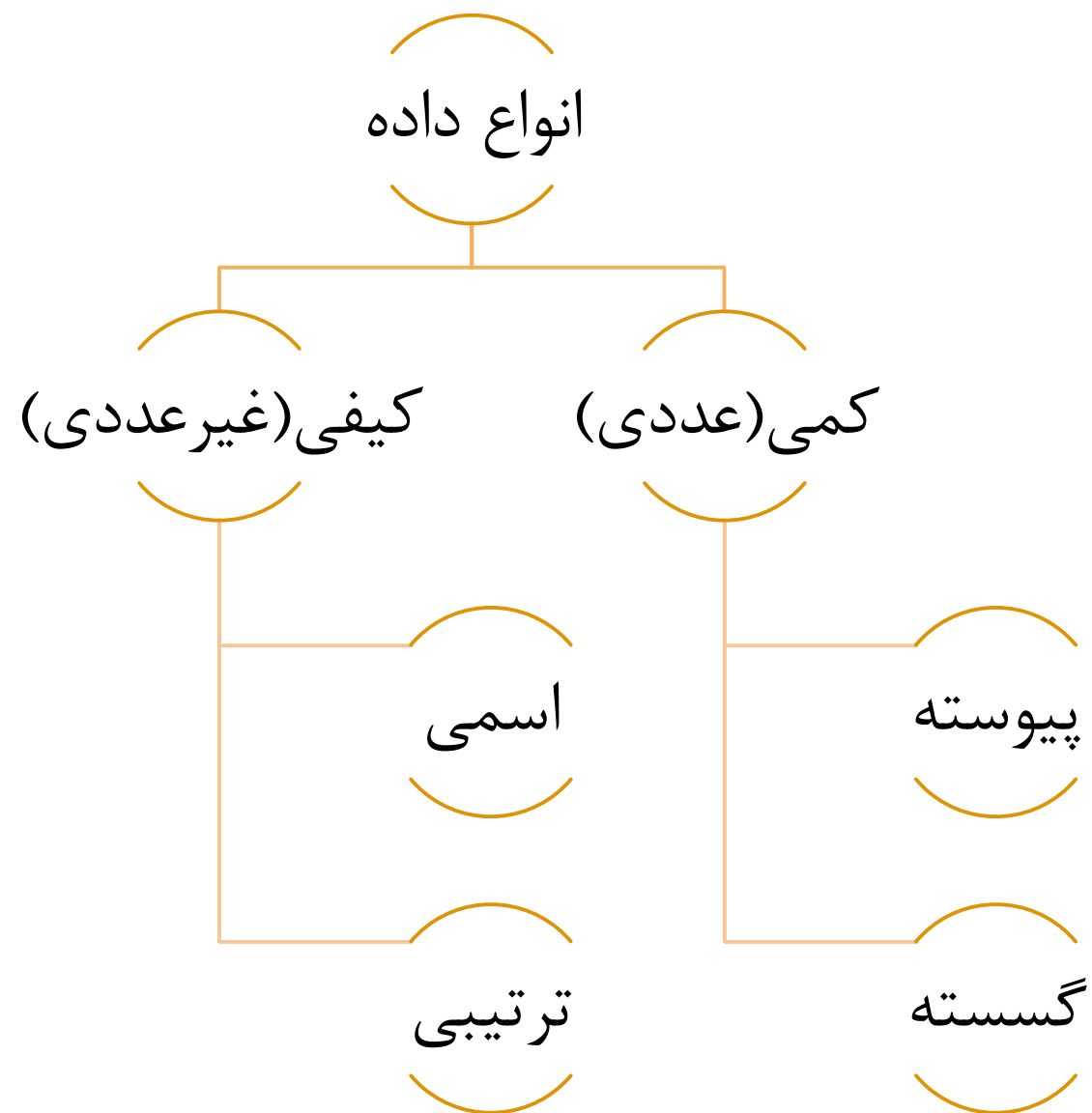
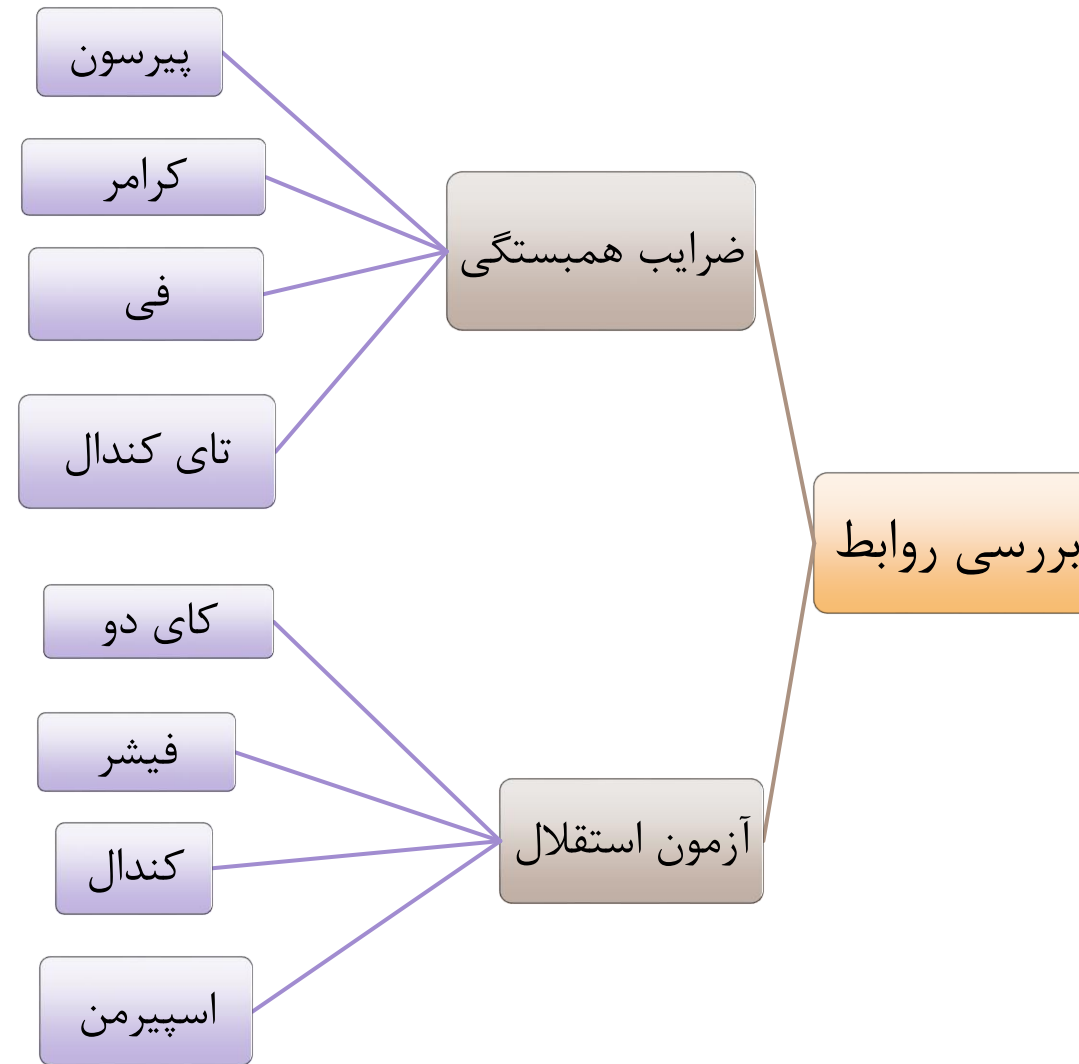


ضروریات آمار برای علم داده

وابستگی بین متغیرها










آزمون استقلال کای دو

▶ دو متغیر گسسته

▶ یک متغیر گسسته و یک متغیر رتبه‌ای

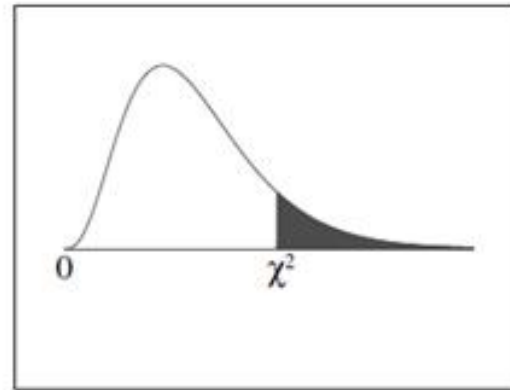
▶ فراوانی هر یک از مقادیر جدول توافقی بالاتر از ۵ باشد

	 Cake	 Ice	 Donut	Total
 Female	4	3	6	13
 Male	5	7	9	21
Total	9	10	15	34

$$\sum_{ij} = \frac{R_i \times C_j}{N} \quad \sum_{i,j=1}^n = \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad df = (r - 1) \times (c - 1)$$

Intervention	Recycles	Does not recycle	Row totals
Flyer (pamphlet)	89	9	98
	$\frac{(98 \times 259)}{300} = 84.61$	$\frac{(98 \times 41)}{300} = 13.39$	
Phone call	84	8	92
	$\frac{(92 \times 259)}{300} = 79.43$	$\frac{(92 \times 41)}{300} = 12.57$	
Control	86	24	110
	$\frac{(110 \times 259)}{300} = 94.97$	$\frac{(110 \times 41)}{300} = 15.03$	
Column totals	259	41	$N = 300$

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_{\alpha}$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

جدول زیر میزان آسیب وارد شده بر اثر تصادف به سرنشینان خودرو را نشان می‌دهد. آیا میزان صدمات وارده به مسافران مستقل از بستن کمربند است؟

میزان آسیب بستن کمربند	هیچ	کم	متوسط	زیاد
بله	۱۲۸۱۳	۶۴۷	۳۵۹	۴۲
خیر	۶۵۹۶۳	۴۰۰۰	۲۶۴۲	۳۰۳

آزمون استقلال فیشر

	Men	Women	Row Total
Studying	<i>a</i>	<i>b</i>	<i>a + b</i>
Non-studying	<i>c</i>	<i>d</i>	<i>c + d</i>
Column Total	<i>a + c</i>	<i>b + d</i>	<i>a + b + c + d (=n)</i>

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

با توجه به داده‌های جدول زیر آیا سیگاری شدن فرزندان مستقل از والدین آنهاست؟

والدین سیگاری	فرزندان سیگاری	
	بله	خیر
بله	۷	۲
خیر	۴	۸

دو متغیر اسمی

	$y = 1$	$y = 0$	total
$x = 1$	n_{11}	n_{10}	$n_{1\bullet}$
$x = 0$	n_{01}	n_{00}	$n_{0\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 0}$	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

ضریب همخوانی فی

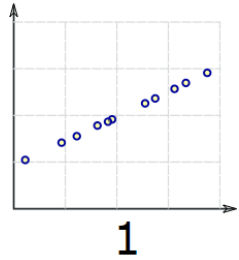
درجه قدرت رابطه را نشان می‌دهد. (بین ۰ تا ۱ است)
زمانی استفاده می‌شود که جدول توافقی داده‌ها ۲ در ۲ باشد.

ضریب همخوانی کرامر

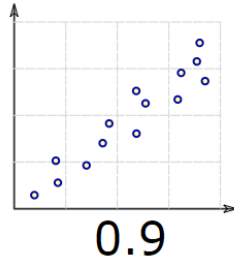
میزان ارتباط دو متغیر را نشان می‌دهد. (بین ۰ تا ۱ است)
برای هر تعداد سطر و ستون جدول توافقی به کار می‌رود.

دو متغیر مقیاسی یا فاصله‌ای

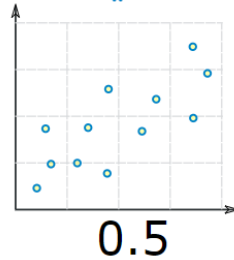
همبستگی
مثبت
کامل



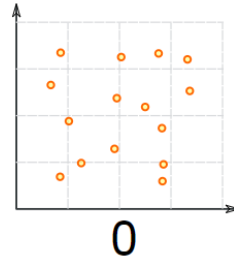
همبستگی
مثبت
کامل



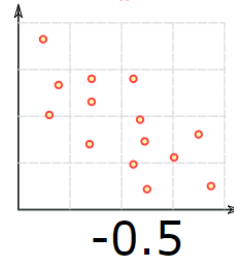
همبستگی
مثبت
ضعیف



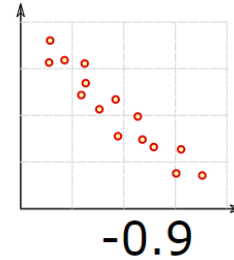
بدون
همبستگی



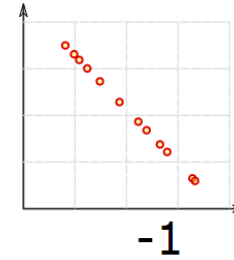
همبستگی
منفی
ضعیف



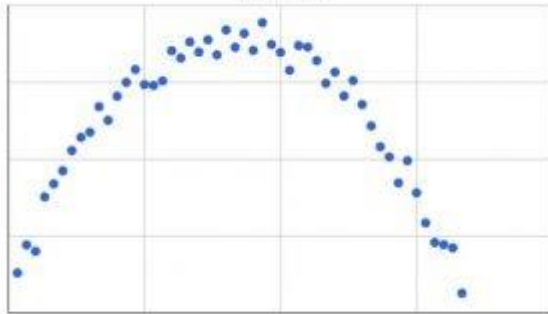
همبستگی
منفی
قوی



همبستگی
منفی
کامل



رابطه غیر خطی



دو متغیر مقیاسی یا فاصله‌ای

$$\begin{cases} H_0 : \rho_s = 0 \\ H_1 : \rho_s \neq 0 \end{cases}$$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

برای آزمون فرض صفر این که بین متغیرهای
فاصله‌ای نرمال رابطه خطی وجود دارد یا نه

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

تعداد سال‌های خدمت و میزان درآمد تعدادی از پاسخ‌گویان یک پژوهش به صورت جدول زیر ارائه شده است. ضریب همبستگی پیرسون این داده‌ها را محاسبه کنید.

دو متغیر رتبه‌ای

$$\tau_A = \frac{n_c - n_d}{n_0}$$

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

$$\tau_C = \frac{2(n_c - n_d)}{n^2 \frac{(m-1)}{m}}$$

ضریب همبستگی تای آ کندال

تعداد سطر و ستون دو متغیر برابر باشد

تعداد مقوله‌های دو متغیر زیاد باشد (بیشتر از ۷ باشد)

اندازه دو نمونه زیاد باشد

بین ۱- تا ۱+ است

ضریب همبستگی تای بی کندال

تعداد سطر و ستون دو متغیر برابر باشد

تعداد مقوله‌های دو متغیر کم باشد (بین ۳ تا ۷ باشد)

بین ۱- تا ۱+ است

ضریب همبستگی تای سی کندال

تعداد سطر و ستون دو متغیر برابر نباشد

تعداد مقوله‌های دو متغیر کم باشد (بین ۳ تا ۷ باشد)

بین ۱- تا ۱+ است

آزمون کندال

$$R_i = \sum_{j=1}^m r_{i,j} \quad \bar{R} = \frac{1}{n} \sum_{i=1}^n R_i$$

$$S = \sum_{i=1}^n (R_i - \bar{R})^2, \quad W = \frac{12S}{m^2(n^3 - n)}$$

آزمون اسپیرمن

رابطه خطی و غیر خطی بین دو متغیر را بررسی می‌کند
نیاز به فرض نرمال بودن ندارد

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$