ضریب همبستگی رشته‌ای

مهران نقی‌زاده قمی [۱] شکوفا کبیری [۲]

چکیده:

در این مقاله، به بررسی رابطه‌ی بین دو متغیر در حالتی که یک متغیر در مقیاس فاصله‌ای یا نسبتی و متغیر دیگر با مقیاس اسمی اندازه‌گیری شده باشد می‌پردازیم. در چنین حالاتی از ضریب همبستگی رشته‌ای استفاده می‌شود. محاسبه‌ی چند ضریب رشته‌ای شامل ضریب دورشته‌ای و ضریب دورشته‌ای- نقطه‌ای با چند مثال و به کمک نرم‌افزار R انجام می‌شود.

واژه‌های کلیدی: ضریب دورشته‌ای، ضریب دورشته‌ای- نقطه‌ای.

۱ مقدمه

یکی از مهمترین شاخص‌های رابطه‌ی بین دو متغیر کمی X و Y، ضریب همبستگی پیرسون است. برای توصیف همبستگی دو متغیر کمی بدون هیچ فرضی می‌توان از ضریب همبستگی خطی پیرسون استفاده کرد. اما برای انجام آزمون معنی‌داری ضریب همبستگی پیرسون و قضاوت در خصوص همبستگی دو متغیر در جامعه، شرط نرمال بودن توزیع توأم دو متغیر کمی در جامعه لازم است. این پذیره به کمک یکی از آزمون‌های شاپیرو- ویلک (قابل انجام در نرم‌افزار R به‌کمک بسته‌ی mvShapiro.Test)، دورنیک- هانسن، هنز- زیکلر و معیار چولگی و کشیدگی ماردیا (قابل انجام در نرم‌افزار STATA 12) قابل بررسی است. در مواردی که یکی از متغیرها دو یا چندحالتی و متغیر دیگر کمی باشد و توزیع آن در جامعه نرمال فرض شود، از ضرایب همبستگی مستخرج از فرمول گشتاوری پیرسون به نام ضرایب رشته‌ای استفاده می‌شود. قصد داریم این ضرایب را با چند مثال مورد بررسی قرار دهیم.

۲ ضریب دورشته‌ای

وقتی از یک جامعه‌ی آماری دو متغیر برای نمونه‌ای استخراج کنیم که یکی در مقیاس نسبتی یا فاصله‌ای و دیگری یک متغیر دوحالتی باشد، برای محاسبه‌ی همبستگی متغیرها از ضریب دورشته‌ای می‌توان استفاده کرد. یکی از پذیره‌های این ضریب همبستگی این است که متغیر دو حالتی باید ذاتاً کمی باشد که عدم وجود وسیله‌ی درست اندازه‌گیری یا سایر محدودیت‌های مطالعه باعث می‌شود نتوان آن را دقیق اندازه‌گیری نمود. بنابراین با فرض این‌که توزیع واقعی آن نرمال باشد می‌توان آزمون معنی‌داری ضریب دورشته‌ای را انجام داد. پذیره‌ی دوم این است که رابطه‌ی بین متغیر پیوسته و متغیر دوحالتی باید خطی باشد. ضریب دورشته‌ای به‌صورت

$$r_{bis} = \frac{d}{s}\frac{pq}{y} \qquad (۱)$$

محاسبه می‌شود که در آن $d = \bar{x}_p - \bar{x}_q$ اختلاف بین میانگین‌های متغیر پیوسته در دو طبقه از متغیر دوحالتی، s انحراف معیار نمونه برای متغیر پیوسته، p و q به ترتیب نسبت حالت‌ها در طبقه‌ی اول و دوم متغیر دو

[۱] استادیار گروه آمار دانشگاه مازندران m.naghizadeh@umz.ac.ir

[۲] کارشناس آمار، دانشگاه مازندران

حالتی و y ارتفاع در نقطه‌ای است که منحنی نرمال استاندارد به دو قسمت p و q تقسیم می‌شود که با دستور $dnorm(qnorm(min(p,q)))$ در نرم‌افزار R قابل محاسبه است. برای آزمون $H_0 : \rho_{bis} = 0$ در مقابل $H_1 : \rho_{bis} \neq 0$ آماره‌ی آزمون به صورت

$$t = \sqrt{\frac{[\frac{y}{\sqrt{pq}}\, r_{bis}]^2 (n-2)}{1 - [\frac{y}{\sqrt{pq}}\, r_{bis}]^2}} \qquad (2)$$

دارای توزیع مجانبی t-استودنت با $n-2$ درجه‌ی آزادی است که در آن n حجم نمونه است.

مثال ۱.۲. *می‌خواهیم همبستگی بین زمان تمرین در هفته و زمانی که برای شنا کردن ۲۵ متر لازم است را تعیین کنیم. تعداد ۸ نفر به مدت ۳ ساعت در هفته و تعداد ۱۶ نفر، ۶ ساعت در هفته تمرین می‌کنند. زمان تمرین یک متغیر دوحالتی غیرواقعی و زمان شنای ۲۵ متر یک متغیر پیوسته است. داده‌ها در جدول ۱ نشان داده شده است.*

جدول ۱. زمان شنای ۲۵ متر و زمان تمرین در هفته

۶ ساعت تمرین هفتگی	۳ ساعت تمرین هفتگی
۱/۹	۳/۱۸
۹/۱۰	۶/۱۷
۸/۱۰	۳/۱۲
۱/۸	۲/۹
۲/۱۱	۱/۱۶
۲/۹	۱/۱۲
۶/۱۲	۲/۱۱
۹/۱۰	۶/۱۰
۲/۹	
۴/۹	
۲/۹	
۴/۱۱	
۹/۸	
۶/۱۴	
۳/۱۱	
۱/۸	

برای محاسبه‌ی ضریب همبستگی رشته‌ای، از دستورهای R بهره می‌گیریم.

```
cor.bis<-function(x,y){
d<-mean(x)-mean(y)
s<-sd(c(x,y))
n=length(x)+length(y)
p<-length(x)/n
q<-length(y)/n
y<-dnorm(qnorm(min(p,q)))
r<-d/s*p*q/y
z=y/sqrt(p*q)*r
cat("r.bis=",r,"\ n")
t<-sqrt(z^2*(n-2)/(1-z^2))
cat("t.statistic=",t,"\ nsig=",
2*min(pt(t,n-2),1-pt(t,n-2)),"\ n")}
y<-c(18.3,17.6,12.3,9.2,16.1,12.1,11.2,10.6)
x<-c(9.1,10.9,10.8,8.1,11.2,9.2,12.6,10.9,9.2,
9.4,9.2,11.4,8.9,14.6,11.3,8.1)
cor.bis(x,y)
```

خروجی R به صورت زیر است:

```
r.bis=-0.6826444
t.statistic= 2.904941
sig= 0.008211725
```

چون گروهی که ۶ ساعت در هفته تمرین می‌کنند ($\bar{x}_p = ۱۰/۳۱$) دارای میانگین زمان شنای کمتری در مقایسه با گروهی که ۳ ساعت در هفته تمرین می‌کنند ($\bar{x}_q = ۱۳/۴۳$) هستند، همبستگی بین زمان تمرین و زمان شنا ۰/۶۸- به دست آمد. چون زمان کمتر در نتیجه‌ی شنای سریع‌تر حاصل می‌شود، همبستگی مثبتی بین زمان تمرین و سرعت شنا وجود دارد. با توجه به مقدار معنی‌داری ($sig = ۰/۰۰۸ < ۰/۰۵$)، فرضیه‌ی صفر در سطح معنی‌داری $\alpha = ۰/۰۵$ رد می‌شود و مقدار $r_{bis} = -۰/۶۸$ نشان می‌دهد که همبستگی مثبت معنی‌داری بین زمان تمرین و سرعت شنا وجود دارد.

تذکر ۲.۲. *توزیع دقیق نمونه‌گیری r_{bis} موجود نیست ولی تحت شرایطی می‌توان از تبدیل لگاریتمی فیشر استفاده کرد. اگر نسبت‌های p و q در جامعه تقریباً برابر با ۰/۵، n (تعداد کل آزمودنی‌ها) بزرگ باشد و r_{bis} به ± 1 نزدیک نباشد می‌توان نشان داد که* $r'_{bis} = \frac{1}{2} \ln \frac{1+r_{bis}}{1-r_{bis}}$

دارای توزیع نرمال با خطای استاندارد نمونه‌گیری $\frac{1}{\sqrt{n}}$ است. همچنین کارایی r_{bis} برای برآورد ρ وقتی $\rho = 0$ باشد خوب و برای مقادیر ρ نزدیک به ± 1 نامناسب می‌شود.

۳ ضریب دورشته‌ای- نقطه‌ای

ضریب دورشته‌ای- نقطه‌ای مانند ضریب دورشته‌ای، شاخص رابطه‌ی بین یک متغیر پیوسته در مقیاس حداقل فاصله‌ای و یک متغیر دوحالتی واقعی مانند جنسیت را فراهم می‌کند. ضریب دورشته‌ای- نقطه‌ای به‌صورت

$$r_{pbis} = \frac{d}{s}\sqrt{pq} \qquad (۳)$$

محاسبه می‌شود که در آن $d = \bar{x}_p - \bar{x}_q$ اختلاف بین میانگین‌های متغیر پیوسته در دو طبقه از متغیر دوحالتی، s انحراف معیار نمونه برای متغیر پیوسته و p و q به ترتیب نسبت حالت‌ها در طبقه‌اول و دوم متغیر دو حالتی هستند. برای آزمون $H_0 : \rho_{bis} = 0$ در مقابل $H_1 : \rho_{bis} \neq 0$ آماره‌ی آزمون به صورت

$$t = \sqrt{\frac{r^2_{pbis}(n-2)}{1-r^2_{pbis}}} \qquad (۴)$$

دارای توزیع مجانبی t- استودنت با $n-2$ درجه‌ی آزادی است که در آن n حجم نمونه است.

مثال ۱.۳. *می‌خواهیم همبستگی بین نمره‌های آزمون مهارت بسکتبال و وسعت دبیرستانی که بازیکنان قبلاً در آن مسابقه داده‌اند را تعیین کنیم. جدول ۲ داده‌ها را نشان می‌دهد.*

جدول ۲. نمره‌های آزمون مهارت و وسعت دبیرستان

وسعت کم	وسعت زیاد
۳۵	۳۰
۳۶	۳۳
۳۸	۳۴
۳۹	۳۶
۵۰	۳۷
۵۱	۴۰
۵۵	۴۲

برنامه‌ی R برای محاسبه‌ی ضریب همبستگی دورشته- نقطه‌ای به صورت زیر است:

```
cor.pbis.test<-function(x,y){
d<-mean(x)-mean(y)
s<-sd(c(x,y))
n=length(x)+length(y)
p<-length(x)/n
q<-length(y)/n
r<-(d/s)*sqrt(p*q)
cat("r.pbis=",r,"\ n")
t<-sqrt(r^2*((n-2)/(1-r^2))
cat("t.statistic=",t,"\ nsig="
,2*min(pt(t,n-2),1-pt(t,n-2)),"\ n")}
x<-c(35,36,38,39,50,51,55)
y<-c(30,33,34,36,37,40,42)
cor.pbis.test(x,y)
```

خروجی R به صورت زیر است:

```
r.pbis= 0.5044248
t.statistic= 2.023704
sig= 0.06585428
```

با توجه به مقدار معنی‌داری ($sig = 0/06 > 0/05$)، فرضیه‌ی صفر در سطح معنی‌داری $\alpha = 0/05$ رد نمی‌شود و نتیجه می‌گیریم که رابطه‌ی معنی‌داری بین وسعت دبیرستان و توانایی در مهارت بسکتبال وجود ندارد.

تذکر ۲.۳. *می‌توان نشان داد که هر چه تفاوت p و q بیشتر باشد، تفاوت r_{bis} و r_{pbis} بیشتر است. در عمل از r_{bis} کمتر استفاده می‌شود و فقط مواقعی کاربرد دارد که از نرمال بودن توزیع واقعی متغیر دوحالتی اطمینان داشته باشیم. در صورت عدم اطمینان، استفاده از r_{pbis} مناسب‌تر است.*

مراجع

[۱] میناسیان، و. (۱۳۸۸)، *آمار در تربیت بدنی و علوم ورزشی*، انتشارات علم و حرکت.

[2] Walker, H. M. (1953), *Statistical Inference*, New York:Holt, Rinehart and Winston.