آزمایشگاه آموزشی
بیست‌ودومین المپیاد
زیست‌شناسی ایران

آمار زیستی

آالی‌ی

روز چهارم
۹۸/۵/۳

اهداف آزمایش:

۱. به دست آوردن نگاهی جامع به آمار و روش‌های آماری مورد استفاده در علوم زیستی

زمان آزمایش: ۹۰ دقیقه

این فایل به منظور آموزش عملی دانش‌پژوهان المپیاد زیست‌شناسی ایران گردآوری شده است.

آمار توصیفی

رسیدن به نتیجه‌گیری‌های قابل استفاده از انبوه داده‌های خام امکان پذیر نیست. برای توصیف داده‌های خام از آماره‌هایی استفاده می‌شود که ویژگی‌های داده‌ها را در بر دارند. به صورت کلی می‌توان آماره‌های توصیفی را به سه دسته تقسیم کرد.

آماره های توصیف مرکزیت

میانگین یا mean

میانگین حسابی از تقسیم مجموع داده ها بر تعداد داده ها به دست میاید.

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$

میانه یا median

پس از مرتب سازی داده ها به ترتیب مقدار، داده‌ی میانی میانه‌ی نمونه است. در صورت زوج بودن بودن تعداد داده میانگین دو داده میانی به عنوان میانه در نظر گرفته میشود. ارجحیت این آماره نسبت به میانگین عدم تغییر آن توسط داده های پرت است.

مد یا mode

پر تکرار ترین داده در بین نمونه

آماره‌های توصیف پراکندگی

واریانس یا variance

وردایی چگونگی پراکندش داده‌ها حول میانگین را نشان می‌دهد و به شکل زیر محاسبه می‌شود.

$$\mathrm{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

σ^2 , s^2

انحراف معیار یا standard deviation

نشان می‌دهد به طور متوسط داده‌ها از میانگین حسابی چه مقدار فاصله دارند. برخلاف واریانس هم‌بعد با داده‌ها می‌باشد.

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

σ, SD

ضریب تغییرات یا coefficient of variation

برای استاندارد سازی و مقایسه‌ی انحراف معیار دو مجموعه داده با میانگین متفاوت استفاده می‌شود.

CV

$$\widehat{c_{\mathrm{v}}} = \frac{s}{\bar{x}}$$

خطای استاندارد یا standard error of mean

با نمونه‌گیری‌های متعدد از یک جامعه، و استفاده از هر نمونه برای تخمین میانگین جامعه، توزیعی از میانگین‌های محاسبه شده به دست خواهد آمد. خطای استاندارد یا standard error of mean انحراف معیار این توزیع است و معیاری‌ست از پراکندگی میانگین نمونه‌ها حول میانگین جامعه.

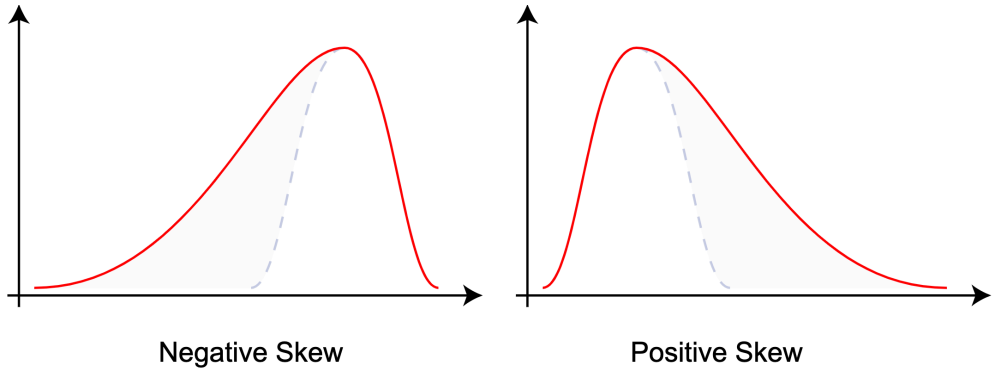
SE or SEM

$$\mathrm{S}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

آماره های توصیف شکل

چولگی

در صورت وجود چولگی در داده ها، میانگین میانه و مد از هم فاصله میگیرند.



توصیف بصری داده‌ها

با استفاده از Box and whiskers plot می توان شهودی کلی از مرکزیت، پراکندگی و شکل داده به دست آورد.

خط میانی: میانه داده‌ها

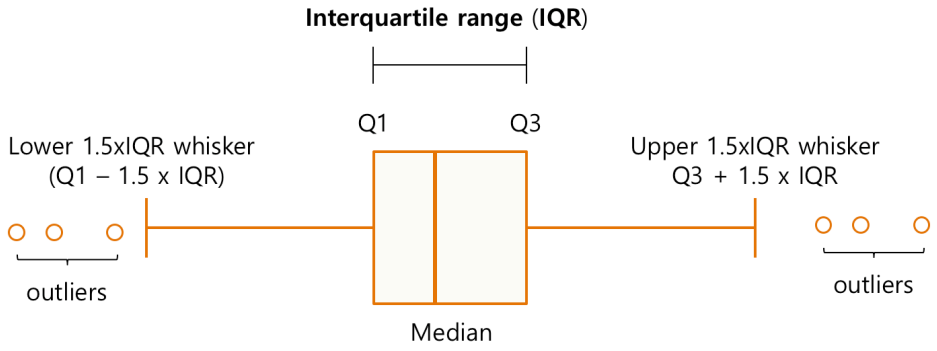
Q1: میانه نیمه پایین مجموعه داده

Q3: میانه نیمه بالای مجموعه داده

IQR: Q3 - Q1

حد بالا و پایین با فاصله یک‌ونیم برابر IQR از Q3 و Q1 به دست می‌آید.

داده‌های خارج از این محدوده به عنوان outlier دسته‌بندی می‌شود.

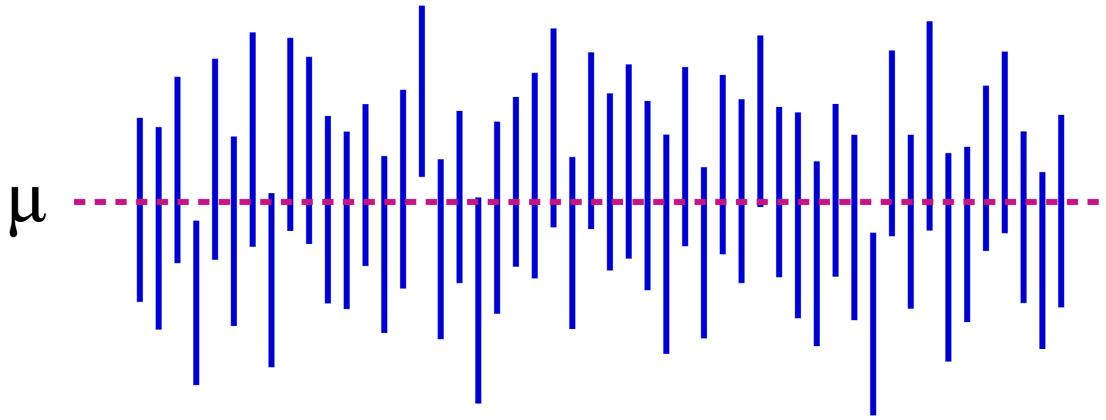


مثال :تخمین میانگین جامعه با استفاده از نمونه‌گیری

پس از اقدام به نمونه‌گیری و محاسبه‌ی آماره‌های نمونه اقدام به تخمین میانگین جامعه می‌کنیم. برای گزارش بازه‌ای حول میانگین نمونه که به احتمال ۹۵ درصد میانگین جامعه را در برداشته باشد، حد بالا و پایین بازه را این‌طور محاسبه می‌کنیم.

حد بالا: $$\bar{x} + (\mathrm{SE} \times 1.96)$$

حد پایین: $$\bar{x} - (\mathrm{SE} \times 1.96)$$



عدد ۱.۹۶ در این مثال با توجه به confidence interval یا CI موردنظر به دست آمده و نقطه‌ای روی نمودار توزیع نرمال استاندارد (با میانگین ۰ و انحراف معیار ۱) است که ۹۵ درصد داده‌ها را در بر دارد.

CI 95%	1.96
CI 99%	2.576
CI 99.9%	3.291

با توجه به داده‌های موجود از وزن چندی از اعضای جامعه به کیلوگرم، بازه میانگین جامعه را با CI 98% به دست آورید.

99	134	119	109	104	93
112	103	150	117	130	124
143	95	146	127	105	110
139	94	135	151	154	140
144	142	98	149	153	116
114	108	100	96	129	145

انواع داده

numerical

هر خوانش داده یک مقدار است.

قد، وزن، نمره، دما و...

categorical

هر خوانش داده دسته‌بندی آن در یک طبقه است. عدد گزارش شده در این نوع داده نه مقدار یک خوانش، بلکه تعداد اعضای یک دسته در تمام نمونه‌گیری است.

nominal

داده‌های categorical کیفی.

جنسیت، زبان مادری و...

ordinal

داده‌های categorical با ارزش کمی، نسبت به هم ترتیب دارند.

میزان تحصیلات، میزان درد از ۱ تا ۱۰

با توجه به نوع داده مورد بررسی، از تست های مختلفی جهت نتیجه‌گیری در مورد آنان استفاده می‌شود.

آمار استنباطی

تحلیل داده موجود برای نتیجه‌گیری در مورد جوامع خواستگاه نمونه‌ها در حوزه آمار استنباطی مطالعه می‌شود. با توجه به نوع داده مورد بررسی، از تست‌های مختلفی جهت نتیجه‌گیری در مورد آن استفاده می‌شود.

هر تست از چند رکن اصلی تشکیل شده است.

فرض صفر یا null hypothesis

گزاره‌ای کلی که اذعان دارد مجموعه داده‌های مورد بررسی فاقد اطلاعات ارزش‌مند است. در صورت وجود شواهد کافی برای رد فرض صفر می‌توان بر وجود اطلاعات جدید در داده استنباط کرد.

آماره یا test statistic

هر تست با انجام محاسبات ویژه خود بر داده‌های مورد بررسی به متغیری نهایی می‌رسد که مقدار آن بر میزان همراستایی داده با فرض صفر دلالت دارد.

مقدار حدی یا critical value

آماره‌های هر تست پیرو توزیع خاصی‌ست که به صورت جدول‌هایی در دسترس همگان است. در صورت حضور آماره در منتهی نقاط توزیع و رد کردن مقداری مرزی می‌توان در حمایت از رد فرض صفر سخن گفت.

p value

احتمال مشاهده پدیده‌ای نوظهور به صورت تصادفی. احتمال به اشتباه رد کردن فرض صفر. هر چه مقدار p value کوچکتر باشد، شواهد بیش‌تری مبنی بر رد فرض صفر موجود است. حد قراردادی آن ۵ درصد است.

F test

از این تست برای بررسی معنی‌دار بودن تفاوت واریانس دو نمونه استفاده می‌شود. نتیجه آن در ادامه روند استنباط و تست‌های مورد استفاده اهمیت دارد. نسبت واریانس بزرگ‌تر به واریانس کوچک‌تر آماره را محاسبه می‌کند.
نکته: df1 تعداد داده‌های X (صورت) منهای یک است.
Df۲ تعداد داده‌های Y (مخرج) منهای یک است.

$$F = \frac{S_X^2}{S_Y^2}$$

T test

۱. One Sample

برای مقایسه یک نمونه با یک جامعه است. با این تست می توانیم بفهمیم نمونه متعلق به جامعه هست یا نه. فرمول آن به شرح زیر است:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

در این جا x̄ میانگین داده‌های نمونه، μ_0 میانگین جمعیت، s انحراف معیار داده‌های نمونه و n تعداد داده‌های نمونه است.درجه آزادی برابر است با ۱ - n.

۲. Two Sample

گاهی اوقات می خواهیم دو نمونه را با هم (و نه جمعیت) مقایسه کنیم:

حالت اول:

Independent Two Sampled t-test (بـا واریانـس یکسان): در صورتی که اختلاف واریانس‌ها از نظر آماری معنادار نباشد (f test)، از فرمول‌های زیر استفاده می کنیم که در آن x̄۱ میانگین داده‌های نمونه اول، x̄۲ میانگین داده‌های نمونه دوم، s_p انحراف معیار تجمیع‌شده، n_1 و n_2 تعداد داده‌های نمونه‌های اول و دوم و s_{X_1} و s_{X_2} انحراف معیار داده‌های نمونه‌های اول و دوم هستند.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

درجه آزادی این تست، برابر است با:

$$n_1 + n_2 - 2$$

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}}$$

حالت دوم:

Independent Two Sample t-test (با واریانس متفاوت) (Welch's t-test):

در صورتی که اختلاف واریانس‌ها از نظر آماری معنادار باشد (f test)، از فرمول زیر استفاده می کنیم که در آن:

$\bar{x}_1$ میانگین داده‌های نمونه اول و $\bar{x}_2$ میانگین داده‌های نمونه دوم است. s_1 و s_2 انحراف معیار داده‌های نمونه‌های اول و دوم و n_1 و n_2 تعداد داده‌های نمونه‌های اول دوم است.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{\Delta}}}$$

$$s_{\bar{\Delta}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

۳. Paired t-test (Dependent t-test)

اگر دو نمونه در حقیقت مربوط به یک نمونه در دو شرایط مختلف باشد، از این تست استفاده می‌کنیم. به زبانی دیگر، نمونه‌گیری‌ها مستقل نیستند. فرمول آن در ادامه آمده است:

$$t = \frac{\bar{X}_D - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

در ابتدا باید کل داده‌های دسته دوم را از داده‌های متناظرشان در دسته اول کم کرده و یک دسته اختلافات بسازیم. $\bar{x}_D$ میانگین داده های دسته اختلافات و s_D انحراف معیار داده‌های دسته اختلافات و n تعداد داده‌های دسته اختلافات را نشان می‌دهد. μ_0 را صفر در نظر می گیریم. سپس با کمک فرمول روبرو، مقدار عددی t به دست می‌آید. درجه آزادی در این جا برابر است با n - ۱ . (n = تعداد داده‌های دسته اختلافات)

نکته: در این تست، همواره تعداد داده‌های دسته اول و دوم برابر است.

نکته: در این تست نمی‌توانیم داده‌های دسته دوم را با هم جابجا کنیم. به زبان دیگر، هر داده یک داده متناظر خودش دارد، مسئله‌ای که برای independent t-test صادق نبود.

مثال: یکی از مسائلی که به معظلی در بین دانش آموزان تبدیل شده است، استفاده از موادی مثل ریتالین در شب امتحان است. یک محقق برای پی بردن به این مسئله که آیا ریتالین واقعا اثرات مثبت بر روی نمره دانش آموزان دارد یا نه، آزمایشی ترتیب داد. به این صورت که یک امتحان از ۸ دانش آموز گرفت و نمرات آن ها را یادداشت کرد. سپس فردای آن روز به آن‌ها مقداری ریتالین داد و یک امتحان در همان سطح از آن‌ها گرفت و نمراتشان را یادداشت کرد. با توجه به جدول زیر، آیا ریتالین به طرز معنا داری باعث بالا رفتن نمره دانش آموزان شده است؟

شماره دانش آموز	نمره دانش آموز از ۱۰۰ در امتحان روز اول	نمره دانش آموز از ۱۰۰ در امتحان روز دوم
۱	۴۷	۴۹
۲	۸۴	۷۳
۳	۱۹	۲۵
۴	۳۴	۳۶
۵	۳۶	۲۹
۶	۴۹	۴۱
۷	۵۸	۶۳
۸	۷۱	۷۰

مثال: گفته می‌شود که تعداد ضربان قلب ورزشکاران نسبت به غیر ورزشکاران، کم‌تر است. شما برای تعیین این که این ادعا حقیقت دارد یا خیر، تعداد ضربان قلب ۱۴ ورزشکار شیرازی در یک دقیقه را گرفته‌اید. داده های آن در جدول زیر آمده است. همچنین از طریق یکی از دوستانتان که قبلا تحقیقی روی تعداد ضربان قلب در شهر شیراز انجام داده بود، می‌دانید که میانگین ضربان قلب در شهر شیراز ۷۷ بار در دقیقه است. حال تعیین کنید که این ادعا حقیقت دارد یا خیر؟

شماره ورزشکار	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴
تعداد ضربان قلب در دقیقه	۸۷	۷۱	۶۹	۷۵	۷۷	۸۲	۷۵	۵۹	۷۰	۶۸	۶۶	۶۷	۷۸	۹۰

مثال: اعداد زیر مربوط به طول دم دو زیرگونه سهره است که در دو اقلیم متفاوت زیست می‌کنند. با توجه به این دو نمونه ۲۶تایی، آیا تفاوت معناداری بین طول دم‌های این دو زیرگونه وجود دارد؟

1												
88	44	66	57	64	71	93	82	56	51	70	55	80
94	79	81	46	77	63	59	50	95	85	52	67	62
2												
71	40	44	35	34	65	27	63	62	39	70	50	29
51	38	52	45	68	64	59	33	57	47	43	54	32

آنُوا ANOVA

برای مقایسه‌ی معنا دار بودن میانگین بیش از دو گروه به به کار میرود.
در آنالیز m گروه که هر کدام حاوی n داده هستند مراحل زیر را طی می‌کنیم:

واریانس بین گروهی	$SS_{between} = \Sigma\, n_j\, (\overline{X}_j - \overline{X})^2$
Df بین گروهی	$df_{between} = m - 1$
میانگین واریانس بین گروهی	$MS_{between} = \dfrac{SS_{between}}{df_{between}}$
واریانس درون گروهی	$SS_{within} = \Sigma\, (X_i - \overline{X}_j)^2$
Df درون گروهی	$df_{within} = n - m$
میانگین واریانس درون گروهی	$MS_{within} = \dfrac{SS_{within}}{df_{within}}$
آماره‌ی نهایی	$F = \dfrac{MS_{between}}{MS_{within}}$

* واریانس بین گروهی: مجموع مربعات اختلافات میانگین هر گروه با میانگین کل
* واریانس درون گروهی: مجموع مربعات اختلافات میانگین هر داده با میانگین گروه مربوطه

نکته: برای تحلیل آماره از جدول F و Df بین گروهی و درون گروهی استفاده می‌شود.
نکته: معنی دار بودن تست ANOVA تنها به این معناست که اختلاف معنی‌داری در بین گروه های مورد بررسی دیده می‌شود. برای تشخیص اینکه میانگین کدام دو گروه با هم اختلاف معنی دار دارد باید از تست های post hoc استفاده شود.

مثال :در یک clinical trial میزان افزایش جریان خون پس از تزریق داروی مربوطه در سه گروه از بیماران اندازه گیری شد. آیا تفاوت معنی داری در بین این سه گروه مشاهده می‌شود؟

Increase in Peak Flow		
Placebo (PL)	Epinephrine (EPI)	Albuterol(ALB)
35	71	77
40	75	70
35	80	60
30	90	80
50	45	85
20	65	90

تست مربع کای Chi Squared

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

بـرای اسـتنباط در مـورد داده categorical اسـتفاده می‌شود. مقایسه مقادیر مشاهده شده با میزان مورد انتظار، آماره را مشخص می‌کند.
در فرمول x_i مقدار مشاهده شده در i امین category و m_i مقدار مورد انتظار در همان category است.

Goodness of fit

با بررسی یک مجموعه داده‌ی nominal و مقایسه مقادیر مشاهده شده در هر category با مقدار مورد انتظار در صورت برقرار بودن شرایطی ابتدایی، آماره را محاسبه می‌کنیم. مقدار بیش‌تر آن نشان‌دهنده مغایرت بین مقادیر مشاهده شده و مورد انتظار است.

مثال: مقادیر زیر در ژنوتیپ های یک ژن دو اللی مشاهده شده است. آیا تعادل هاردی واینبرگ برقرار است؟

AA	۳۶۹
Aa	۹۵۳
aa	۶۹۱

نکته: df در تست هاردی‌واینبرگ همواره برابر با ۱ است.

Test of independence

با مقایسه‌ی هم‌زمان دو داده categorical در مورد مستقل بودن آن اظهار نظر می‌کند. داده مورد بررسی contingency table است.

	Smoking +	Smoking -	Total
Lung cancer +	A	B	a+b
Lung cancer -	C	D	c+d
Total	a+c	b+d	a+b+c+d

برای محاسبه مقادیر مورد انتظار از فرمول زیر استفاده می‌کنیم.

$$E(i,j) = \frac{\text{(row total)(column total)}}{\text{total number surveyed}}$$

	Smoking +	Smoking -	Total
Lung cancer +	(a+b).(a+c)/(a+b+c+d)	(a+b).(b+d)/(a+b+c+d)	a+b
Lung cancer -	(c+d).(a+c)/(a+b+c+d)	(c+d).(b+d)/(a+b+c+d)	c+d
Total	a+c	b+d	a+b+c+d

نکته: df در این تست برابر با $(m-1)(n-1)$ است که هر کدام تعداد category در یکی از دو متغیر است.

مثال: آیا جنسیت و ترجیح حزب از هم مستقل‌اند؟

	Voting Preferences		
	Rep	**Dem**	**Ind**
Male	200	150	50
Female	250	300	50

Fisher's exact test of independence

در تحلیل contingency table های ۲ در ۲ و زمانی که خانه‌ای با مقدار کمتر از ۵ داشته باشیم استفاده می‌شود.
نکته: آماره این تست p value است.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

مثال: آیا طبق جدول زیر رابطه معنی‌دار بین مبتلا شدن به ویروس و بروز سرطان وجود دارد؟

	Viral infection +	Viral infection -
Cancer +	9	1
Cancer -	3	11

تست‌های ناپارامتری

در صورت عدم پیروی داده از توزیع نرمال یا عدم اطلاع در مورد توزیع آن از تست‌های Non parametric استفاده می‌کنیم. در این تست‌ها از پارامترهای توصیف داده استفاده نمی‌شود و تنها رنک داده در بین مجموعه مورد بررسی است. مقداری از اطلاعات موجود در داده در این فرآیند از دست می‌رود.

Mann Whitney U test

معادل independant t test در تست‌های ناپارامتری‌ست.
در ابتدا داده‌های هر دو گروه تجمیع شده و به ترتیب مقدار مرتب می‌شود سپس رنک هر داده مشخص می‌شود. در صورتی که یک داده چند بار تکرار شده باشد میانگین رنک آن‌ها به تمام آن‌ها اطلاق می‌شود.

به عنوان مثال

داده	۱	۳	۵	۵	۵	۹	۹
رنک اولیه	۱	۲	۳	۴	۵	۶	۷
رنک تصحیح شده	۱	۲	۴	۴	۴	۶.۵	۶.۵
گروه	A	A	B	A	A	B	B

آماره برای داده های هر گروه محاسبه می‌شود.

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

که R جمع رنک پس از تصحیح و n تعداد نمونه در گروه است.
از بین این دو مقدار کوچک‌تر به عنوان آماره نهایی در نظر گرفته می‌شود.

مثال: تعداد مراجعه به اورژانس در ماه گذشته در دو گروه سنجیده شد. گروه اول داروی Z را مصرف کرده‌اند و گروه دوم این دارو را مصرف نکرده‌اند. آیا اختلاف این دو گروه معنی‌دار است؟

	مریض ۱	مریض ۲	مریض ۳	مریض ۴	مریض ۵	مریض ۶
گروه اول	۰	۰	۰	۱	۰	۱
گروه دوم	۳	۲	۰	۳	۲	۱