more positive than negative, which happens whenever x_i, y_i are both above, or both below, the mean together
- $\operatorname{cov}(X, X) = \operatorname{Var}(X) = \sigma_X^2$
- Whenever X, Y are independent variables, then $\operatorname{cov}(X, Y) = 0$ however the converse is not true!

Definition. The **pearson's correlation**, r, is the covariance of the standardized versions of X and Y:

$$r(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_x \sigma_Y} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_X}{\sigma_X}\right)\left(\frac{y_i - \mu_Y}{\sigma_Y}\right)$$

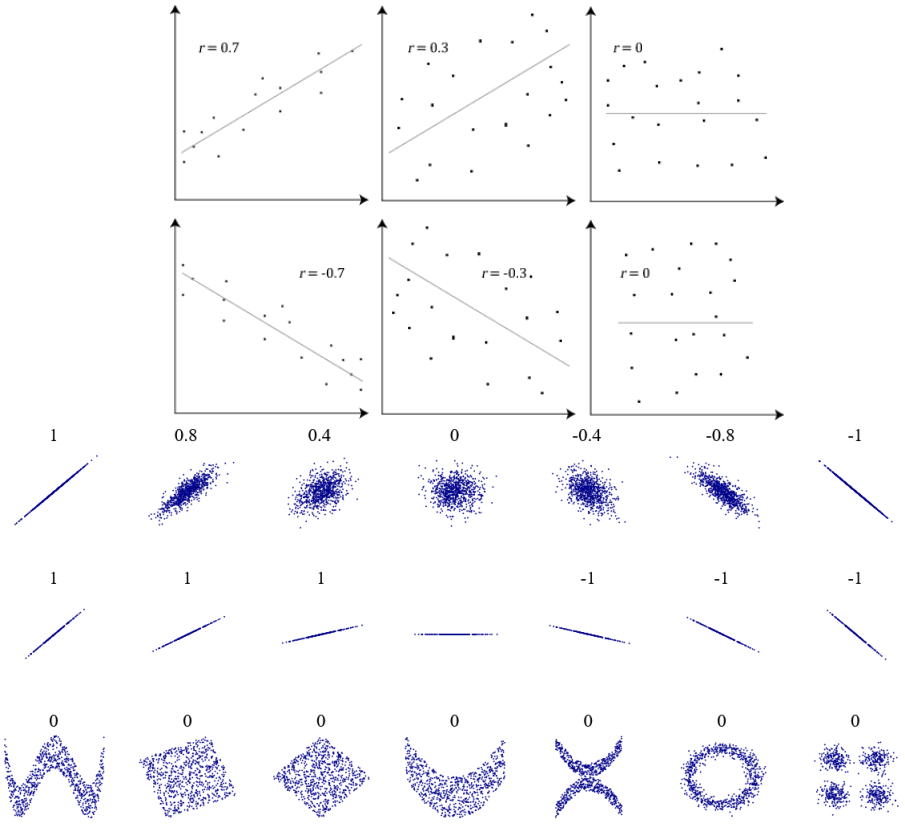
- Whenever $r(X, Y) = 0$, we say X and Y are uncorrelated

- r is a number that is always sitting $-1 \leq r \leq 1$.[1]

The values of r:

- as $r \to 1$, the data tends to more **positive linear correlation**
- as $r \to -1$, the data tends to become more **negative linear correlation**
- as $r \to 0$, the data does **not** have any linear correlation

Below are pictures of different scatterplots with different r values (credit from Wikipedia)



Whenever commenting on the strength of correlation, always note the two things:

(1) If r is positive or negative
(2) The strength of the correlation:

$r-$ value	$0 \leq \lvert r \rvert \leq 0.25$	$0.25 \leq \lvert r \rvert \leq 0.5$	$0.5 \leq \lvert r \rvert \leq 0.8$	$0.8 \leq \lvert r \rvert \leq 1$
correlation strength	very weak	weak	moderate	strong

[1] This comes from *Cauchy-Schwartz inequality*!