

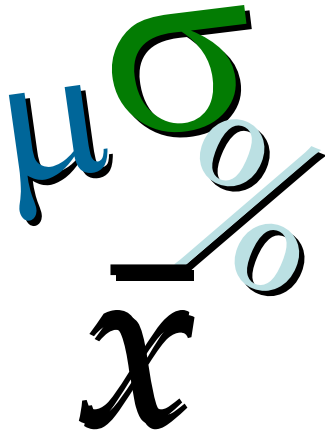
آمار توصیفی

قسمت دوم: روشهای عددی

آمار توصیفی

قسمت دوم: روشهای عددی

- معیارهای نسبی مکانی و شناسایی داده های پرت
- بررسی تحلیل داده ها
- اندازه ارتباط بین دو متغیر
- میانگین وزنی و کار با داده های گروهی



معیارهای نسبی مکانی و شناسایی داده های پرت

- امتیازات Z
- قضیه چیبیشف
- قاعده تجربی
- شناسایی داده های پرت

امتیازات Z

- امتیازات Z را اغلب مقادیر استاندارد شده می نامند.
- میزان انحراف معیار مقدار یک داده X_i از میانگین نشان می دهد.

$$Z_i = \frac{X_i - \bar{X}}{S}$$

- یک مقدار داده کمتر از میانگین نمونه، مقدار Z کمتر از صفر را خواهد داشت.
- یک مقدار داده بیشتر از میانگین نمونه، مقدار Z بیشتر از صفر را خواهد داشت.
- یک مقدار داده مساوی با میانگین نمونه، مقدار Z صفر را خواهد داشت.

مثال: کرایه های آپارتمان

- امتیاز z کوچکترین مقدار (425)

$$z = \frac{x_i - \bar{x}}{s} = \frac{425 - 490.80}{54.74} = -1.20$$

مقادیر استاندارد شده برای کرایه های آپارتمان

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

قضیه چبیشف

حداقل $(1 - 1/z^2)$ از داده ها در محدوده انحراف معیار z از میانگین نخواهد بود، در حالیکه z هیچ مقدار بزرگتر از 1 را ندارد.

– حداقل 75٪ از موارد باید در محدوده 2 انحراف معیار از میانگین $z =$ باشند.

– حداقل 89٪ از موارد باید در محدوده 3 انحراف معیار از میانگین $z =$ باشند.

– حداقل 94٪ از موارد باید در محدوده 4 انحراف معیار از میانگین $z =$ باشند.

مثال: کرایه های آپارتمان

- قضیه چیبیشف

فرض کنید $z = 1.5$ با $\bar{x} = 490.80$ و $s = 54.74$

حداقل $(1 - 1/(1.5)^2) = 1 - 0.44 = 0.56$ یا 56%

از مقادیر کرایه ها باید در محدوده های زیر باشند:

$$\bar{x} - z(s) = 490.80 - 1.5(54.74) = 409$$

و

$$\bar{x} + z(s) = 490.80 + 1.5(54.74) = 573$$

مثال: کرایه های آپارتمان

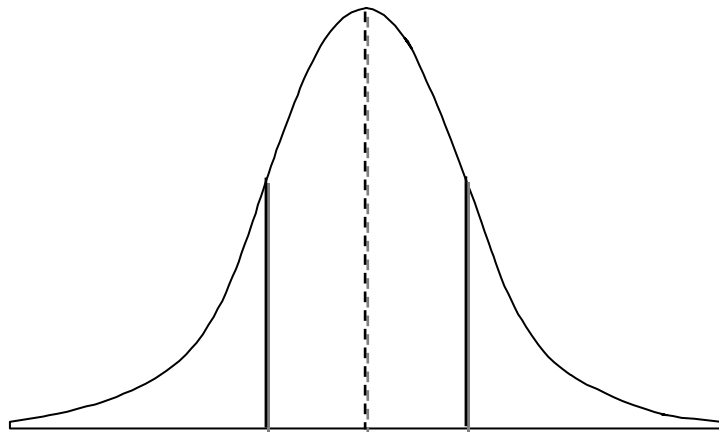
- قضیه چیبیشف (ادامه)

در واقع ، 86٪ از مقادیر اجاره بین 409 و 573 است.

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

قاعده تجربی

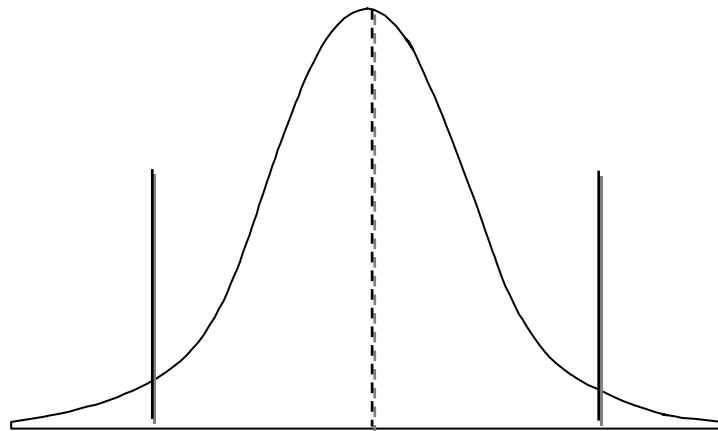
برای داده هایی که توزیع زنگوله ای شکل دارند:



تقریباً 68% از مقادیر داده در محدوده یک انحراف معیار از میانگین خواهند بود.

قاعده تجربی

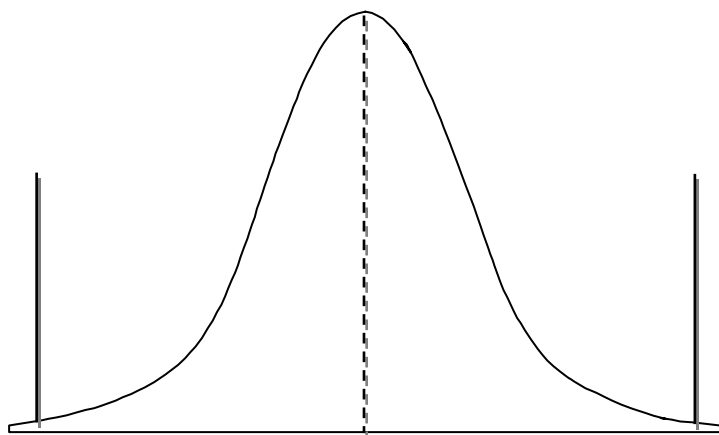
برای داده هایی که توزیع زنگوله ای شکل دارند:



تقریباً 95% از مقدار داده ها در محدوده دو انحراف معیار از میانگین خواهند بود.

قاعده تجربی

برای داده هایی که توزیع زنگوله ای شکل دارند:



تقریباً تمام داده ها (99.7%) در محدوده سه انحراف معیار از میانگین خواهند بود.

مثال: کرایه های آپارتمان

- قاعده تجربی

	فاصله	% در فاصله
$\pm 1s$ محدوده	43.06 to 545.54	$48/70 = 69\%$
$\pm 2s$ محدوده	381.32 to 600.28	$68/70 = 97\%$
$\pm 3s$ محدوده	326.58 to 655.02	$70/70 = 100\%$

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

شناسایی داده های پرت

- یک داده پرت یک مقدار کوچک یا بزرگ غیر معمول در یک سری داده ها است.
- یک مقدار داده با مقدار Z کمتر از -3 یا بیشتر از $+3$ ممکن است یک داده پرت در نظر گرفته شود.
- داده پرت ممکن است یکی از موارد زیر باشد:
 - مقدار داده ای که به اشتباه ثبت شده است.
 - مقدار داده ای که به غلط در مجموعه داده ها گنجانده شده است.
 - مقدار داده ای که درست ثبت شده و متعلق به مجموعه داده هاست.

مثال: کرایه های آپارتمان

- شناسایی داده های پرت

مقادیر انتهایی z (2.27 و -1.20) می باشد.

$|z| \geq 3$ را به عنوان معیاری برای داده های پرت بکار می بریم، هیچ داده پرتی در این مجموعه داده ها وجود ندارد.

انحراف معیار برای کرایه های آپارتمان

-1.20	-1.11	-1.11	-1.02	-1.02	-1.02	-1.02	-1.02	-0.93	-0.93
-0.93	-0.93	-0.93	-0.84	-0.84	-0.84	-0.84	-0.84	-0.75	-0.75
-0.75	-0.75	-0.75	-0.75	-0.75	-0.56	-0.56	-0.56	-0.47	-0.47
-0.47	-0.38	-0.38	-0.34	-0.29	-0.29	-0.29	-0.20	-0.20	-0.20
-0.20	-0.11	-0.01	-0.01	-0.01	0.17	0.17	0.17	0.17	0.35
0.35	0.44	0.62	0.62	0.62	0.81	1.06	1.08	1.45	1.45
1.54	1.54	1.63	1.81	1.99	1.99	1.99	1.99	2.27	2.27

بررسی تحلیل داده ها

- خلاصه های پنج عددی
- نمودار جعبه ای

خلاصه های پنج عددی

- کوچکترین مقدار
- چارک اول
- میانه
- چارک سوم
- بزرگترین مقدار

مثال: کرایه های آپارتمان

- خلاصه های پنج عددی

کمترین مقدار = 425

چارک اول = 450

میانه = 475

چارک سوم = 525

بیشترین مقدار = 615

425	430	430	435	435	435	435	435	440	440
440	440	440	445	445	445	445	445	450	450
450	450	450	450	450	460	460	460	465	465
465	470	470	472	475	475	475	480	480	480
480	485	490	490	490	500	500	500	500	510
510	515	525	525	525	535	549	550	570	570
575	575	580	590	600	600	600	600	615	615

نمودار جعبه ای

- جعبه بین چارک اول و سوم کشیده شده است.
- یک خط عمودی روی جعبه در محل میانه کشیده شده است.
- محدوده ها با استفاده از دامنه میان چارکی (IQR) تعیین می شوند (رسم نمی شوند).
- حد پایین با استفاده از (IQR) 1.5 پایین چارک اول تعیین می شود.
- حد بالا با استفاده از (IQR) 1.5 بالای چارک سوم تعیین می شود.
- داده های خارج از این محدوده داده های پرت در نظر گرفته می شوند.
- ریشها (خطهای تیره) از انتهای جعبه به کوچکترین و بزرگترین مقادیر داده در کنار حدود رسم شده اند.
- محل هر داده پرت با علامت * نشان داده شده است.

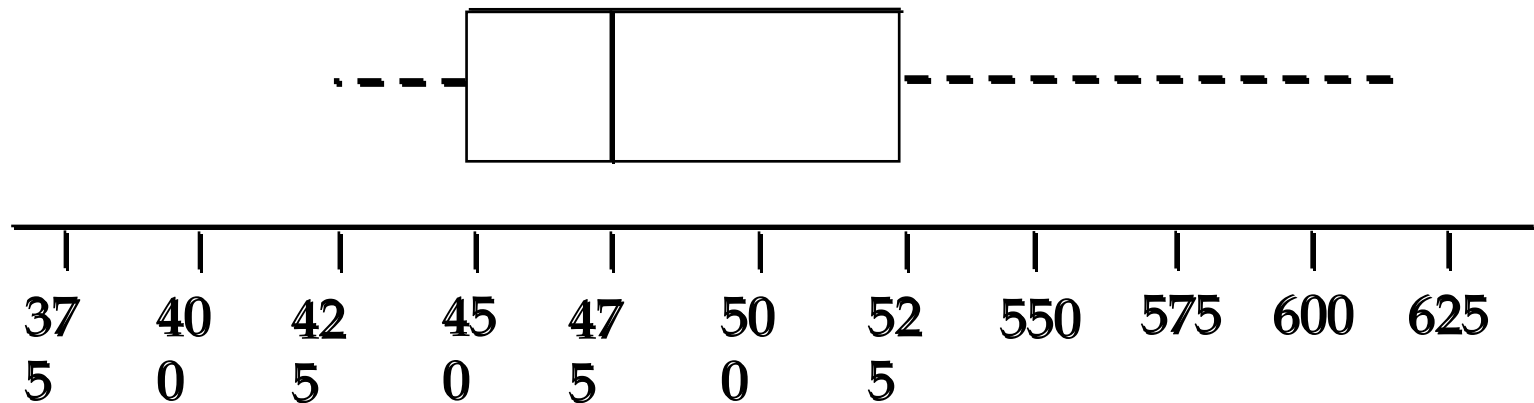
مثال: کرایه های آپارتمان

• نمودار جعبه ای

$$\text{حد پایین} : Q1 - 1.5(IQR) = 450 - 1.5(75) = 337.5$$

$$\text{حد بالا} : Q3 + 1.5(IQR) = 525 + 1.5(75) = 637.5$$

داده پرتی وجود ندارد.



اندازه ارتباط بین دو متغیر

- کواریانس
- ضریب همبستگی

کواریانس

- کواریانس میزان ارتباط خطی بین دو متغیر است.
- مقادیر مثبت همبستگی مثبت را نشان می دهند.
- مقادیر منفی همبستگی منفی را نشان می دهد.

کواریانس

- اگر مجموعه داده ها نمونه باشند، کواریانس با S_{xy} نشان داده می شود.

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- اگر مجموعه داده ها جامعه باشند، کواریانس با σ_{xy} نشان داده می شود.

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

ضریب همبستگی

- ضریب همبستگی می تواند مقادیر بین -1 و $+1$ را بگیرد.
- مقادیر نزدیک -1 یک همبستگی خطی منفی شدید را نشان می دهد.
- مقادیر نزدیک $+1$ یک همبستگی خطی مثبت شدید را نشان می دهد.
- اگر مجموعه داده ها نمونه باشند، ضریب همبستگی r_{xy} است.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- اگر مجموعه داده ها جامعه باشند، ضریب همبستگی ρ_{xy} است.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

میانگین وزنی و کار با داده های گروهی

- میانگین وزنی
- میانگین داده های گروهی
- واریانس داده های گروهی
- انحراف معیار داده های گروهی

میانگین وزنی

- زمانی که میانگین با تاثیر وزنی که هر یک از مقادیر می گیرند محاسبه می شود، میانگین وزنی نامیده می شود.
- در محاسبه میانگین نقطه پایه (GPA)، وزنها تعداد فواصل زمانی اطمینانی هستند که برای هر پایه بدست می آیند.
- زمانی که مقادیر داده ها به صورت با معنایی تغییر می کند، تحلیل گر باید وزنی را انتخاب کند که بهترین تاثیر را در ارزش هر مقدار داشته باشد.

میانگین وزنی

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

که

x_i = مقدار مشاهده i ام

w_i = وزن مشاهده i ام

داده های گروهی

- محاسبه میانگین وزنی می تواند برای بدست آوردن برآورد میانگین، واریانس و انحراف معیار داده های گروهی به کار رود.
- برای محاسبه میانگین وزنی، نقطه میانی هر رده را به عنوان میانگین تمام بخشهای رده به حساب می آوریم.
- در محاسبه میانگین وزنی نقاط میانی رده، از فراوانیهای رده به عنوان وزن استفاده می کنیم.
- بطور مشابه، در محاسبه واریانس و انحراف معیار فراوانیهای رده به عنوان وزنها استفاده می شود.

میانگین داده های گروهی

- داده های نمونه

$$\bar{x} = \frac{\sum f_i M_i}{\sum f_i}$$

- داده های جامعه

$$\mu = \frac{\sum f_i M_i}{N}$$

که:

f_i = فراوانی رده i

M_i = نقطه میانی رده i

مثال: کرایه های آپارتمان

داده های زیر کرایه های ماهانه آپارتمانهای یک خوابه در نمونه قبلی است که در اینجا به عنوان داده های گروهی از توزیع فراوانی نشان داده شده است.

اجاره	فراوانی
420-439	8
440-459	17
460-479	12
480-499	8
500-519	7
520-539	4
540-559	2
560-579	4
580-599	2
600-619	6

مثال: کرایه های آپارتمان

- میانگین داده های گروهی

اجاره	f_i	M_i	$f_i M_i$
420-439	8	429.5	3436.0
440-459	17	449.5	7641.5
460-479	12	469.5	5634.0
480-499	8	489.5	3916.0
500-519	7	509.5	3566.5
520-539	4	529.5	2118.0
540-559	2	549.5	1099.0
560-579	4	569.5	2278.0
580-599	2	589.5	1179.0
600-619	6	609.5	3657.0
Total	70		34525.0

$$\bar{x} = \frac{34,525}{70} = 493.21$$

این تقریب 2.41 از

میانگین واقعی 490.80

تفاوت دارد.

واریانس داده های گروهی

- داده های نمونه

$$s^2 = \frac{\sum f_i (M_i - \bar{x})^2}{n-1}$$

- داده های جامعه

$$\sigma^2 = \frac{\sum f_i (M_i - \mu)^2}{N}$$

مثال: کرایه های آپارتمان

- واریانس داده های گروهی

$$s^2 = 3,017.89$$

- انحراف معیار داده های گروهی

$$s = \sqrt{3,017.89} = 54.94$$

این تقریب فقط 0.20 از مقدار واقعی انحراف معیار یعنی 54.74 متفاوت است.

پایان، قسمت دوم

