Mathematical Statistics and Data Analysis: Lection 6. Joint distributions and variables correlation

Aleksei Zhukov

Data Scientist, Research Fellow

BI BRICS,
2020

Outline

Definition

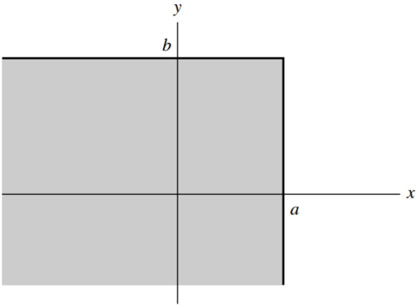
Sometimes we need to analyze a joint probability structure of two or more random variables defined on the same sample space. Joint distributions arise naturally in many applications:

- The joint probability distribution of the some socio-economical factors (like income and health).
- The joint distribution of the values of various physiological variables in a population of patients is often of interest in medical studies
- The joint probability distribution of the x, y, and z components of wind velocity can be experimentally measured in studies of atmospheric turbulence.

The joint behavior of two random variables, X and Y, is determined by the cumulative distribution function

$$F(x, y) = P(X \leq x, Y \leq y)$$

regardless of whether X and Y are continuous or discrete. The cdf gives the probability that the point (X, Y) belongs to a semi-infinite rectangle in the plane, as shown.



$F(a, b)$ gives the probability of the shaded rectangle.

Suppose that X and Y are discrete random variables defined on the same sample space and that they take on values $x_1, x_2, \ldots,$ and $y_1, y_2, \ldots,$ respectively. Their **joint frequency function,** or joint probability mass function, $p(x, y)$, is

$$p(x_i, y_j) = P(X = x_i, Y = y_j)$$

A simple example will illustrate this concept. A fair coin is tossed three times; let X denote the number of heads on the first toss and Y the total number of heads. From the sample space, which is

$$\Omega = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}$$

we see that the joint frequency function of X and Y is as given in the following table:

	y			
x	0	1	2	3
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$

Thus, for example, $p(0, 2) = P(X = 0, Y = 2) = \frac{1}{8}$. Note that the probabilities in the preceding table sum to 1.

Suppose that we wish to find the frequency function of Y from the joint frequency function. This is straightforward:

$$
\begin{aligned}
p_Y(0) &= P(Y = 0) \\
&= P(Y = 0, X = 0) + P(Y = 0, X = 1) \\
&= \frac{1}{8} + 0 \\
&= \frac{1}{8} \\
p_Y(1) &= P(Y = 1) \\
&= P(Y = 1, X = 0) + P(Y = 1, X = 1) \\
&= \frac{3}{8}
\end{aligned}
$$

In general, to find the frequency function of Y, we simply sum down the appropriate column of the table. For this reason, p_Y is called the **marginal frequency function** of Y. Similarly, summing across the rows gives

$$p_X(x) = \sum_i p(x, y_i)$$

which is the marginal frequency function of X.

The case for several random variables is analogous. If $X_1, \ldots, X_m$ are discrete random variables defined on the same sample space, their joint frequency function is

$$p(x_1, \ldots, x_m) = P(X_1 = x_1, \ldots, X_m = x_m)$$

The marginal frequency function of X_1, for example, is

$$p_{X_1}(x_1) = \sum_{x_2 \cdots x_m} p(x_1, x_2, \ldots, x_m)$$

The two-dimensional marginal frequency function of X_1 and X_2, for example, is

$$p_{X_1 X_2}(x_1, x_2) = \sum_{x_3 \cdots x_m} p(x_1, x_2, \ldots, x_m)$$

Suppose that X and Y are continuous random variables with a joint cdf, $F(x, y)$. Their **joint density function** is a piecewise continuous function of two variables, $f(x, y)$. The density function $f(x, y)$ is nonnegative and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dy\, dx = 1$. For any "reasonable" two-dimensional set A

$$P((X, Y) \in A) = \iint\limits_{A} f(x, y)\, dy\, dx$$

In particular, if $A = \{(X, Y) | X \leq x \text{ and } Y \leq y\}$,

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, dv\, du$$

Thus, the probability that (X, Y) is in a small neighborhood of (x, y) is proportional to $f(x, y)$. Differential notation is sometimes useful:

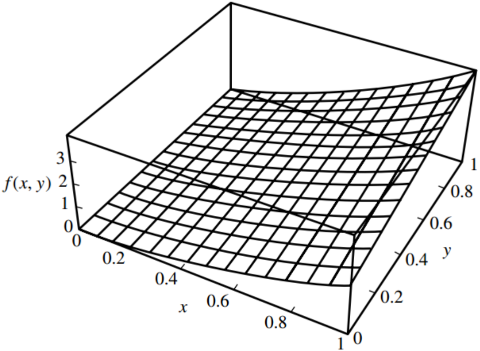
$$P(x \leq X \leq x + dx, y \leq Y \leq y + dy) = f(x, y)\, dx\, dy$$

Consider the bivariate density function

$$f(x, y) = \frac{12}{7}(x^2 + xy), \qquad 0 \le x \le 1, \qquad 0 \le y \le 1$$

which is plotted in Figure. $P(X > Y)$ can be found by integrating f over the set $\{(x, y)|0 \le y \le x \le 1\}$:

$$P(X > Y) = \frac{12}{7}\int_0^1 \int_0^x (x^2 + xy)\, dy\, dx = \frac{9}{14}$$



The **marginal cdf** of X, or F_X, is

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \lim_{y \to \infty} F(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u, y)\, dy\, du \end{aligned}$$

From this, it follows that the density function of X alone, known as the **marginal density** of X, is

$$f_X(x) = F'_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$$

In the discrete case, the marginal frequency function was found by summing the joint frequency function over the other variable; in the continuous case, it is found by integration.

Continuing previous example, the marginal density of X is

$$\begin{aligned} f_X(x) &= \frac{12}{7} \int_{0}^{1} (x^2 + xy)\, dy \\ &= \frac{12}{7} \left(x^2 + \frac{x}{2} \right) \end{aligned}$$

A similar calculation shows that the marginal density of Y is $f_Y(y) = \frac{12}{7}(\frac{1}{3} + y/2)$.

The variance of a random variable is a measure of its variability, and the covariance of two random variables is a measure of their joint variability, or their degree of association.

DEFINITION

If X and Y are jointly distributed random variables with expectations μ_X and μ_Y, respectively, the covariance of X and Y is

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

provided that the expectation exists. ■

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

The covariance is the average value of the product of the deviation of X from its mean and the deviation of Y from its mean. If the random variables are positively associated – that is, when X is larger than its mean, Y tends to be larger than its mean as well – the covariance will be positive. If the association is negative – that is, when X is larger than its mean, Y tends to be smaller than its mean – the covariance is negative.

By expanding the product and using the linearity of the expectation, we obtain an alternative expression for the covariance

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

In particular, if X and Y are independent, then $E(XY) = E(X)E(Y)$ and $\operatorname{Cov}(X,\ Y) = 0$ (but the converse is not true).

The **correlation coefficient** is defined in terms of the covariance.

DEFINITION

If X and Y are jointly distributed random variables and the variances and covariances of both X and Y exist and the variances are nonzero, then the correlation of X and Y, denoted by ρ, is

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \qquad \blacksquare$$

Investment Portfolio

We first consider the simple example of two securities, assuming that they have the same expected returns $\mu_1 = \mu_2 = \mu$ and their returns are uncorrelated: $\sigma_{ij} = \mathrm{Cov}(R_i, R_j) = 0$. For a portfolio $(\pi, 1 - \pi)$, the expected return is

$$E(R(\pi)) = \pi\mu + (1 - \pi)\mu = \mu$$

so that when considering expected return only, the choice of π makes no difference. However, taking risk into account,

$$\mathrm{Var}(R(\pi)) = \pi^2\sigma_1^2 + (1 - \pi)^2\sigma_2^2.$$

Minimizing this with respect to π gives the optimal portfolio

$$\pi_{\text{opt}} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

For example, if the investments are equally risky, $\sigma_1 = \sigma_2 = \sigma$, then $\pi = 1/2$, so the best strategy is to split total investment equally between the two securities.

$$\mathrm{Var}\left(R\left(\tfrac{1}{2}\right)\right) = \frac{\sigma^2}{2}$$

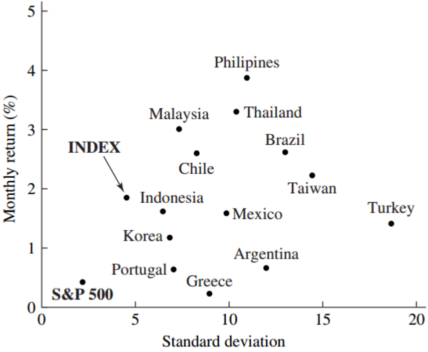
whereas if we put all her money in one security, the variance of the return would be σ^2. The expected return is the same in both cases. This is a particularly simple example of the value of diversification of investments.

Suppose now that the two securities do not have the same expected returns, $\mu_1 < \mu_2$. Let the standard deviations of the returns be σ_1 and σ_2; usually less risky investments have lower expected returns, $\sigma_1 < \sigma_2$. Furthermore, the two returns may be correlated: $\mathrm{Cov}(R_1, R_2) = \rho\sigma_1\sigma_2$. Corresponding to the portfolio $(\pi, 1-\pi)$, we have expected return

$$E(R(\pi)) = \pi\mu_1 + (1-\pi)\mu_2$$

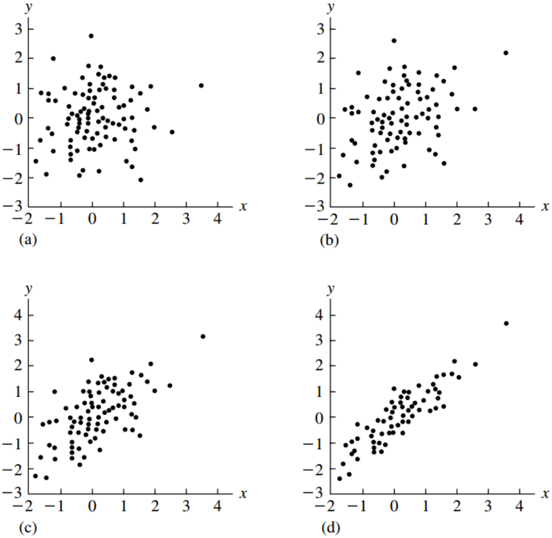
and the variance of the return is

$$\mathrm{Var}(R(\pi)) = \pi^2\sigma_1^2 + 2\pi(1-\pi)\rho\sigma_1\sigma_2 + (1-\pi)^2\sigma_2^2$$



As a general rule, risk is reduced by diversification and can be decreased with only a small sacrifice of returns. Bernstein (1996, p. 254) illustrates this point empirically.

The point labeled "Index" shows the monthly average versus standard deviation for an investment that was equally weighted across all the markets. A reasonably high return with relatively little risk would thus have been obtained by spreading investments equally over the 13 stock markets. In fact, the risk is less than that of any of the individual markets. Note that these emerging markets were riskier than the U.S. market, but that they were more profitable.



Scatterplots of 100 independent pairs of bivariate normal random variables, (a) $\rho = 0$, (b) $\rho = .3$, (c) $\rho = .6$, (d) $\rho = .9$.

Statistical hypothesis

There are methods called statistical tests. Such a methods are usually applied to check if random variables is related to particular distribution or random varible independence.

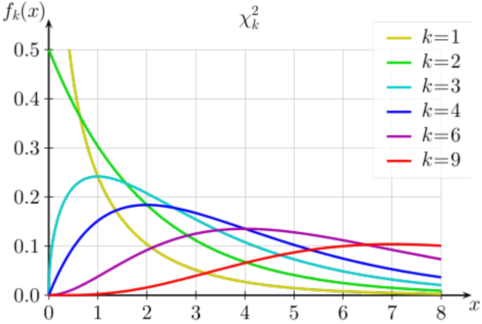
1. Define the problem, a null hypothesis and alternative hypothesis
2. Define propper test and its parameters (degree of freedom in case of χ^2)
3. Define α - hypothesis rejection threshold
4. Propose test, approve or reject null hypothesis

One of the most widely used test is χ^2.

1. independence of random variables
2. goodness-of-fit (quality of the real data representation by model distribution)

Appliable to the binomial, Poisson or normal distributed variables.

If a random variable Z has the standard normal distribution ($Z \sim \mathbb{N}(0,1)$), then Z^2 has the χ^2 distribution with one degree of freedom.



if $Z_1 + Z_2 + \ldots + Z_n$ are independent standard random variables, χ^2_n is a χ^2 distribution with n degree of freedom

$$\chi^2_n = Z_1 + Z_2 + \ldots + Z_n$$

We can get χ^2 test significance value (p-value)

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$$\textit{p-value}(\chi^2) = 1 - CDF(\chi^2)$$

Actually p-value is a quantile of the distribution.
See examples on the practical lesson.

Big test is coming! Be prepared!

Brilliant Statistics explanation here (god bless VPN)
https://www.youtube.com/user/jbstatistics/videos

Todays topics

1. Joint distributions
2. Covariance and Correlation
3. Statistical hypothesis and statistical tests
4. Conclusions

Questions

Time for your questions!

zhukovalex13@gmail.com