Week 8: Bivariate random variables, Covariance, Correlation

Armenak Petrosyan

Tuesday class

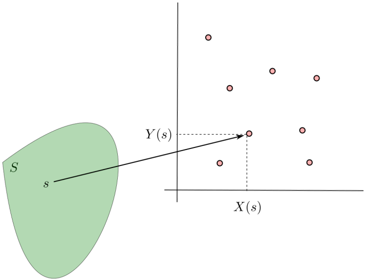
Bivariate random variable

- Random variable $X(s)$ - measurement associated with the outcome $s \in S$.
- Sometimes we can have multiple measurements associated with the same outcome

$$s \mapsto (X(s), Y(s)).$$

Example

Let S be the set of all students in our class. $X(s)$ denote the height and $Y(s)$ denote the weight of student s.



Joint pmf of bivariate random variable

- We will consider discrete bivariate random variables.
- Denote
$$\mathrm{Range}(X, Y) = \{(X(s), Y(s)) : \ s \in S\}.$$

Definition

Let (X, Y) be discrete bivariate random variable on outcome set S. For every $(x, y) \in \mathrm{Range}(X, Y)$, denote

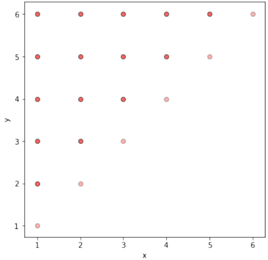
$$f(x, y) = P(X = x \text{ and } Y = y).$$

$f(x, y)$ is called the **joint probability mass function** of random variables X and Y.

We typically define $f(x, y) = 0$ for $(x, y) \in \mathbb{R}^2$ that are not in $\mathrm{Range}(X, Y)$.

Example

- Roll a fair dice twice.
- Let X be the smaller value and Y be the larger value.
- $S = \{(i,j) : \; i,j = 1,\ldots,6\}$.
- $\mathrm{Range}(X,Y) = \{(x,y) : \; 1 \le x \le y \le 6\}$.

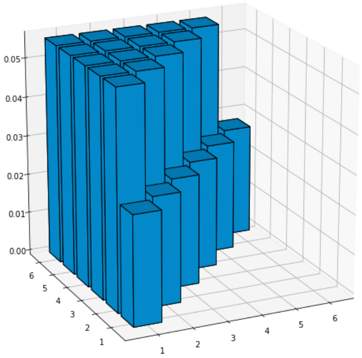


- For $1 \le x \le y \le 6$,

$$f(x,y) = \begin{cases} \frac{1}{36} & x = y \\ \frac{2}{36} & x \neq y \end{cases}.$$

3D histogram

In the previous dice roll example:



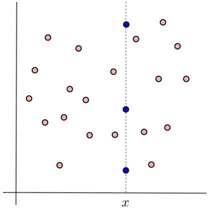
Marginal pmf

Definition

For any $x \in \mathrm{Range}(X)$, let

$$f_X(x) = \sum_{y:\ (x,y) \in \mathrm{Range}(X,Y)} f(x,y).$$

$f_X(x)$ is called the **marginal pmf** of X .

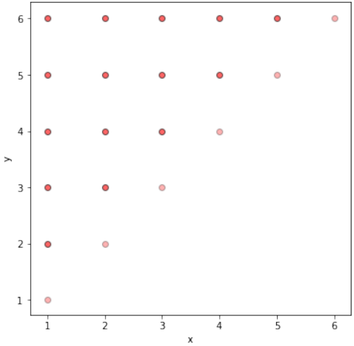


- Marginal pmf of Y is defined similarly.
- Notice that,

$$\boxed{f_X(x) = P(X = x).}$$

In the dice roll example

- $f_X(x) = \sum_{y=x}^{6} f(x,y) = \frac{1}{36} + \frac{2(6-x)}{36}$
- $f_Y(y) = \sum_{x=y}^{6} f(x,y) = \frac{1}{36} + \frac{2(y-1)}{36}$



Theorem

If $f(x,y)$ is a joint pmf then

1. $f(x,y) \geq 0$,
2. $\sum_{(x,y)\in \mathrm{Range}(X,Y)} f(x,y) = 1$.

Proof.

1. $f(x,y)$ is a probability and so is non-negative.
2. Notice that

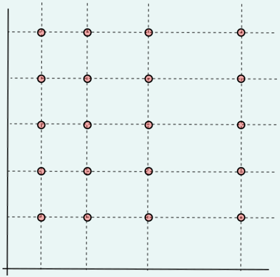
$$\begin{aligned}\sum_{(x,y)\in \mathrm{Range}(X,Y)} f(x,y) &= \sum_{x\in \mathrm{Range}(X)} \sum_{y\colon (x,y)\in \mathrm{Range}(X,Y)} f(x,y) \\ &= \sum_{x\in \mathrm{Range}(X)} f_X(x) \\ &= 1.\end{aligned}$$

□

We say that a set $A \subset \mathbb{R}^2$ is rectangular if there exists subsets $B, C \subset \mathbb{R}$ such that

$$A = \{(x, y) : \ x \in B, y \in C\}.$$

Or equivalently, if $(x_1, y_1), (x_1, y_1) \in A$ then also $(x_1, y_2), (x_2, y_1) \in A$.



Definition (Independence)

Random variables X and Y are called **independent** if, for any $x, y \in \mathbb{R}^2$,

$$\boxed{f(x,y) = f_X(x) \cdot f_Y(y)} \quad \text{for any } (x,y) \in \mathbb{R}^2.$$

Otherwise they are called **dependent**.

- In terms of probabilities this is equivalent to, for any $x, y \in \mathrm{Range}(X, Y)$

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y).$$

- If X, Y are independent then $\mathrm{Range}(X, Y)$ (rather the set $\{(x,y) : \ f(x,y) > 0\}$) is rectangular:

$$\mathrm{Range}(X, Y) = \{(x,y) : \ x \in \mathrm{Range}(X), y \in \mathrm{Range}(Y)\}.$$

- In the dice roll example X and Y are not independent (the range is not rectangular).
- Typically, dependence of two random variables is easier to show than independence.

Exercise 1

Problem (4.1-3 in the textbook)

Let the joint pmf of X and Y be defined by $f(x, y) = \frac{x+y}{32}$, $x = 1, 2, y = 1, 2, 3, 4.$

(a) *Find $f_X(x)$, the marginal pmf of X.*

(b) *Find $f_Y(y)$, the marginal pmf of Y.*

(c) *Find $P(X > Y)$.*

(d) *Find $P(Y = 2X)$.*

(e) *Find $P(X + Y = 3)$.*

(f) *Find $P(X \leq 3 - Y)$.*

(g) *Are X and Y independent or dependent? Why or why not?*

(h) *Find the means and the variances of X and Y.*

Solution

(a) $f_X(x) = \sum_{y=1}^{4} \frac{x+y}{32} = \frac{x+1}{32} + \frac{x+2}{32} + \frac{x+3}{32} + \frac{x+4}{32} = \boxed{\dfrac{4x+10}{32}}$

(b) $f_Y(y) = \sum_{x=1}^{2} \frac{x+y}{32} = \frac{1+y}{32} + \frac{2+y}{32} = \boxed{\dfrac{3+2y}{32}}$

Solution (cont.)

(c) $P(X > Y) = f(2,1) = \boxed{\dfrac{3}{32}}$

(d) $P(Y = 2X) = f(1,2) + f(2,4) = \frac{3}{32} + \frac{6}{32} = \boxed{\dfrac{9}{32}}$

(e) $P(X + Y = 3) = f(1,2) + f(2,1) = \frac{3}{32} + \frac{3}{32} = \boxed{\dfrac{6}{32}}$

(f) $P(X \le 3 - Y) = P(X + Y \le 3) = f(1,1) + f(1,2) + f(2,1) = \frac{2}{32} + \frac{3}{32} + \frac{3}{32} = \boxed{\dfrac{8}{32}}$

(g) *The range is rectangular. But*

$$f(1,1) = \frac{2}{32} \neq f_X(1) f_Y(1) = \frac{14}{32}\frac{5}{32}.$$

(h)

$$E_X[X] = \sum_{x \in \mathrm{Range}(X)} x f_X(x) = 1 \cdot \frac{14}{32} + 2 \cdot \frac{18}{32} = \boxed{\frac{50}{32}}.$$

$$E_Y[Y] = 1 \cdot \frac{3+2}{32} + 2 \cdot \frac{3+4}{32} + 3 \cdot \frac{3+6}{32} + 4 \cdot \frac{3+8}{32} = \frac{5 + 14 + 27 + 44}{32} = \boxed{\frac{90}{32}}.$$

- Let $u : \mathbb{R}^2 \to \mathbb{R}$ be any function.
- Then $Z = u(X, Y)$ is a random variable on S.

Theorem

Let X, Y be discrete random variables and $u : \mathbb{R}^2 \to \mathbb{R}$ be a function. If

$$\sum_{(x,y)\in \mathrm{Range}(X,Y)} |u(x, y)| f(x, y) < \infty.$$

Then the expected value of the random variable $Z = u(X, Y)$ exists and

$$E[Z] = \sum_{(x,y)\in \mathrm{Range}(X,Y)} u(x, y) f(x, y).$$

- In the previous exercise, $u(X, Y) = X$ and

$$\mu_X = E[X] = \sum_{(x,y)\in \mathrm{Range}(X,Y)} x f(x, y).$$

In the dice roll example

$$\mu_X = \sum_{x=1}^{6} \sum_{y=x}^{6} x f(x,y) = \sum_{x=1}^{6} x f_X(x) = \sum_{x=1}^{6} x \left(\frac{1}{36} + \frac{2(6-x)}{36} \right) = 2.84375.$$

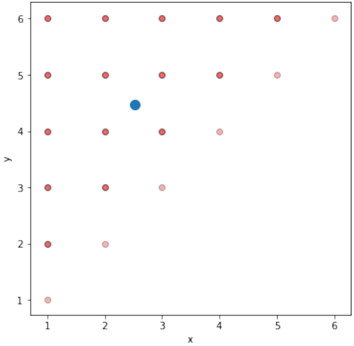
Similarly,

$$\mu_Y = \sum_{y=1}^{6} \sum_{x=y}^{6} x f(x,y) = \sum_{y=1}^{6} y f_Y(y) = \sum_{y=1}^{6} y \left(\frac{1}{36} + \frac{2(y-1)}{36} \right) = 5.03125.$$

If we denote

$$
\begin{aligned}
g(a,b) &= E[|X-b|^2 + |Y-a|^2] \\
&= E[|X-b|^2] + E[|Y-a|^2] \\
&= E_X[|x-b|^2] + E_Y[|y-a|^2].
\end{aligned}
$$

- The minimum of $g(a,b)$ is attained when $E_X[|x-b|^2]$ and $E_Y[|y-a|^2]$ are minimized.
- Minimum at (μ_X, μ_Y).
- (μ_X, μ_Y) is the center of the histogram.



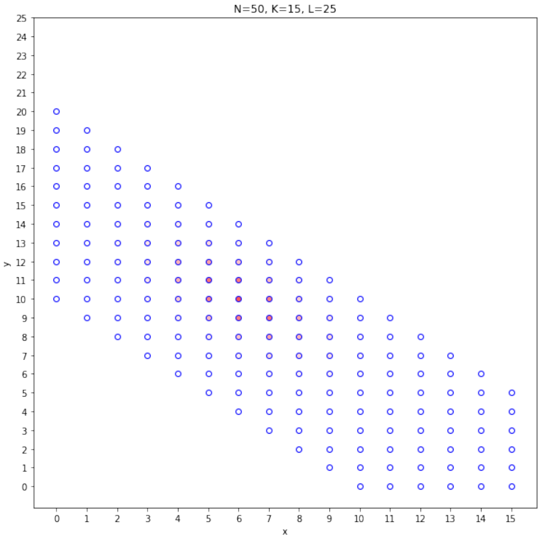
Bivariate hypergeometric distribution

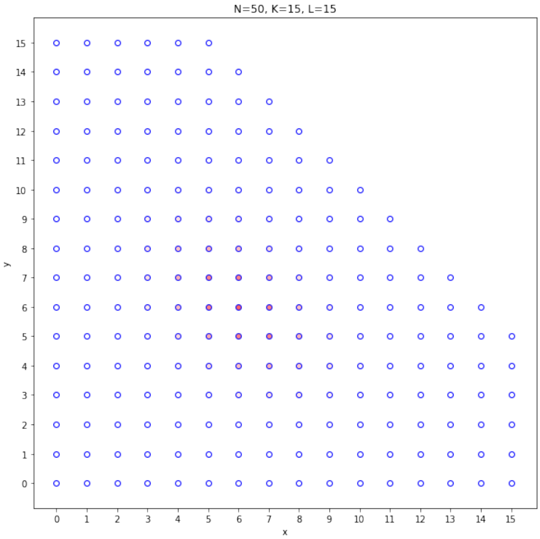
- Suppose there are N balls in an urn, K are red, L are blue and the rest $N-K-L$ are orange.
- S is the set of n balls selected without order and without replacement.
- For $s \in S$, let $X(s)$ be the number of red balls in s and $Y(s)$ be the number of blue balls in s.

Definition (Bivariate hypergeometric distribution)

The joint pmf of above X, Y is called **bivariate hypergeometric distribution** with parameters (N, K, L, n).

- $\mathrm{Range}(X,Y) = \{(x,y) :\ 0 \le x \le K,\ 0 \le y \le L,\ 0 \le n-x-y \le N-K-L\}$.
- $f(x,y) = \dfrac{\binom{K}{x}\binom{L}{y}\binom{N-K-L}{n-x-y}}{\binom{N}{n}}$.
- $f_X(x) = \dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$.
- X, Y are dependent (range is not rectangular).


N=50, K=15, L=25
y
x
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25


N=50, K=15, L=15
x
y
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

Trinomial distribution

- The experiment terminates in three ways: **Success, Inconclusive, Failure**.
- p_S, p_I, p_F are corresponding probabilities (Note: $p_F = 1 - p_S - p_I$).
- We have n trials of the experiment.
- $X(s)$ is the the number of successes in n-trials.
- $Y(s)$ is the number of inconclusive in n trials.
- $\mathrm{Range}(X, Y) = \{(x, y) : \; x + y \leq n\}$.

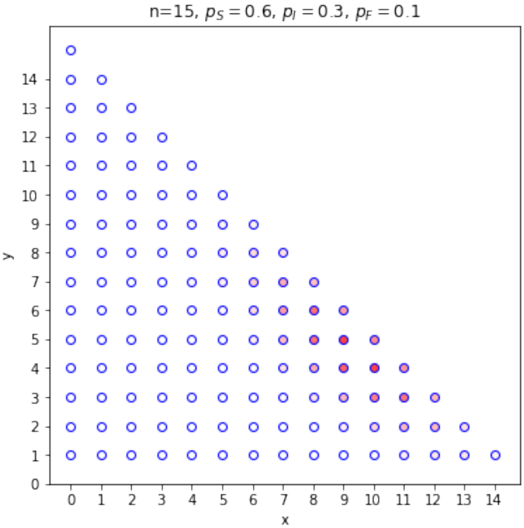
Definition

The joint distribution of (X, Y) is called **trinomial distribution** with parameters (n, p_S, p_I).

- Can be seen that

$$f(x, y) = P(X = x, Y = y) = \frac{n!}{x!y!(n-x-y)!} p_S^x p_I^y (1 - p_S - p_I)^{n-x-y}.$$

- X and Y are not independent because $\mathrm{Range}(X, Y)$ is not rectangular.
- $f_X(x) = \binom{n}{x} p_S^x (1 - p_S)^{n-x}$.


n=15, $p_S = 0.6$, $p_I = 0.3$, $p_F = 0.1$
y
x
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14

Exercise 2

Problem (4.1-9 from the textbook)

A manufactured item is classified as good, a "second," or defective with probabilities 6/10, 3/10, and 1/10, respectively. Fifteen such items are selected at random from the production line. Let X denote the number of good items, Y the number of seconds, and $15 - X - Y$ the number of defective items.

(a) *Give the joint pmf of X and Y, $f(x, y)$.*

(b) *Sketch the set of integers (x, y) for which $f(x, y) > 0$. From the shape of this region, can X and Y be independent? Why or why not?*

(c) *Find $P(X = 10, Y = 4)$.*

(d) *Give the marginal pmf of X.*

(e) *Find $P(X \leq 11)$.*

Solution

(X, Y) *has trinomial distribution with parameters* $n = 15$, $p_S = 0.6, p_I = 0.3, p_F = 0.1$.

(a) $f(x, y) = \frac{15!}{x!y!(15-x-y)!} \cdot 0.6^x \cdot 0.3^y \cdot 0.1^{15-x-y}$.

Solution (cont.)

(b) *See previous page for sketch. It is not independent because not rectangular.*

(c) $P(X = 10, Y = 4) = \frac{15!}{10!4!1!} \cdot 0.6^{10} \cdot 0.3^4 \cdot 0.1 \boxed{\approx 0.0735}$.

(d) *We noticed already that* $\boxed{f_X(x) = \binom{15}{x} \cdot 0.6^x \cdot 0.4^{15-x}}$.

(e) *Since* X *has binomial distribution with parameters* $(n = 15, p = 0.6)$ *we use the fact that*

$$P(X \le 11) = P(Z > 3) = 1 - P(Z \le 3)$$

where Z *is binomial with parameters* $n = 15, p = 0.4$.
By checking at the back of the book $\boxed{P(X \le 11) = 1 - 0.0905 = 0.9095}$.

Thursday class

Covariance

Definition (Covariance)

The expected value

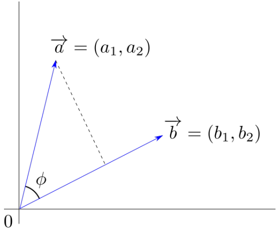
$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)],$$

if it exists, is called the **covariance** of two random variables X, Y. It is also denoted by σ_{XY}.

- Expectation is computed in terms of joint pmf of X, Y for the function

$$u(x, y) = (x - \mu_X)(y - \mu_Y).$$

- $\mathrm{Cov}(X, X) = E[(X - \mu_X)^2] = \mathrm{Var}(X)$.
- Geometrically, it is the weighted dot-product (projection of one on the other) of the centralized random variables $X - \mu_X$ and $Y - \mu_Y$.



$$\boxed{\|\overrightarrow{a}\|\|\overrightarrow{b}\|\cos\phi = a_1b_1 + a_2b_2}.$$

Example

- Let $X(s)$ be the height of student s in this class and $Y(s)$ be the weight of student s in the class of 60 students.
- Let $\mathrm{Range}(X,Y) = \{(x_1,y_1),\dots,(x_{60},y_{60})\}$.
- Assuming the (weight,height) combinations are unique so that $f(x_i,y_i) = \frac{1}{60}$ for every i.
- $\mathrm{Cov}(X,Y) = \sum_{i=1}^{60}(x_i-\mu_X)(y_i-\mu_Y)\frac{1}{60}$.
- $\mathrm{Cov}(X,Y)$ is the weighted inner product between vectors

$$\overrightarrow{a} = (x_1-\mu_X,\dots,x_{60}-\mu_X), \quad \overrightarrow{b} = (y_1-\mu_Y,\dots,y_{60}-\mu_Y).$$

Correlation

Definition (Correlation coefficient)

Let $\sigma_X, \sigma_Y > 0$. The quantity

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

is called the **correlation coefficient** of random variables X and Y.

- ρ measures the cosine of the "angle" between X and Y.
- As such, $-1 \leq \rho \leq 1$.

Definition

1. If $\rho > 0$ then X and Y are said to be **positively correlated**.
2. If $\rho < 0$ then X and Y are said to be **negatively correlated**.
3. If $\rho = 0$ then X and Y are said to be **uncorrelated**.

- $\mathrm{Cov}(X, Y)$ represents by how much changing the value of X affects the value of Y.
- ρ represents whether the change of X affects negatively or positively the value of Y.

Theorem

$$\operatorname{Cov}(X,Y) = E[XY] - E[X]E[Y].$$

Proof.

$$\begin{aligned}
\operatorname{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\
&= E[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\
&= E[XY] - E[\mu_Y X] - E[\mu_X Y] + E[\mu_X \mu_Y] \\
&= E[XY] - \mu_Y E[X] - \mu_X E[Y] + \mu_X \mu_Y \\
&= E[XY] - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\
&= E[XY] - \mu_Y \mu_X .
\end{aligned}$$

$\square$

Theorem

If X and Y are independent then X and Y are uncorrelated ($\rho = 0$).

Proof.

- Equivalent to $\mathrm{Cov}(X, Y) = 0$.
- X and Y are independent then $\mathrm{Range}(X, Y)$ is rectangular and $f(x, y) = f_X(x) f_Y(y)$.
- Notice that

$$
\begin{aligned}
E[XY] &= \sum_{(x,y)\in \mathrm{Range}(X,Y)} xyf(x, y) \\
&= \sum_{x\in \mathrm{Range}(X), y\in \mathrm{Range}(Y)} xyf_X(x)f_Y(y) \\
&= \sum_{x\in \mathrm{Range}(X)} \sum_{y\in \mathrm{Range}(Y)} xyf_X(x)f_Y(y) \\
&= \sum_{x\in \mathrm{Range}(X)} xf_X(x) \sum_{y\in \mathrm{Range}(Y)} yf_Y(y) \\
&= E[X]E[Y].
\end{aligned}
$$

- From previous theorem

$$\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0.$$

$$\boxed{\text{Independent} \Rightarrow \text{uncorrelated}}$$

$$\boxed{\text{Uncorrelated} \not\Rightarrow \text{independent}}$$

Example

- Let
$$\text{Range}(X,Y) = \{(0,1),(1,0),(2,1)\}.$$
- $f(x,y) = \frac{1}{3}$.
- Not independent as $\text{Range}(X,Y)$ is not rectangular.
- $\mu_X = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = 1$.
- $\mu_Y = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}$.
- $\text{Cov}(X,Y) = E[XY] - \mu_X\mu_Y = 0 \cdot 1 \cdot \frac{1}{3} + 1 \cdot 0 \cdot \frac{1}{3} + 2 \cdot 1 \cdot \frac{1}{3} - 1 \cdot \frac{2}{3} = 0$.

The smaller the $|\rho|$ is the weaker the relationship between X and Y is.

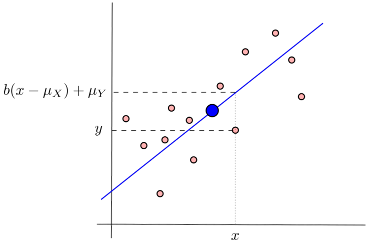
- We mentioned that the (μ_X, μ_Y) is the center of the histogram in a sense.
- Any line that passes though the center has the formula

$$y = b(x - \mu_X) + \mu_Y$$

  for some $b \in \mathbb{R}$.
- The quantity

$$g(b) = E[|Y - b(X - \mu_X) - \mu_Y|^2]$$

  measures the average distance of the range to the line.
- By minimizing $g(b)$, we are looking at the line passing through the center that has the "closest" fit to the (X, Y) distribution.



Theorem

The minimum of the functional

$$g(b) = E[|Y - b(X - \mu_X) - \mu_Y|^2]$$

is attained at

$$b = \rho \frac{\sigma_Y}{\sigma_X}.$$

Proof.

$$\begin{aligned}
g(b) &= E[(Y - \mu_Y - b(X - \mu_X))^2] \\
&= E[(Y - \mu_Y)^2 - 2b(X - \mu_X)(Y - \mu_Y) + b^2(X - \mu_X)^2] \\
&= E[(Y - \mu_Y)^2] - 2b\, E[(X - \mu_X)(Y - \mu_Y)] + b^2 E[(X - \mu_X)^2] \\
&= \sigma_Y^2 - 2b\, \mathrm{Cov}(X, Y) + b^2 \sigma_X^2.
\end{aligned}$$

The minimum is attained when

$$g'(b) = -2\mathrm{Cov}(X, Y) + 2b\sigma_X^2 = 0.$$

Hence

$$b = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X}.$$

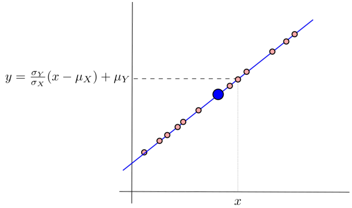
□

- At the minimum value,

$$
\begin{aligned}
g\left(\rho\frac{\sigma_Y}{\sigma_X}\right) &= \sigma_Y^2 - 2\rho\frac{\sigma_Y}{\sigma_X}\,\mathrm{Cov}(X,Y) + \left(\rho\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 \\
&= \sigma_Y^2 - 2\rho^2\sigma_Y^2 + \rho^2\sigma_Y^2 \\
&= \sigma_Y^2(1-\rho^2).
\end{aligned}
$$

- We again conclude $1-\rho^2 \geq 0$ or $-1 \leq \rho \leq 1$.
- $|\rho| = 1$ if and only if X and Y are perfectly correlated:

$$
Y = \pm\frac{\sigma_Y}{\sigma_X}(x-\mu_X)+\mu_Y.
$$



- **The larger the $|\rho|$, the more linear the relationship between X and Y is.**

Linear least squares regression

Definition

The line

$$y = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}(x-\mu_X)+\mu_Y = \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)+\mu_Y$$

is called the **least squares regression line**.

Theorem (Linear least squares regression problem)

The least squares regression line is the line that minimizes the function

$$g(a,b) = E[|Y-aX-b|^2].$$

- This is stronger statement than what we already proved.
- We are looking at the best fit among all possible lines, not just the ones that pass trough the center.
- Will be left as homework (Problem 4.2-5 in the textbook).

Correlation vs causation

- When X and Y are positively correlated, the increase in X corresponds to the increase in Y when ρ is closer to 1.
- When X and Y are negatively correlated, the increase in X corresponds to the decrease in Y when ρ is close to -1.
- This leads to the "correlation implies causation" logical fallacy.

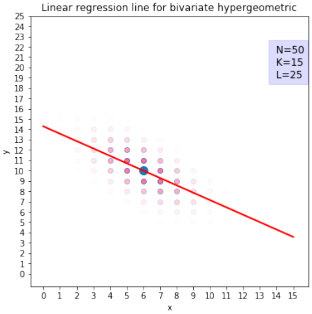
Example

People with higher height tend to have higher weight. That means height and weight are positively correlated. High height is not the cause of high weight.

Correlation does NOT imply causation

Least squares regression line for bivariate hypergeometric

- $\mu_X = \frac{nK}{N}$, $\mu_Y = \frac{nL}{N}$,
- $\sigma_X^2 = n\frac{K}{N}\left(1 - \frac{K}{N}\right)\frac{N-n}{N-1}$,
- $\sigma_Y^2 = n\frac{K}{N}\left(1 - \frac{L}{N}\right)\frac{N-n}{N-1}$,
- $\mathrm{Cov}(X, Y) = -n\frac{N-n}{N-1}\frac{KL}{N^2}$ (we didn't prove this but take as given)



The color intensity is proportional to the joint pmf value.