Basic Concepts of Correlation

Definition 1: The **covariance** between two sample random variables x and y is a measure of the linear association between the two variables, and is defined by the formula

$$cov(x, y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \Big/ (n-1)$$

vation: The covariance is similar to the variance, except that the covariance is l for two variables (x and y above) whereas the variance is defined for only one e. In fact, $cov(x, x) = var(x)$.

Real Statistics Resources

Free Download

Follow @Real1Statistics

Current Section

- Correlation
  - Basic Concepts
    - Advanced
  - Scatter Diagrams
  - One Sample Testing
  - Two Sample Testing
  - Multiple Correlation
  - Spearman's Rho

variance can be thought of as the sum of matches and mismatches among the pairs of data elements for x and y: a match occurs when both elements in the pair are on the same side of their mean; a mismatch occurs when one element in the pair is above its mean and the other is below its mean.

The covariance is positive when the matches outweigh the mismatches and is negative when the mismatches outweigh the matches. The size of the covariance in absolute value indicates the intensity of the linear relationship between x and y: the stronger the linear relationship the larger the value of the covariance will be. The size of the covariance is also influenced by the scale of the data elements, and so in order to eliminate the scale factor the correlation coefficient is used as the scale-free metric of linear relationship.

Definition 2: The **correlation coefficient** between two sample variables x and y is a scale-free measure of linear association between the two variables, and is given by the formula

$$r = cov(x,y)/s_x s_y$$

If necessary we can write r as r_{xy} to explicitly show the two variables.

We also use the term **coefficient of determination** for r^2

Observation: Just as we saw for the variance in Measures of Variability, the covariance can be calculated as

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}(\sum_{i=1}^{n}x_i y_i - \bar{x}\sum_{i=1}^{n}y_i - \bar{y}\sum_{i=1}^{n}x_i + n\bar{x}\bar{y}) = \frac{1}{n-1}(\sum_{i=1}^{n}x_i y_i - n\bar{x}\bar{y})$$

As a result, we can also calculate the correlation coefficient as

Charles Zaiontz

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

Property 1:

$$-1 \leq r \leq 1$$

Observation: If r is close to 1 then x and y are positively correlated. A **positive linear correlation** means that high values of x are associated with high values of y and low values of x are associated with low values of y.

If r is close to -1 then x and y are negatively correlated. A **negative linear correlation** means that high values of x are associated with low values of y, and low values of x are associated with high values of y.

When r is close to 0 there is little linear relationship between x and y.

Observation: We have defined covariance and the correlation coefficient for data samples. We can also define covariance and correlation coefficient for populations, based on their pdf.

Definition 3: The **covariance** between two random variables x and y for a population with discrete or continuous pdf is

$$cov(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

Definition 4: The (**Pearson's product moment**) **correlation coefficient** for two variables x and y for a population with discrete or continuous pdf is

$$\rho = cov(x,y)/\sigma_x\sigma_y$$

Property 2:

$$-1 \leq \rho \leq 1$$

Property 3:

$$cov(x,y) = E[xy] - \mu_x\mu_y$$

Property 4: The following is true for both for the sample and population definitions of covariance:

If x and y are independent then $cov(x, y) = 0$

Property 5: The following are true both for samples and populations:

$$var(x+y) = var(x) + var(y) + 2cov(x,y)$$
$$var(x-y) = var(x) + var(y) - 2cov(x,y)$$

Observation: Click here for additional properties of covariance and correlation, as well as the proofs of the properties given above.

Observation: It turns out that r is not an unbiased estimate of ρ. A relatively unbiased estimate of ρ^2 is given by the **adjusted coefficient of determination** r^2_{adj}:

$$r^2_{adj} = 1 - \frac{(1-r^2)(n-1)}{n-2}$$

While r^2_{adj} is a better estimate of the population coefficient of determination, especially for small values of n, for large values of n it is easy to see that $r^2_{adj} \approx r^2$. Note too that $r^2_{adj} \leq r^2$,

and while r^2_{adj} can be negative, this is relatively rare.

An even more unbiased estimate of the population correlation coefficient associated with normally distributed data is given by

$$\rho_{est} = r\left[1 + \frac{(1 - r^2)}{2(n - 3)}\right]$$

Excel Functions: Excel provides the following functions regarding the covariance and correlation coefficient:

COVAR(R1, R2) = the population covariance between the data in arrays R1 and R2. If R1 contains data $\{x_1,\ldots,x_n\}$, R2 contains $\{y_1,\ldots,y_n\}$, $\bar{x}$ = AVERAGE(R1) and $\bar{y}$ = AVERAGE(R2), then COVAR(R1, R2) has the value

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/n$$

This is the same as the formula given in Definition 1, with n replaced by $n - 1$. Excel doesn't have a sample version of the covariance, although this can be calculated using the formula:

$$n * \text{COVAR(R1, R2)} / (n - 1)$$

CORREL(R1, R2) = the correlation coefficient of data in arrays R1 and R2. This function can be used for both the sample and population versions of the correlation coefficient. Note that:

CORREL(R1, R2) = COVAR(R1. R2) / (STDEVP(R1) * STDEVP(R2)) = the population version of the correlation coefficient

CORREL(R1, R2) = n * COVAR(R1. R2) / (STDEV(R1) * STDEV(R2) * (n − 1)) = the sample version of the correlation coefficient

Excel also provides the following, less useful, functions:

PEARSON(R1, R2) = CORREL(R1, R2)

RSQ(R1, R2) = CORREL(R1, R2) ^ 2

Excel 2010/2013 also provide **COVARIANCE.S**(R1, R2) to compute the sample covariance as well as **COVARIANCE.P**(R1, R2) which is equivalent to COVAR(R1, R2). Also, the Real Statistics supplemental functions **COVARP**(R1, R2) and **COVARS**(R1, R2) compute the population and sample covariances respectively.

Finally there is a **Correlation** data analysis tool which we demonstrate in the Example 1 of Multiple Correlation.

Real Statistics Functions: The Real Statistics Resource Pack contains the following functions:

RSQ_ADJ(R1, R2) = adjusted coefficient of determination r^2_{adj} for the data sets contained in ranges R1 and R2.

CORREL_ADJ (R1, R2) = estimated correlation coefficient ρ_{est} for the data sets contained in ranges R1 and R2.

RSQ_ADJ(r, n) = adjusted coefficient of determination r^2_{adj} corresponding to the sample

CORREL_ADJ (r, n) = estimated correlation coefficient ρ_{est} corresponding to a sample correlation coefficient for a sample of size n.

45 Responses to *Basic Concepts of Correlation*

AQ *says:*

February 27, 2016 at 6:38 pm

Dear Charles,

I am conducting a study which measures the relationship between three variables; quality of life, medication adherence and healthcare satisfaction. Research suggests that all three variables directly affect one another (a triangular-shaped relationship). I am wondering what a relationship between three variables is called?

Many thanks

Reply

Charles *says:*

February 27, 2016 at 6:43 pm

AQ,

Of course it depends on the relationship that you are referring to, but probably you are looking for “they are correlated” or “they have an association”.
Charles

Reply

Martin *says:*

January 26, 2016 at 9:35 pm

Hello Charles,

I have a quick question regarding ANOVA and correlation factor, I am trying to analyze different experiments to test treatments. I get inconsistent and not high enough correlation factors to prove linear relationship among the variables. I also ran a one way anova and I get a p value of 0.000 using minitab.

Can we only use ANOVA when there is a linear relationship? or can I trust my results and if so what can I determine from them?

I would appreciate any help, thanks!

Reply

Charles *says:*

January 27, 2016 at 9:23 am

Martin,
You should be able to run ANOVA without making a separate test for

correlation. This should come out of the ANOVA results anyway. If you are getting inconsistent results perhaps you have made an error in conducting one of the tests. Without better information about your scenario I am unable to comment further.
Charles

Reply

Nirosha *says:*
January 7, 2016 at 11:23 am

Dear Charles,
I am having 30 sample size and need to test relationship with individual age, education level with their perception towards several variables which measures using likert scale.(+1 strongly agree to -1 strongly disagree).
can I use pearson correlation test to measure correlation between two group of this sample:
for example my hypothesis will be:
educated officers have best choice of selecting best employee or
experiences of officers have positive relationship with best practices of officers etc.

I have data on age and education level as categorical data and perception as ranking data.

hope you can understand my issue

thanks
Nirosha

Reply

Charles *says:*

January 11, 2016 at 8:17 pm

Nirosha,

The more likert scales you have, the more accurate tests that are designed for continuous data. With 7 scales (e.g. strongly agree, fairly strongly agree, mildly agree, neutral, mildly disagree, fairly strongly disagree, strongly disagree), a continuous test should generally work fine. It is also common to use such a test with a 5-point scale, although there is more risk. Better yet would to assign any value between -1 and +1.

You can certainly use pearson's correlation to measure the associations that you have listed. You can also test whether these correlation coefficients are significantly different from zero. Provided the data is at reasonably normally distributed this is equivalent to conducting a t test. See the webpage Relationship between correlation and t test.

You have stated that you plan to compare two groups. You can also compare more than two groups. This is equivalent to running ANOVA.

Charles

Reply

jane Israel *says:*

December 27, 2015 at 8:04 pm

am hapi wt d work ur doing, pls I'm working on gender and socioeconomic status as correlates of students's academic achievement. pls what statistical tool should I use to analyze the data..tanx in advance

Reply

Charles *says:*

December 28, 2015 at 9:30 am

Jane, it really depends on what hypothesis you are trying to test. The Correlation sections describes a number of tests that you could use. See the webpage

Correlation

Charles

Reply

Eric W. *says:*

December 9, 2015 at 2:23 pm

I have a large data set. I am trying to determine the correlation a distance variable and a probability variable. The distance is in increments of 5 (there are 1000+ data points for each distance increment). Most of the probabilities are zero (~10%). If I run Excel Correl() on the complete data, there is very little correlation. If I run Correl() on the average probability for each distance, there is strong correlation. Am I using Correl() in some way that is violating the built in assumptions?

Reply

Charles *says:*

December 9, 2015 at 6:37 pm

Eric,
There are no assumptions for using the Correl function. There may be some assumptions when you test the correlation value.
Without seeing your data, I can't tell you why you are getting such different results.
Charles

Reply

Chiranjit Chakraborty *says:*

December 6, 2015 at 7:23 pm

Dear, Excuse me, I am very confused that How to find R1 and R2 please details inform me. thanks

Reply

Charles *says:*

December 7, 2015 at 9:07 am

R1 and R2 are two ranges on an Excel spreadsheet which contain the data for the two samples.
Charles

Reply

Steve Thomas *says:*

December 3, 2015 at 4:19 pm

Im trying to work out the Standard deviation on Excel, and some of the cells contain a (0) which results in the function returning with an error code.

Does anyone know how I can get around this?

Thanks,

Steve

Reply

Charles *says:*

December 4, 2015 at 12:17 am

Steve,
Just because some of the cells contain a zero shouldn't necessarily result in the standard deviation function STDEV.S returning an error.
If you send me an Excel file with your data, I can try to figure out where the problem is.
Charles

Reply

Akira *says:*

October 26, 2015 at 4:30 am

Greetings Charles,

I've more than 500 relationship (one to one function) to be study either they have a possible relationship or not. Therefore my first step is by using Correlation Coefficient to segregate the possible relationship or not before move to the next step.

My question, is that okay to used that method to find the possible relationship or there is another method of that more reliable to segregate those one to one into possible relationship group and vise versa.

Reply

Charles *says:*

October 26, 2015 at 5:34 am

Sorry, but you haven't provided enough information for me to answer your question.
Charles

Reply

Akira *says:*

October 26, 2015 at 7:09 am

For example, in crude assays there are hundred of parameter (properties, such as sulfur content, Specific Gravity, viscosity and etc).

This one to one function, for example Sulfur vs SPG or Sulfur vs Viscosity and so on. The study is to find any possible relationship

among the properties of crude oil.

Since the study doesn't really wide and only few people attempt to do the statistical analysis on whole crude oil, I have to start with random variables.

So, I wonder if I can segregate the possible relationship just by using coefficient of determination or there is another method that much better compared to R-square.

Sorry but thank you in advance.

Reply

Charles *says:*

November 3, 2015 at 4:32 pm

Sorry Akira, but I don't understand your question.
Charles

Reply

Alessio *says:*

September 23, 2015 at 1:07 pm

Hi, something is the matter with the radj formula: (i) it cannot give negative r coefficients (I guess one needs to add a "sign(r)" factor before the sqrt; (ii) the content of the square root can be negative. Eg. when N=4 all r between -0.58 and 0.58 produce imaginary sqrt.

Reply

Charles *says:*

September 26, 2015 at 9:32 am

Alessio,
You are correct. For this reason it is better to speak about the adjusted coefficient of determination (the square of the correlation coefficient). I have now changed the webpage to reflect this. Thanks very much for identifying this problem.
Charles

Reply

Emeka *says:*

August 6, 2015 at 12:36 am

Greetings Charles!
Weldone for this rich site. Pls can I run CORREL on two sets of data with different units. Eg. X has units in molecules/cm^2 while Y has units in molecules/cm^3.
Thanks in advance

Reply

Charles *says:*

August 6, 2015 at 7:55 am

Emeka,
Yes, the units won't matter. The correlation coefficient is independent of the units.

Charles

Reply

Ramin *says:*

June 8, 2015 at 12:37 pm

Hi!

I have an important correlation related question:

I am analyzing spiking Neurons.
I have 4 Island each contains 16 spiking neurons. Each neuron fires spikes randomly in a time frame of 250 us.

I want to find the correlation between this 4 islands, how can i do it ?

Reply

Charles *says:*

June 8, 2015 at 4:03 pm

Hi Ramon,
I suggest that you look at the Multiple Correlation – Advanced webpage.
Charles

Reply

Tara *says:*

June 4, 2015 at 10:51 am

Thanks a lot Charles. Now I can find my way better.

Reply

---

mohsen *says:*

May 16, 2015 at 4:32 pm

hi
may you please explain about correlation coeficient in multi variables i.e. y and x1,x2,x3,...
y=ax1+bx2+bx3+.... how to find a,b,c,... so that we attain best fitting.

Reply

Charles *says:*

May 17, 2015 at 9:14 am

This is what regression is all about. Please look at the Linear Regression and especially the Multiple Regression webpages.
Charles

Reply

Tanya *says:*

May 1, 2015 at 1:13 pm

Hello,

I'm using excel to do a quick correlation. I was reading through the variables. I'm trying to make a correlation between performance metrics (rating scale is 1-5) and Versant Exam scores (rating is 1-100). Would it matter if the scales are different when I do the correlation?

Reply

Charles *says:*

May 1, 2015 at 10:29 pm

Tanya,
You can calculate the correlation coefficient even if the scales are different.
Charles

Reply

Tara *says:*

April 29, 2015 at 2:55 pm

Hi
Please can you help about finding correlation coefficient between two dependent variables, each variable with four level but I want to find over all correlation between the two dependent variables without making regards for the levels which I do not know how to do it. Any insight will be helpful. Many thanks.

Reply

Charles *says:*

April 29, 2015 at 3:50 pm

Tara,
Sorry, but I don't understand your question. In particular I don't understand what levels you are referring to. Are these part of an ANOVA?
Charles

Reply

Tara *says:*

June 2, 2015 at 3:28 pm

Thanks Charles
Yes the levels are part of an ANOVA. I meant I want to find correlation between two dependent variables over the 4 levels that factor have. And My question is that can I use mean value for each level when I calculate correlation between ? if so I think I will be able to have correlation over that 4b levels of the two dependent variables.

Reply

Charles *says:*

June 3, 2015 at 10:21 am

Tara,
I am afraid that I still don't understand your question.
Charles

Reply

Tara *says:*

June 3, 2015 at 2:12 pm

I am sorry that I have not been able to explain my question.
For each dependent variable there are 2 factors one factor has 4 levels and the other factor has 2 levels. I can separate the factor with two levels when I test correlation but I want to keep the 4 levels together of the other factor when I test correlation. So I want to test correlation for factor 1 (a,b,c,d) with factor 2(a) then find correlation between factor 1(a,b,c,d) with factor 2(b). I test correlation between two dependent variables. Is this possible?
If so, can I use mean value of levels(a,b,c,d) when I test correlation?
I hope I could explained my question well.
Thanks a lot.

Charles *says:*

June 3, 2015 at 4:14 pm

Tara,

I'm not sure why you want to do this, but in any case here my response to your question based on my understanding of what you are asking.

Suppose the data for 4 variables x1, x2, x3 and x4 are contained in the range R1 (with 4 columns, one

for each variable) and the data for another variable y is contained in the range R2 (with 1 column and the same number of rows as R1). The correlation of x1, x2, x3 and x4 with y can be calculated by the Real Statistics formula MultipleR(R1, R2). This is essentially the R value in multiple linear regression.

The Correlation test described in Correlation Testing is between two variables x and y. If you define the x sample values as the mean of the corresponding values of x1, x2, x3 and x4, you can then test the correlation of x with y. It is not clear to me why this would be useful though.

Charles

question then is wh

Ni *says:*

April 25, 2015 at 3:56 pm

Thank you for your prompt response.

If I don't possess entity level information for any participants within category subgroups can I really correlate subgroups between categories?

Though I possess the standard deviations and means of the categories and subgroups within categories, I don't see how I can calculate covariance. If I can't calculate covariance is there another way to calculate correlation?

Reply

Charles *says:*

April 25, 2015 at 4:11 pm

You clearly need more than just the means and standard deviations of the samples to calculate the covariance, and, as you observed, you need to know the covariance to calculate the correlation.
Charles

Reply

Ni *says:*

April 24, 2015 at 3:13 pm

Mr. Zaiontz,

Great website in so many respects.

Have a correlation question for you.

Here is my data structure:
1. Over fifty categories with the same two subgroups per category. Subgroup 1 Passes and Subgroup 2 Fails.
2. Not all categories possess the same size subgroups and not all categories are the

same size.
3. Data for each category contains both subgroup means and standard deviations as well as the overall category mean and standard deviation.
4. The same participant population was evaluated in all categories. A fail in one category is also a fail in all the other categories.

Question:
With data formatted in this manner is it possible to correlate the categories?

Any insights would be helpful.

Reply

Charles *says:*
April 25, 2015 at 7:31 am

Sorry, but I don't completely understand the premise.
Charles

Reply

jerome *says:*
April 6, 2015 at 6:46 pm

can you please explain pair wise correlation ?

Reply

Charles *says:*
April 7, 2015 at 8:31 am

Jerome,
If you have say 4 variables A, B, C and D, there are C(4,2) = 6 different pairs of these variables, namely AB, AC, AD, BC, BD, CD. Correlation coefficients are calculated on pairs of variables. Thus with 4 variables there are are 6 pairwise correlations, namely correlation(A,B), correlation(A,C), etc.
Charles

Reply

John Rodri *says:*
March 19, 2015 at 3:41 pm

Hi Charles,

Is it possible obtain any extra information from the correlation value? For example, if I have a correction value of 0,85 could we say that 85% of the values correlate. Or can we say that all values correlate with 15% error?

Thanks in advance.

Reply

Charles *says:*
March 19, 2015 at 5:30 pm

John,
Neither of these is true, although you could say that it is 15% short of perfect correlation.

Charles

Reply

John Gonzales *says:*

March 13, 2015 at 6:47 am

what is meant by the definition of correlation coefficient ” The correlation coefficient between two sample variables x and y is a scale-free measure of linear association between the two variables, and is given by the formula,” specifically scale-free measure? Please respond as soon as possible as this is for a project due this Sunday. Thank you for your time. -John G.

Reply

Charles *says:*

March 13, 2015 at 8:15 am

John,

The correlation coefficient is a measure of the linear association between the two variables, but it is not scale-free. E.g. if the sample for variable x is {3,4,5,1,5} and the sample for variable y is {5,2,7,3,4}, then the covariance coefficient is 1.08. If instead I multiply each of the sample elements by 10, the covariance coefficient will be 108, i.e. 10 x 10 = 100 times higher. Thus the covariance coefficient is not scale-free since scale matters (here scale means the size of the input data, not just their relationship to each other).

The correlation coefficient is an attempt to make the covariance coefficient

scale-free. In this way only the relationship between the two variables is captured. Using the above example, the correlation coefficient for the original samples is .419425, the same as the correlation coefficient for the samples that are 10 times bigger. This is a scale-free measure. In fact, no matter what the size of the original data the correlation coefficient has a value between -1 and +1. The closer the correlation coefficient is to +1 the better (higher) the linear association between the two variables (i.e. when x is high, y tends to be high too and when x is low, y tends to be low). The closer the correlation coefficient is to 0 the worse (lower) the linear association between the two variables.

The same is true in the negative range, namely the closer the correlation coefficient is to -1 the better (higher) the linear association between the two variables, except that this time the association is the inverse of the positive association (i.e. when x is high, y tends to be low and when x is low, y tends to be high).

Charles

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

Post Comment