روش تحقیق و آمار کاربردی

همبستگی و رگرسیون

تهیه کننده: دکتر طیبه موسوی میانگاه
گروه زبانشناسی دانشگاه پیام نور
mosavit@pnu.ac.ir

ضریب همبستگی

میزان رابطه میان دو دسته متفاوت از نمرات که متعلق به گروه واحدی از افراد باشند را با ضریب همبستگی (correlation coefficient) تعیین می کنند. مثلاً میتوان همبستگی بین نمرات درس تربیت بدنی و درس ریاضی یک گروه از دانش آموزان را محاسبه نمود. همبستگی را می توان برای دو گروه مختلف که براساس یک منطق در یک متغیر همتراز شده باشند نیز محاسبه کرد. مثلاً می توان همبستگی بین بهره هوشی زنان و بهره هوشی شوهرانشان را محاسبه کرد تا مشخص شود که آیا زنان باهوش با مردان باهوش ازدواج می کنند یا نه.

ضریب همبستگی گشتاوری پیرسون

متداول ترین روش محاسبه ضریب همبستگی بوده ، مقدار آن از -۱ تا ۱_ متغیر است.
بیشترین استفاده از این ضریب همبستگی زمانی است که متغیرها بااستفاده از **مقیاس فاصله ای یا نسبتی** اندازه گیری شده باشند.
ضریب همبستگی پیرسون از راه انحراف از میانگین بااستفاه از فرمول زیر محاسبه می شود:

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}$$

$$x = X - \bar{X} \qquad y = Y - \bar{Y}$$

مثال: نمرات دو درس آمار و کامپیوتر یک گروه ۵ نفری از دانشجویان را درنظر بگیرید. برای تعیین ضریب همبستگی میان نمرات این دو درس ابتدا جدولی مانند جدول زیر را تهیه کرده و داده ها را در آن وارد می کنیم. میانگین نمرات هر درس را محاسبه کرده، نمرات انحراف (فاصله هرنمره از میانگین) را نیز مشخص کرده در ستونهای x و y درج می کنیم. سپس نمرات انحراف را مجذور کرده و مقدار xy را نیز بدست می آوریم.

دانشجو	آمارX	کامپیوترY	x	y	X^2	Y^2	xy
1	19	16	15-19=-4	13-16=-3	361	256	**12**
2	18	14	15-18=-3	13-14=-1	324	196	**3**
3	12	9	15-12=3	13-9=4	144	81	**12**
4	14	10	15-14=1	13-10=3	196	100	**3**
5	12	16	15-12=3	13-16=-3	144	256	**9**
	$\bar{X}$ = 15	$\bar{X}$ = 13			Σ 1169	Σ 889	**Σ 39**

$$r_{xy} = \frac{39}{\sqrt{1169 \times 889}} = \frac{39}{1019.431} = 0.038$$

$$r_{xy} = +0.038$$

ضریب همبستگی رتبه ای اسپیرمن

زمانی که تعداد نمرات کم بوده و نمرات به صورت **رتبه ای** هستند (رتبه بندی شده اند) یعنی به جای اعداد تنها رتبه های آنها در دسترس باشند می توان از ضریب همبستگیِ رتبه ای اسپیرمن استفاده نمود. در این نوع همبستگی فواصل بین رتبه ها ضرورتاً مساوی نیستند.
برای محاسبه این نوع همبستگی از فرمول زیر استفاده می شود:

$$r_s = 1 - \frac{6\ \Sigma\ D^2}{n(n^2-1)}$$

D^2 = مجذور تفاوت بین رتبه های هر جفت نمرات
n = تعداد نمره های هر یک از دو گروه

مثال: به جدول زیر توجه کنید که شیوه محاسبه ضریب همبستگی برای دو دسته نمره رتبه بندی شده را نشان می دهد:

X	Y	d	D^2
1	1	0	0
3	2	1	1
5	3	2	4
4	4	0	0
2	5	-3	9
			Σ 14

$$r_s = 1 - \frac{6\,\Sigma\,D^2}{n(n^2-1)} = \quad r_s = 1 - \frac{6 \times 14}{5 \times 24} = 1 - 0.7 = +0.3$$

همبستگی دورشته ای نقطه ای

ضریب همیستگی دورشته ای نقطه ای مورد خاصی از ضریب همبستگی گشتاوری است. وقتی دو متغیر داشته باشیم که یکی پیوسته و دیگری دوارزشی (یعنی دارای دو مقدار صفر یا یک) باشد میتوان از روش همبستگی دورشته ای نقطه ای استفاده نمود. برلی محاسبه ضریب همیستگی دورشته ای نقطه ای بااستفاده از فرمول زیر محاسبه می شود:

$$r_{pbis} = \frac{\bar{X}_p - \bar{X}_t}{\sigma_t}\sqrt{\frac{p}{q}}$$

$\bar{X}_p$ = میانگین نمرات متغیر پیوسته برای همه آزمون شوندگان با نمره ۱ در آزمون دوارزشی گرفته اند

$\bar{X}_t$ = میانگین نمرات برای همه آزمون شوندگان در متغیر پیوسته

σ_t = انحراف معیار همه نمرات در متغیر پیوسته

p = نسبت همه آزمون شوندگانی که در متغیر دوارزشی نمره ۱ گرفته اند

$q = 1 - p$ = نسبت آزمون شوندگانی که در متغیر دوارزشی نمره ۰ گرفته اند

مثال: در جدول زیر داده های مربوط به تعداد برگه های جریمه یک راننده و نیز عملکرد او در نخستین آیین نامه رانندگی (بصورت قبول (۱) و رد (۰)) وارد شده است.

شماره راننده	آزمون آئین نامه X	تعداد برگه های جریمه Y
1	1	5
2	0	3
3	1	1
4	1	4
5	1	7
6	0	0
7	1	1
8	1	6
9	0	5
10	1	2
11	1	6
12	0	3
13	0	1
14	1	4
15	1	2
	Σ= 10	Σ= 50

با قراردادن داده های جدول در فرمول مربوطه ضریب همبستگی دورشته ای نقطه ای این دو متغیر چنین بدست می آید:

$$\bar{X}_t = \frac{50}{15} = 3.33 \qquad p = \frac{10}{15} = 0.67$$

$$\bar{X}_p = \frac{5+1+4+7+1+6+2+6+4+2}{10} = 3.80$$

$$\sigma_t = \sqrt{\frac{(5-33.3)^2 + (3-33.3)^2 + \cdots + (2-33.3)^2}{N}} = 3.80$$

$$= \sqrt{\frac{232}{15} - (3.33)^2} = 2.1$$

$$r_{pbis} = \frac{3.80 - 3.33}{2.1}\sqrt{\frac{0.67}{0.33}} = 0.31$$

با توجه به ضریب همبستگی بدست آمده می توان نتیجه گرفت که کسانی که در اولین آیین نامه رانندگی قبول شده اند بیشتر از کسانی که در اولین آزمون موفق نبوده اند مرتکب جرائم رانندگی شده اند، ولی این همبستگی زیاد نیست.

همبستگی فای

ضریب همبستگی فای مورد خاصی از ضریب همبستگی گشتاوری پیرسون است. زمانی از این نوع همبستگی استفاده می شود که هر دو متغیر دوارزشی واقعی باشند.
با استفاده از فرمول زیر ضریب همبستگی فای دو متغیر را می توان محاسبه نمود:

$$\emptyset = \frac{p_{ij} - p_i\, p_i}{\sqrt{(p_i\, q_i)\,(p_j\, q_j)}}$$

که در آن:

p_{ij} = نسبت افرادی که در هردو متغیر نمره قبولی گرفته اند.
p_i = نسبت افرادی که در متغیر X نمره قبولی گرفته اند.
p_j = نسبت افرادی که در متغیر Y نمره قبولی گرفته اند.

به عنوان مثال می توان ضریب همبستگی متغیر جنسیت را با متغیر شکست / موفقیت در یک امر اندازه گیری نمود.

در جدول زیر داده های مربوط به دو متغیر جنسیت و نتایج آزمون رانندگی وارد شده است. افراد مذکر با ۱ و افراد مؤنث با ۰ ارزش گذاری شده اند. قبولی در آزمون با ۱ و مردودی با ۰ نشان داده شده است.

متغیرها										
جنسیت	1	0	1	1	0	0	0	0	1	1
آزمون	0	0	1	0	0	1	0	0	1	1

$$\emptyset = \frac{p_{ij} - p_i\, p_i}{\sqrt{(p_i\, q_i)\,(p_j\, q_j)}}$$

$$p_i\ \frac{5}{10} = 0.5$$

$$p_j\ \frac{4}{10} = 0.4$$

$$p_{ij}$$

$$\frac{3}{10} = 0.3$$

$$q_i = 0.5$$

$$q_j = 0.6$$

$$\emptyset = \frac{0.3 - (0.5)(0.4)}{\sqrt{[(0.5)(0.5)][(0.4)(0.6)]}} = 0.42$$

رگرسیون (همبستگی از نوع پیش بینی)

اگر بین دو متغیر نوعی همبستگی وجود داشته باشد می توان صفتی از یک متغیر را ازطریق متغیر دیگر پیش بینی نمود.

رگرسیون روشی است برای پیش بینی متغیر وابسته (y) یا متغیر ملاک ازطریق یک یا چند متغیر مستقل (x_i) یا متغیرهای پیش بین.

مثلاً برای پیش بینی میزان موفقیت دانشجویی در دانشگاه می توان از معدل نمره های او در سال آخر دبیرستان استفاده نمود.

دقت پیش بینی به میزان همبستگی بین دو متغیر ملاک و پیش بین بستگی دارد.

پیش بینی نمره های استاندارد (Z)

در پیش بینی نمره های استاندارد می توان از ضریب همبستگی پیرسون استفاده نمود بدین معنا که نمره استاندارد متغیری که می خواهیم پیش بینی کنیم - یعنی برابر است با حاصل ضرب نمره استاندارد متغیر $Z_{y\acute{e}}$ - پیش بینی کننده (پیش بین) در ضریب همبستگی دو متغیر پیش بین و ملاک بصورت زیر:

$$Z_{y\acute{e}} = (Z_x)(Z_{xy})$$

Z_y = نمره پیش بینی شده برای متغیر y

Z_x= نمره استاندارد متغیر x که برای پیش بینی بکار برده می شود

Z_{xy}= ضریب همبستگی بین متغیرهای x و y

مثال: اگر همبستگی بین دو درس مبانی زبانشناسی و ساخت واژه 0.70 و نمره استاندارد یک دانشجو در درس مبانی زبانشناسی 2.5 باشد، پیش بینی می شود که نمره این دانشجو در درس ساخت‌واژه 1.75 باشد، بصورت زیر: $2.5 \times 0.70 = 1.75$

- اگر ضریب همبستگی بین دو متغیر کامل و مثبت باشد ، می توان پیش بینی $r_{xy} = 1$ نمود که نمره استاندارد هر فرد در متغیر x برابر است با نمره استاندارد او در متغیر y.
- اگر ضریب همبستگی بین دو متغیر کامل و منفی باشد ، می توان پیش بینی $r_{xy} = -1$ نمود که نمره استاندارد هر فرد در متغیر x برابر است با نمره استاندارد او در متغیر y منتها با علامت مخالف. در این حالت و حالت فوق پدیده رگرسیون اتفاق نمی افتد. تنها زمانی رگرسیون وجود دارد که همبستگی بین دو متغیر کامل نباشد.
- هرچه شدت همبستگی بین دو متغیر کمتر باشد نمره های پیش بینی شده به میانگین نزدیکتر هستند و بالعکس هرچه شدت همبسگی بیشتر باشد رگرسیون به طرف میانگین کمتر است.

پیش بینی با نمره های خام

برای پیش بینی y از روی x بااستفاده از نمره های خام باید دو عامل را در نظر بگیریم: اول شیب رگرسیون (ضریب b) و سپس مقدار y در صورتی که مقدار x صفر باشد یعنی محل تلاقی خط رگرسیون با محور y(ضریب a) .

برای محاسبه معادله پیش بینی بااستفاده از نمره های خام باید ابتدا خطای پیش بینی محاسبه شود که لازمه آن دانستن مقدار a و bمی باشد.

خطای پیش بینی عبارتند از اختلاف بین نمره مشاهده شده (y) و نمره پیش بینی شده ($\acute{y}$).

محاسبه نمره پیش بینی شده بصورت زیر:

$$\acute{y} = a + bx$$

و فرمول خطای پیش بینی بصورت زیر است:

$$y - \acute{y} = y - (a + bx)$$

برای محاسبه ضرایب خط رگرسیون (مقادیر a و b) از راه نمرات خام می توان از فرمول های زیر استفاده نمود (و در صورتیکه بخواهیم مقدار y را از روی x پیش بینی کنیم):

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$a_{yx} = \frac{\Sigma y - b_{yx}\,\Sigma x}{n}$$

b_{xy} = ضریب برای پیش بینی y از روی x

n= تعداد آزمودنیها

Σxy= مجموع حاصل ضرب هر نمره x در هر نمره y

Σx = مجموع نمره های x

Σy = مجموع نمره های y

مثال: بااستفاده از داده های جدول زیر شیوه استفاده از فرمول های زیر تا رسیدن به معادله پیش بینی را بررسی می کنیم:

X	y	x^2	y^2	xy
10	8	100	64	**80**
9	8	81	64	**72**
8	4	64	16	**32**
5	9	25	81	**45**
6	6	36	36	**36**
4	6	16	36	**24**
3	7	9	49	**21**
2	6	4	36	**12**
2	3	3	9	**6**
1	3	1	9	**3**

$$n = 10\,, \qquad \Sigma x = 50, \qquad \Sigma y = 60, \qquad \Sigma x^2 = 340,$$

$$\Sigma y^2 = 400, \qquad \Sigma xy = 331$$

حال مقادیر بدست آمده را در فرمول های b و a جاگذاری بصورت زیر می کنیم:

$$b_{yx} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b_{yx} = \frac{(10)(331) - (50)(60)}{(10)(340) - (50)^2} = \frac{310}{90} = 0.344$$

$$a_{yx} = \frac{\Sigma y - b_{xy}\,\Sigma x}{n}$$

$$a_{yx} = \frac{60 - (0.344)(50)}{10} = \frac{42.8}{10} = 4.28$$

و در نهایت مقادیر a و b را جایگزین نموده و معادله پیش بینی را بصورت زیر خواهیم داشت:

$$\acute{y} = a + bx$$
$$\acute{y} = 4.28 + 0.344\,x$$