Lab 12

Correlation and Covariance

Applied Statistics and Experiments

November, 2024

Agenda

1. Some definitions.
2. Covariance vs. Variance
3. Pearson correlation coefficient (r)
4. Testing the significance of r

Lecture Recap

https://quizizz.com/join

Join and enter game code

Definitions

- A **Random Variable** X, is a set of numeric outcomes assigned to probabilistic events. X: {(H, H), (H, T), (T, H), (T, T)} → {0, 1, 2}
- **Expectation E(X)**, is the outcomes of a *Random Variable* weighted by their probabilities.
- **Variance Var(X)**, is the difference between Expectation of a squared Random Variable and the Expectation of that Random Variable squared: $Var(X) = E(X^2) - (E(X))^2$

Definitions

- **Variance Var(X)**, is the difference between Expectation of a squared Random Variable and the Expectation of that Random Variable squared..
- **Covariance Cov(X, Y)**, is a measure how two variables vary together:

$$Var(X) = E(X^2) - (E(X))^2$$

$$Cov(X, Y) = E((X - \overline{X})(Y - \overline{Y})) = E(XY) - E(X)E(Y)$$

Definitions

- **Correlation Corr(X, Y)**, is the normalized covariance.

$$Var(X) = E\left(X^2\right) - (E(X))^2$$

$$Cov(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Variance

- The Variance is defined as the average of the squared differences from the Mean

$$E(X) = \frac{\sum x_i}{n}$$

$$\sigma^2 = \frac{\sum (x_i - E(X))^2}{n}$$

Range is]-∞, +∞[

Covariance

- Covariance indicates how two variables are related.
- A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related.

$$E(X) = \frac{\sum x_i}{n}$$

$$Cov(X,Y) = \frac{\sum (x_i - E(X))(y_i - E(Y))}{n}$$

Range is]-∞, +∞[

Covariance

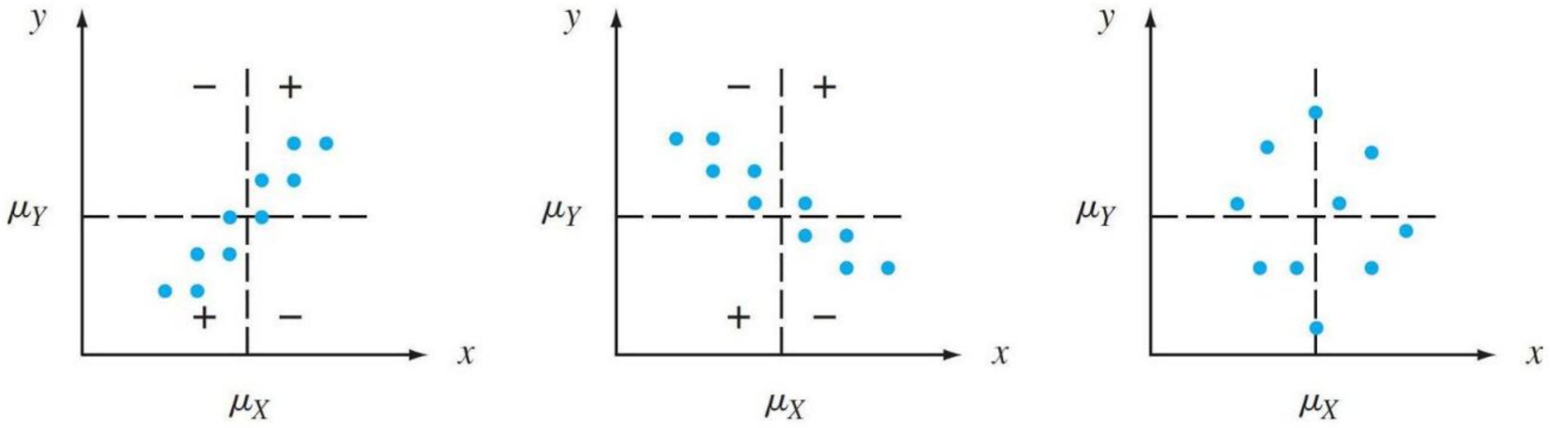
- Covariance indicates how two variables are related.
- A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related.

$$Cov(X,Y) = \begin{cases} \sum_{x \in X} \sum_{y \in Y} (x - E(X))(y - E(Y))p(x,y) & X,\ Y \text{ are discrete} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - E(X))(y - E(Y)) f(x,y)\, dx\, dy & X,\ Y \text{ are continuous} \end{cases}$$

Covariance

We have 3 types of "co-varying"

- If both variables tend to deviate in the same direction (both go above their means or below their means at the same time), then the covariance will be positive. If the opposite is true, the covariance will be negative.
- If X and Y are not strongly related, the covariance will be near 0.



(a) positive covariance; (b) negative covariance; (c) covariance near zero

Bessel's Correction

- While calculating a **sample** variance in order to estimate a population variance, the denominator of the variance equation becomes (n-1)
- This removes bias from the estimation, as it prohibits the researcher from underestimating the population variance.

$$Cov(X,Y) = \frac{\sum(x_i - E(X))(y_i - E(Y))}{n-1}$$

$$\sigma^2 = \frac{\sum(x_i - E(X))^2}{n-1}$$

Covariance – Example

Given the height and weight of 3 top participants in a sport race.

Height (cm) x	160	164	171
Weight (kg) y	53	57	60

Covariance – Example

$$E(X) = \bar{x} = \frac{160 + 164 + 171}{3} = 165$$

$$E(Y) = \bar{y} = \frac{53 + 57 + 60}{3} = 56.67$$

$$Cov(X,\ Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$= \frac{(160 - 165)(53 - 56.67) + (164 - 165)(57 - 56.67) + (171 - 165)(60 - 56.67)}{3}$$

$$= \frac{38}{3} = 12.67\ (\textit{population covariance})$$

Covariance – Example

For the given probability distributions, find the covariance of X and Y.

	X = 0	X = 1	X = 2
Y = 1	0.1	0.2	0.3
Y = 2	0.05	0.15	0.2

Covariance – Example

	X = 0	X = 1	X = 2
Y = 1	P(X=0,Y=1)	P(X=1,Y=1)	P(X=2,Y=1)
Y = 2	P(X=0,Y=2)	P(X=1,Y=2)	P(X=2,Y=2)

Law of total probability. If $C_1, \ldots, C_k$ are disjoint with $C_1 \cup \cdots \cup C_k = \Omega$, then

$$\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(A \cap C_i) = \sum_{i=1}^{k} \mathrm{P}(A|C_i)\mathrm{P}(C_i)$$

Conditional probability

Joint probability

Conditional Probability Formula

Probability of A and B

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A given B

Probability of B

Marginal probability

Bayes Theorem

Covariance – Example

For the given probability distributions, find the covariance of X and Y.

	X = 0	X = 1	X = 2	Total
Y = 1	0.1	0.2	0.3	0.6
Y = 2	0.05	0.15	0.2	0.4
Total	0.15	0.35	0.5	1

Joint probability

`P(X=0,Y=1)=0.1, P(X=0,Y=2)=0.05`

Marginal probability

`P(X=0)= P(X=0,Y=1)+P(X=0,Y=2)=0.1+0.05=0.15`

Covariance – Example

$$E(X) = \bar{x} = \sum xP(X = x) = 0*0.15 + 1*0.35 + 2*0.5 = 1.35$$

$$E(Y) = \bar{y} = \sum yP(Y = y) = 1*0.6 + 2*0.4 = 1.4$$

$$Cov(X,\ Y) = \sum_i \sum_j (x_i - E(X))(y_j - E(Y))p(x, y)$$

$$= (0-1.35)(1-1.4)(0.1) + (0-1.35)(2-1.4)(0.05) + (1-1.35)(1-1.4)(0.2) + (1-1.35)(2-1.4)(0.15)$$
$$+ (2-1.35)(1-1.4)(0.3) + (2-1.35)(2-1.4)(0.2)$$
$$= 0.01$$

Covariance – Example

Let the joint probability density function be

$$f(x, y) = \frac{1}{e^{x+y}} \qquad (x, y \geqslant 0)$$

Calculate Cov(X, Y) where X and Y are continuous random variables.

Covariance – Example

$$E(X) = \int_0^{+\infty} x f(x)\, dx$$

$$f(x) = \int_{y=0}^{+\infty} f(x,y)\, dy \;= \int_0^{+\infty} \frac{1}{e^{x+y}}\, dy = \int_0^{+\infty} e^{-x-y}\, dy$$

$$= \int_0^{+\infty} e^{-x-y}\, dy = e^{-x}\left[-e^{-y}\right]_0^{\infty} = e^{-x}(0+1) = e^{-x}$$

$$E(X) = \int_0^{+\infty} x f(x)\, dx = \int_0^{+\infty} x\, e^{-x}\, dx = \left[-xe^{-x}\right]_0^{\infty} + \int_0^{+\infty} e^{-x}\, dx = 0 + \left[-e^{-x}\right]_0^{\infty} = 1$$

$$f(y) = e^{-y}$$
$$E(Y) = 1$$

Covariance – Example

$$
\begin{aligned}
Cov(X, Y) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - E(X))(y - E(Y))f(x, y)\,dx\,dy \\
&= \int_{0}^{+\infty}\int_{0}^{+\infty}(x-1)(y-1)e^{-x-y}\,dx\,dy \\
&= \int_{0}^{+\infty}(y-1)e^{-y}\left[\int_{0}^{+\infty}(x-1)e^{-x}\,dx\right]dy \\
&= \int_{0}^{+\infty}(y-1)e^{-y}\left[\int_{0}^{+\infty}xe^{-x}\,dx - \int_{0}^{+\infty}e^{-x}\,dx\right]dy \\
&= \int_{0}^{+\infty}(y-1)e^{-y}\,[1-1]\,dy \\
&= 0
\end{aligned}
$$

The Pearson Correlation Coefficient

- The correlation coefficient of X and Y, denoted by Corr(X, Y) or $\rho_{X,Y}$ is defined by:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- It represents a “scaled” covariance – correlation ranges between -1 and +1.
- Correlation is another way to determine how two variables are related.
- In addition to telling you whether variables are positively or inversely related, correlation also tells you the degree to which the variables tend to move together.

The Pearson Correlation Coefficient (r)

- The correlation coefficient of X and Y, denoted by Corr(X, Y) or $\rho_{X,Y}$ is defined by:

n only

Should we use n or n-1?

$$\rho = \frac{n * \sum_i x_i * y_i - \sum_i x_i \sum_i y_i}{\sqrt{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} * \sqrt{n \sum_i y_i^2 - \left(\sum_i y_i\right)^2}}$$

n or n-1

$$Cov(X,Y) = \frac{\sum (x_i - E(X))(y_i - E(Y))}{n}$$

$$\rho = \frac{SS_{xy}}{\sqrt{SS_{xx}} * \sqrt{SS_{yy}}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} * \sqrt{\sum (y_i - \overline{y})^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)} * \sqrt{Var(Y)}}$$

Correlation – Example

- Is there a relationship between the age at which a child first begins to speak and his or her mental ability later on?
- To answer this question a study was conducted in which the age (in months) at which a child first spoke and the child's score on an aptitude test as a teenager were recorded:

Age (X)	15	20	8	10
Score (Y)	100	40	60	80

Correlation – Example

					Total
Age (X)	15	20	8	10	53
Score (Y)	100	40	60	80	280
X^2	225	400	64	100	789
Y^2	10000	1600	3600	6400	21600
XY	1500	800	480	800	3580

$$\rho = \frac{n * \sum_i x_i * y_i - \sum_i x_i \sum_i y_i}{\sqrt{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2} * \sqrt{n \sum_i y_i^2 - \left(\sum_i y_i\right)^2}} = \frac{4*3580 - 53*280}{\sqrt{4*789 - (53)^2} * \sqrt{4*21600 - (280)^2}} = -0.3121$$

Testing the significance of correlation coefficient (ϱ)

- The significance of correlation coefficient (r) can be tested by Student's t test.
- The p value is calculated using a t distribution with (n−2) degrees of freedom.
- The test statistics is given by

$$t = \frac{|\rho|}{\sqrt{\frac{1-\rho^2}{n-2}}}$$

Testing the significance of ϱ – Example

Compute Pearson's coefficient of correlation between advertising costs and sales as per the data given below:

Cost (X)	39	65	62	90	82	75	25	98	36	78
Sales (Y)	47	53	58	86	62	68	60	91	51	84

And test its significance at $\alpha = 5\%$

Testing the significance of ϱ – Example

$H_0 : \rho = 0$ (*The correlation coefficient IS NOT significantly different from zero.*)
$H_A : \rho \neq 0$ (*The correlation coefficient IS significantly DIFFERENT FROM zero*)

$n = 10$

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

$$Cov(X,Y) = E((X - E(X))(Y - E(Y)))$$

$$Var(X) = E\left(X^2\right) - E(X)^2$$

$$E(X) = \frac{39 + 65 + 62 + 90 + 82 + 75 + 25 + 98 + 36 + 78}{10} = 65$$

$$E(Y) = \frac{47 + 53 + 58 + 86 + 62 + 68 + 60 + 91 + 51 + 84}{10} = 66$$

Testing the significance of ϱ – Example

											Total
Cost (X)	39	65	62	90	82	75	25	98	36	78	650
Sales (Y)	47	53	58	86	62	68	60	91	51	84	660
$X - E(X)$	-26	0	-3	25	17	10	-40	33	-29	13	
$Y - E(Y)$	-18	-12	-7	21	-3	3	-5	26	-14	19	
$(X - E(X)) * (Y - E(Y))$	468	0	21	525	-51	30	200	858	406	247	2704
X^2	1521	4225	3844	8100	6724	5625	625	9604	1296	6084	47648
Y^2	2209	2809	3364	7396	3844	4624	3600	8281	2601	7056	45784

Testing the significance of ϱ – Example

$$Var(X) = E(X^2) - E(X)^2 = \frac{47648}{10} - \left(\frac{650}{10}\right)^2 = 539.8$$

$$Var(Y) = E(Y^2) - E(Y)^2 = \frac{45784}{10} - \left(\frac{660}{10}\right)^2 = 222.4$$

$$Cov(X, Y) = E[(X - E(X)) * (Y - E(Y))] = \frac{2704}{10} = 270.4$$

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X) * Var(Y)}} = \frac{270.4}{\sqrt{539.8 * 222.4}} = 0.780410054$$

$$t = \frac{0.780410054}{\sqrt{\frac{1 - (0.780410054)^2}{10 - 2}}} = \frac{0.780410054}{0.221065643} = 3.53021864$$

Testing the significance of ρ – Example

$Two - tailed\ test$
$df = n - 2 = 10 - 2 = 8$

We have two critical values which are: $t_{\alpha/2,\,df} = t_{0.025,\,8} = -2.306$
$t_{1-\alpha/2,\,df} = t_{0.975,\,8} = 2.306$

The critical region (CR) is $(-\infty, -2.306] \cup [2.306, +\infty)$
Since $t \in CR$, we reject the null hypothesis, the correlation coefficient ρ is significant.

Conclusion. There is sufficient evidence to conclude that there is a significant linear relationship between between advertisement costs (X) and the sales (Y) because the correlation coefficient is significantly different from zero.

References

- https://www.probabilitycourse.com/chapter5/5_3_1_covariance_correlation.php
- https://www.colorado.edu/amath/sites/default/files/attached-files/ch5_covariance_0.pdf
- https://online.stat.psu.edu/stat414/book/export/html/728
- https://www.wolframalpha.com/input?i=Covariance
- https://statisticseasily.com/coefficient-of-determination-vs-coefficient-of-correlation/

Attendance

https://baam.duckdns.org

Questions?