



به نام خدا

تحلیل آماری توصیفی : تحلیل انشائی داده ها

Descriptive Statistics: Exploratory Analysis Data

UNIVERSITY OF SISTAN AND BALUCHESTAN

Department of Statistics, Dr. Dehghan

مهارت‌های لازم و مفید برای دانشجویان

- (1) نرم افزار R ، SPSS و Minitab
- (2) نرم افزار Word و یا Excel
- (3) نرم افزار زی - پرشین یا X-Persian
- (4) توصیه های اساتیدتان را جدی بگیرید

فهرست مطالب

- 1) مقدمه و منابع داده‌های میدانی، ثبتی، فراداده/متاداده، کلان‌داده
- ۲) داده (انواع متغیر و سطوح سنجش (اسمی/ترتیبی/فاصله‌ای/نسبتی)
- ۳) جدول‌ها و نمودارها (جدول فراوانی، هیستوگرام، ساقه‌وبرگ، نمودار جعبه‌ای، منحنی فراوانی)
- ۴) شاخص‌های گرایش مرکزی (میانگین، میانه، نما)
- ۵) شاخص‌های پراکندگی (دامنه، IQR، واریانس، انحراف معیار، ضریب تغییرات)
- ۶) شکل توزیع (چولگی، کشیدگی) و ناهنجاری‌ها (دورافتاده‌ها)
- ۷) توزیع تجربی (ECDF و منحنی نرمال به‌عنوان تقریب
- ۸) EDA: جریان کار، پاکسازی و داده‌های گم‌شده، مستندسازی و بهترین شیوه‌ها

تقسیم بندی آمار

- (1) آمار (Statistics): علم جمع آوری، تحلیل، تفسیر، ارائه و سازماندهی داده‌هاست. این علم به ما کمک می‌کند تا از داده‌های پراکنده و پیچیده، نتیجه‌گیری‌های معنادار و تصمیم‌های منطقی بگیریم.
- (2) آمار توصیفی Descriptive Statistics به توصیف جامعه (با شاخص یا نمودار) می‌پردازد و در مورد تعمیم اطلاعات به جامعه اظهار نظر نمی‌کند.
- (3) آمار استنباطی Inferential Statistics هدف آن برآورد، فاصله اطمینان، تعمیم نتایج به جامعه و آزمون فرضیه‌هاست.
- (1) آمار استنباطی به آمار پارامتری، ناپارامتری، بیزی، کلاسیک، پیشرفته یا تعمیم یافته تقسیم بندی می‌شود.

فصل اول : آمار توصیفی

Descriptive Statistics

➤ تعاریف اولیه Basic definitions

➤ **جامعه آماری Statistical population** : به مجموعه ای از اشیاء، یا افراد گفته می شود که حداقل یک ویژگی مشترک آنها مورد مطالعه قرار می گیرد.

➤ مثال : جامعه دانشجویان دانشگاه سیستان و بلوچستان در سال 1403، جامعه درختان یک باغستان، جامعه لامپهای تولیدی کارخانه.

➤ **نمونه آماری Statistical sample** : بخشی از جامعه آماری را که به روش خاصی انتخاب می شود، نمونه آماری گویند. انتظار می رود نمونه خواص جامعه را داشته باشد، در اینصورت نمونه را یک نمونه خوب نامند.

فصل اول: آمار توصیفی Descriptive Statistics

تعاریف اولیه Basic definitions

- **صفت آماری Statistical adjective:** هر مشخصه از جامعه آماری را یک صفت آماری گویند که مورد بررسی قرار می گیرد.
- **متغیر کمی Quantity variable:** متغیر برای اندازه گیری صفت آماری بکار می رود. که معمولاً قابل اندازه گیری هستند و می توان مقدار آنها را به صورت عددی نمایش داد. و مقدار آن از فردی به فرد دیگر ممکن است متفاوت باشد.

فصل اول : آمار توصیفی

Descriptive Statistics

➤ صفت آماری مثل: قد، وزن، دندانهای تعمیر شده یک فرد، اعضای خانوار ...

➤ متغیر کمی Quantity variable : به دو دسته تقسیم می شوند.

1. متغیر گسسته Discrete variable : متغیرهایی که بین دو مقدار متصور آنها هیچ عدد دیگری وجود نداشته باشد. تعداد اعضای خانواده، تعداد کپی

2. متغیر پیوسته Continuous variable : متغیرهایی که بین هر دو مقدار متصور آنها همواره عددی دیگری وجود دارد. قد، وزن

➤ متغیر کیفی Quality variable : متغیرهایی که مستقیماً توسط اعداد و ارقام قابل اندازه گیری نیستند. مانند گروه خون افراد A,B,AB,O یا ملیت افراد، رنگ سبب درختی

مقیاس‌های اندازه گیری در آمار Measurement scales in statistics

➤ مقیاس اندازه گیری چیست؟

➤ آنچه در هر مطالعه مورد توجه قرار می گیرد، اندازه گیری ویژگی ها یا صفات (اشخاص و اشیا) است که بوسیله اعداد یا نمرات نسبت داده می شود.

➤ مثلاً وقتی برای سنجش نگرش مردم نسبت به یک مجموعه تلویزیونی از پرسشنامه استفاده می کنیم، برای تعیین نمره ی هر شخص (نظر) به برنامه عدد یا اعدادی را در نظر می گیریم. بسته به اینکه چه چیزی و چگونه اندازه گیری می نمایم اعداد به دست آمده ویژگی های متفاوتی خواهند داشت. این ویژگیهای متفاوت اعداد را مقیاس های اندازه گیری می نامیم.

انواع مقیاس های اندازه گیری All Types of measurement scales

➤ انواع مقیاس های اندازه گیری

➤ مقیاس های اندازه گیری را در چهار سطح می توان دسته بندی کرد:

➤ اسمی

➤ ترتیبی

➤ نسبی

➤ فاصله ای

انواع مقیاس های اندازه گیری All Types of measurement scales

- **مقیاس اسمی (برچسب)**
- اندازه گیری اسمی یعنی جای دادن اشیا یا افراد در طبقاتی است که از نظر کیفی با هم متفاوت هستند. (درون گروه یکسان یا بسیار نزدیک به هم)
- یک مثال ساده در این مورد متغیر جنسیت است. در سطح اندازه گیری اسمی استفاده از اعداد تنها به منظور مشخص کردن طبقات می باشد و هیچ رابطه ی مقایسه ای که با اعمال ریاضی مطابقت کند بین آن ها وجود ندارد.
- اعداد به کار رفته در یک مقیاس اسمی معرف مقدار مطلق یا نسبی هیچ ویژگی نیستند.
- مثال رنگ، گروه خونی افراد- نام شهرها، کد ملی
- برای مقیاس میانگین بی معناست.

انواع مقیاس های اندازه گیری All Types of measurement scales

- **مقیاس ترتیبی (رتبه پذیرند)**
- مقیاس های ترتیبی برای ما امکان رتبه بندی و ترتیب سطوح متغیر مورد بررسی را فراهم می آورند.
- مثال 10 برای بررسی میزان رضایت، مهارت یک شخص، از متخصصین می خواهیم این ویژگی یا صفت را برحسب درجه اهمیت آن از 1 تا 5 رتبه بندی کنند.
- یا میزان تحصیلات را با اعداد 1 تا 5 را معمولاً رتبه بندی می کنیم
- سه نفر اول دومی دانی، سه نفر اول ورودی رشته آمار 1404 دانشگاه س ب
- با این حال یکی از مشکلات مقیاس ترتیبی آن است که هرچند محقق می تواند تفاوت های مورد نظر را برحسب بزرگتر- کوچکتر بودن، بهتر- بدتر بودن و غیره رتبه کند ولی قادر نیست فاصله ی این تفاوت ها را بدست آورد. به عبارت دیگر مقادیر حاصل در مقیاس ترتیبی فاقد معنای کمی و عددی هستند. علت این امر آن است که در مقیاس ترتیبی هیچ مبدای وجود ندارد تا ارزش ها نسبت به آن سنجیده شوند.
- برای این مقیاس میانه یا صدکها بهتر از میانگین است.

انواع مقیاس های اندازه گیری

All Types of measurement scales

➤ **مقیاس فاصله ای:** ویژگی عمده ی مقیاس فاصله ای آن است که مقدار ارزشها را با نمره های عددی می سنجد. این مقیاس علاوه بر آن که اطلاعات (خواص) مربوط به مقیاس های اسمی و ترتیبی را در بر دارد، اطلاعات بیشتری در مورد تفاوت ارزشها به دست می دهد.

➤ برای مثال 11 دو دمای 30 و 40 درجه در نظر بگیرید. این دو دما با یکدیگر تفاوت دارند، دمای 40 بیشتر از 30 است و همچنین دقیقا 10 درجه تفاوت وجود دارد. به این ترتیب در این مقیاس می توان روشهای آماری مختلف مانند محاسبه میانگین، میانه، مد، واریانس، همبستگی و... را انجام داد. در این مقیاس نسبت فاصله ها ثابت است.

$$F = \frac{9}{5}C + 32. \quad \rightarrow$$

5	10	15	سیلسی وس
41	50	59	فارنهایت

انواع مقیاس های اندازه گیری

All Types of measurement scales

- لازم به ذکر است که این مقیاس دارای نقطه ی صفر مطلق نیست ولی می توان برای آن صفر قراردادی برگزید. در مثال فوق دو تا صفر داریم 0 و 32. به طور مثال اگر سه نفر در یک آزمون روانشناسی نمرات 20، 30 و 15 گرفته باشند نمی توان گفت نمره 30 معرف دو برابر بودن دانش فرد نسبت به فرد با نمره 15 است. چون مبنا را هنوز نمی دانیم.
- مقیاس فاصله ای ، اندازه گیری با فواصل برابر و بدون صفر مطلق است.

انواع مقیاس های اندازه گیری All Types of measurement scales

➤ مقیاس نسبتي

- این مقیاس نه تنها دارای تمام ویژگی های مقیاس های قبلی است بلکه دارای مبدأ مطلق یا صفر حقیقی نیز می باشد و بنابراین امکان محاسبه نسبت های مختلف از لحاظ صفت مورد نظر در این مقیاس وجود دارد.
- متغیرهایی مانند سن، نرخ مرگ و میر، وزن، قد افراد، جمعیت و ... از این دست هستند. تمام اعمال آماری و ریاضی را می توان بر روی این مقیاس انجام داد.
- تذکر: اگر خواص مقیاسهای فوق را در نظر بگیریم، آنها را بترتیب زیر می توان مرتب کرد:
- مقیاس نسبتي < مقیاس فاصله ای < مقیاس ترتیبی < مقیاس اسمی

انواع مقیاس های اندازه گیری All Types of measurement scales

- مثال‌هایی 12 از مقیاس‌ها:
- اسمی : رنگ چشم شما چیست؟
- کدامیک از مجموعه های تلویزیونی را تماشا می کنید؟
- ترتیبی : مجموعه های تلویزیونی مورد توجه شما تا چه حد منعکس کننده واقعیت های زندگی است؟
- خیلی زیاد، زیاد، متوسط ، کم، بسیار کم
- فاصله ای : بهره هوشی افراد مجموعه تلویزیونی مورد علاقه خود را حدس بزنید؟
- نسبی : هفته ای چند ساعت مجموعه ی تلویزیونی تماشا می کنید؟
- هفته ای چند ساعت درس می خوانید؟

جدول آماری Statistical table: در جداول آماری علاوه بر داده ها، تعداد و درصد تکرار آنها نمایش داده می شود بطوریکه با یک نگاه به جدول می توان اطلاعات مفیدی در مورد پراکندگی و توزیع داده ها بدست آورد.

(1) **فراوانی Frequency:** با شمردن تعداد دفعات تکرار هر داده در میان تمامی داده ها فراوانی آن داده بدست می آید. معمولاً برای داده های کیفی یا گسسته یا طبقه بندی شده بکار می رود.

فراوانی طبقه ام را به f_i :

$$\sum_{i=1}^k f_i = N \text{ و } 1 \leq f_i \leq N \text{ و } 1 \leq i \leq k \text{ (تعداد کل داده ها)}$$

مثال 1: گروه خون

(2) **فراوانی نسبی Relative frequency:** حاصل تقسیم فراوانی هر طبقه بر تعداد کل داده هاست.

$$r_i = \frac{f_i}{N}, i = 1, \dots, k$$

(2) فراوانی تجمعی :Cumulative frequency

حاصل جمع فراوانی هر طبقه با طبقات قبل از آن را فراوانی تجمعی نامند.

$$g_i = f_1 + f_2 + \cdots + f_i = \sum_{j=1}^i f_j$$

(2) فراوانی نسبی تجمعی :Cumulative Relative frequency

حاصل جمع فراوانی نسبی هر طبقه با طبقات قبل از آن را فراوانی نسبی تجمعی نامند.

$$r_i = \frac{f_i}{N}, i = 1, \cdots, k$$

$$s_i = r_1 + r_2 + \cdots + r_i = \sum_{j=1}^i s_j = s_i = \frac{1}{N} g_i$$

جدول فراوانی برای داده های گسسته

مثال 2: داده های زیر تعداد فرزندان تحت پوشش بیمه در ۳۰ خانواده را نشان می دهد.

4 1 5 0 3 2 2 2 3 0 1 4 5 2 5 3 4 0 2 1 5 2 3 3 1 5 4 3 0

ابتدا داده ها را از کوچک به بزرگ برای تشکیل جداول فراوانی مرتب می کنیم:

0 0 0 0 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 5 5 5

جدول فراوانی برای داده های گسته

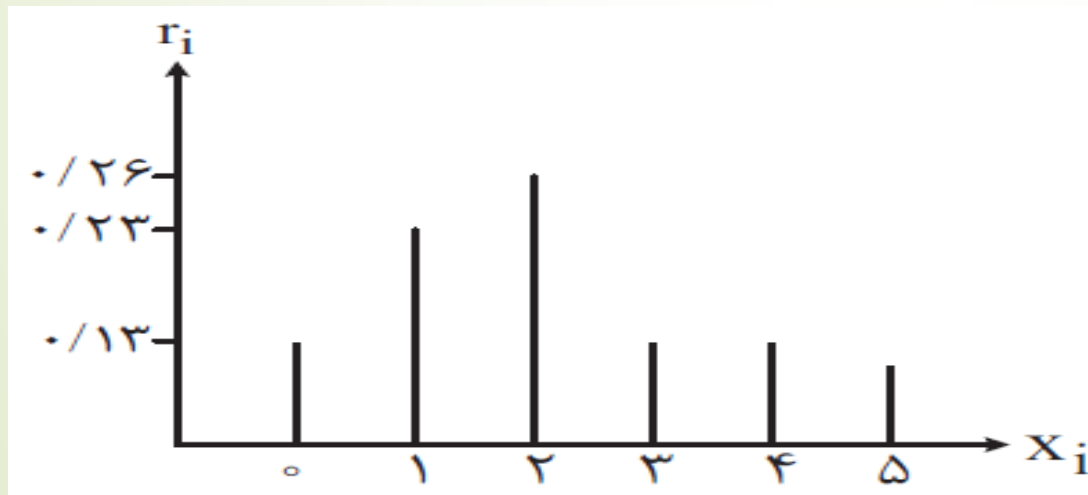
0 0 0 0 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 5 5 5 5 5

xi	fi	gi	ri	si
0	4	4	0.1333	0.1333
1	4	8	0.1333	0.2666
2	7	15	0.2333	0.5
3	6	21	0.2	0.7
4	4	25	0.1333	0.8333
5	5	30	0.1666	1

نمودارهای آماری

Statistical graphs

➤ نمودارهای میله‌ای Bar chart: در این نمودار محور X ها نمایش دهنده مقادیر داده‌ها و محور Y ها نمایش دهنده فراوانی یا فراوانی نسبی می‌باشد. مثال تعداد بیمه‌شده‌ها

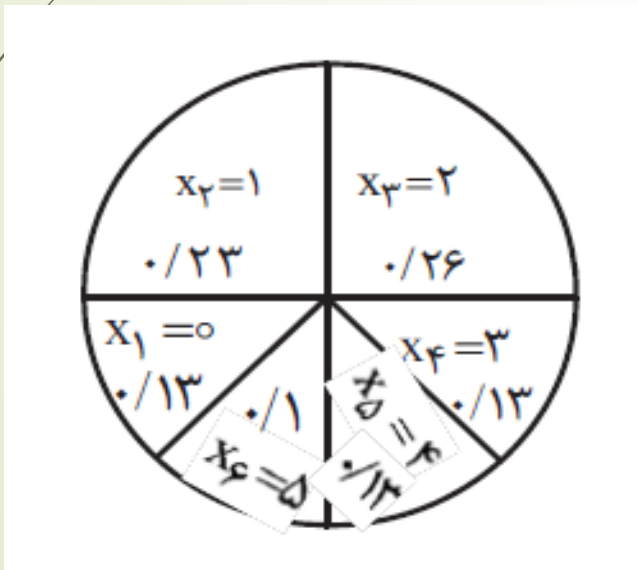


نمودارهای آماری

Statistical pie

نمودار دایره ای pie chart : این نمودار درون دایره ای رسم می شود که به قطاع هایی تقسیم شده است و هر قطاع متناسب با مساحت خود معرف یکی از فراوانی های نسبی در جدول فراوانی می باشد.

$$\theta_i = 0.23 * 360 = 82.8$$



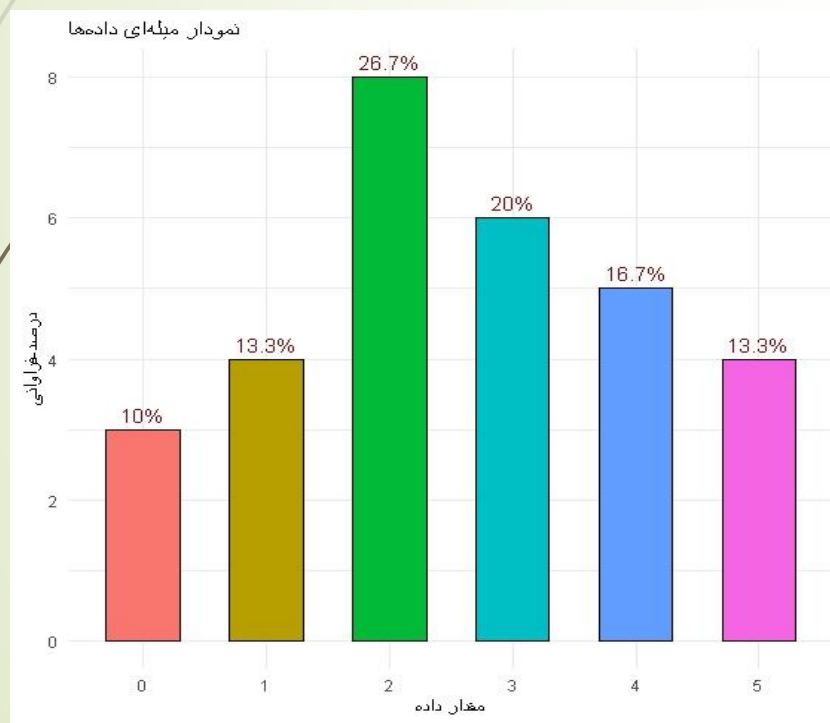
نمودارهای آماری

Statistical pie

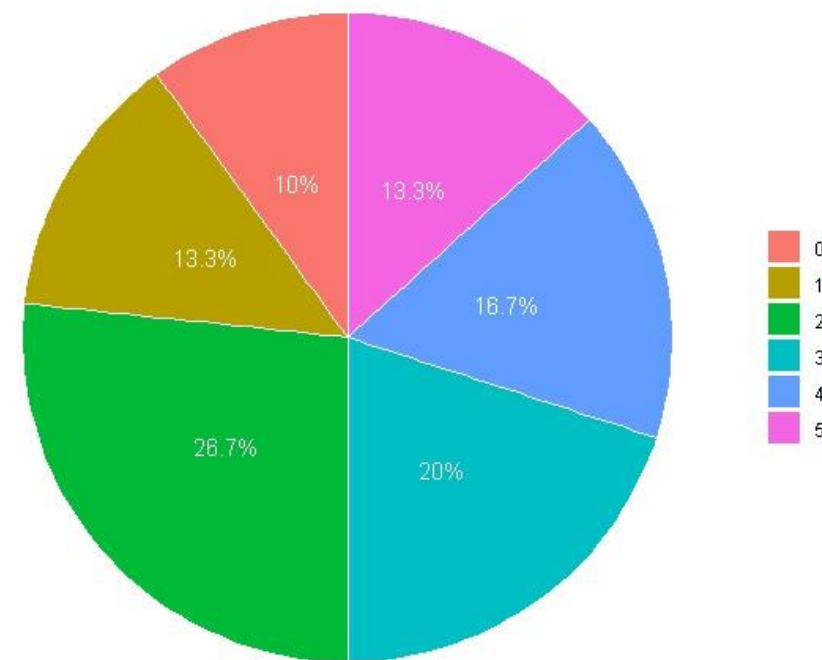
مثال:

5 4 3 2 1 0

4 5 6 8 4 3



نمودار دایره‌ای داده‌ها



نمودارهای آماری

Statistical pie

➤ **نمودار شاخه و برگ Stem and leaf:** در این نمودار عدد دهگان داده ها را به عنوان شاخه و رقم یگان آنها را برگ می نامند.

➤ **مثال:** این نمودار را برای داده های زیر رسم کنید.

62 75 31 44 56 54 30 81 88 6 35 11 91 63 32 42 65 69 39 6 36 76 75 27 20 42 56 75 39 40 62 47
50 68 50 53 61 59 29 65 29 27 83 56 52 51 64 35 33 52 62 45 5 75 22 28 36 42 72 36 37

```

0 | 66
1 | 1
2 | 07799
3 | 012355699
4 | 02247
5 | 00122346669
6 | 122345589
7 | 5556
8 | 138
9 | 1

```

جدول فراوانی برای داده های پیوسته

➤ مثال 4: داده های زیر طول عمر ۴۰ عدد لامپ را نشان می‌دهند که به نزدیکترین عدد صحیح گرد شده‌اند.

11	9	12	15	20	13	14	17	23	22
8	16	17	21	11	18	21	12	11	10
14	13	19	16	15	17	20	8	7	13
15	17	16	14	22	12	11	9	18	19

جدول فراوانی برای داده های پیوسته، محاسبه تعداد طبقات برای داده های خام

25

$$K = 1 + 3.222 \log_{10}^{40} = 6.322 \cong 7 \Rightarrow \omega = \frac{17}{7} = 2.4 \cong 3$$

$$\text{تعداد رده ها} = K \quad \text{طول رده} = \omega = \frac{R}{K}$$

$$S = \frac{\text{واحد گرد شده داده ها}}{2} = \frac{1}{2} = 0.5$$

برای داده های خام به صورت زیر عمل می کنیم:

کمترین داده = 7

بیشترین داده = 23

کران پایین طبقه (رده) اول $7 - 0.5 = 6.5$

کران بالای طبقه (رده) آخر $23 + 0.5 = 23.5$

$$\text{کوچکترین داده} = \min \quad -S = 7 - 0.5 = 6.5$$

$$\text{بزرگترین داده} = \max \quad +S = 23 + 0.5 = 23.5$$

$$\text{دامنه تغییرات داده ها} \quad R = \max - \min = 23.5 - 6.5 = 17$$

جدول فراوانی برای داده های پیوسته

26

ب عگرد شده اند. رابلالبی بیی برای

برای داده های خام به صورت زیر عمل می کنیم:

کمترین داده=7

بیشترین داده=23

کران پایین طبقه (رده) اول $7-0.5=6.6$

کران بالای طبقه (رده) آخر $23+0.5= 23.5$

دامنه تغییرات:

$$\text{Range} = \text{Max} - \text{Min} + 1$$

$$\text{Range} = C_{k+1} - C_1$$

نماینده طبقه:

$$X_i = (C_{i+1} + C_i) / 2$$

$$X_i = (A_i + B_i) / 2$$

جدول فراوانی برای داده های پیوسته

ردیف	حدود طبقات $A_i - B_i$	کران طبقات $C_i - C_{i+1}$	نماینده طبقه X_i	فراوانی مطلق f_i	فراوانی نسبی r_i	فراوانی تجمعی g_i	فراوانی نسبی تجمعی s_i	درصد فراوانی
1	7- 9	6.5-9.5	8	5	0.125	5	0.125	12.5
2	10-12	9.5-12.5	11	8	0.20	13	0.325	20
3	13-15	12.5-15.5	14	9	0.225	22	0.55	22.5
4	16-18	15.5-18.5	17	9	0.225	31	0.775	22.5
5	19-21	18.5-21.5	20	6	0.15	37	0.925	15
6	22-24	21.5-24.5	23	3	0.075	40	1	7.5

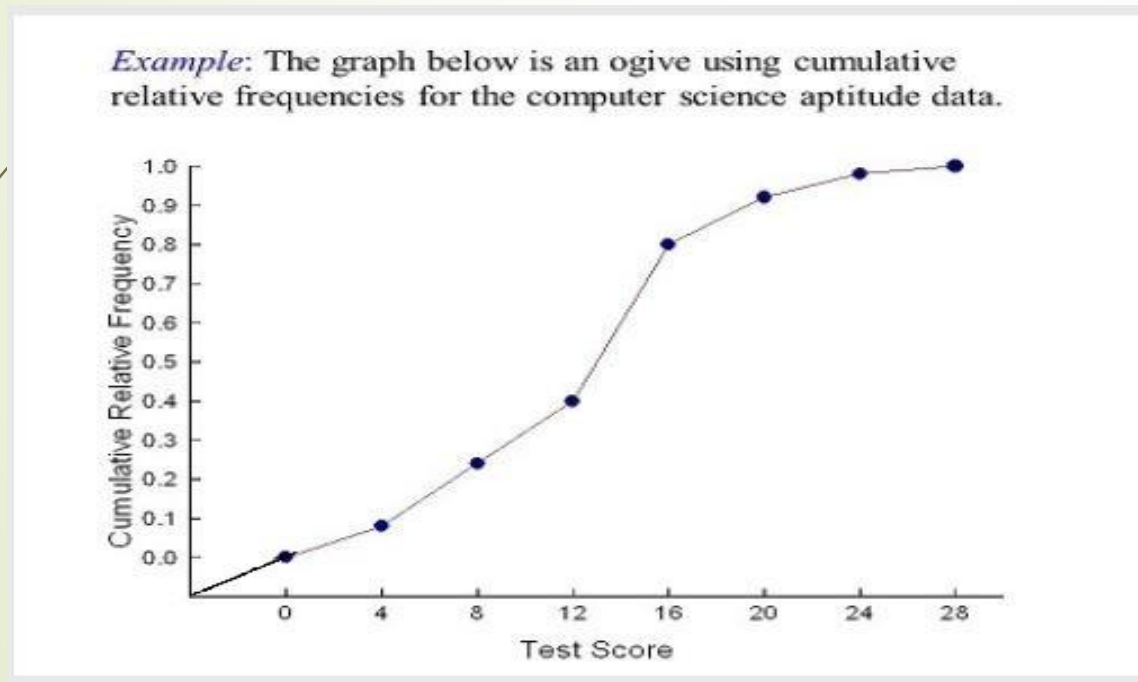
تذکر: X_i معرف نماینده هر رده است که عضو رده می باشد.

✓ چون آخرین رده شامل هیچ عضوی نداشت آنرا حذف می کنیم
 رده 27/5 - 24/5 شامل هیچ عضوی نیست بنابراین حذف
 می شود. نمودار دایره ای را رسم کنید.

رده‌ها	خط و نشان	X_i	f_i	r_i	g_i	s_i
6/5-9/5		۸	۵	0/125	۵	0/125
9/5-12/5		۱۱	۸	0/2	۱۳	0/325
12/5-15/5		۱۴	۹	0/225	۲۲	0/55
15/5-18/5		۱۷	۹	0/225	۳۱	0/775
18/5-21/5		۲۰	۶	0/15	۳۷	0/925
21/5-24/5		۲۳	۳	0/075	۴۰	1/0
جمع			۴۰			

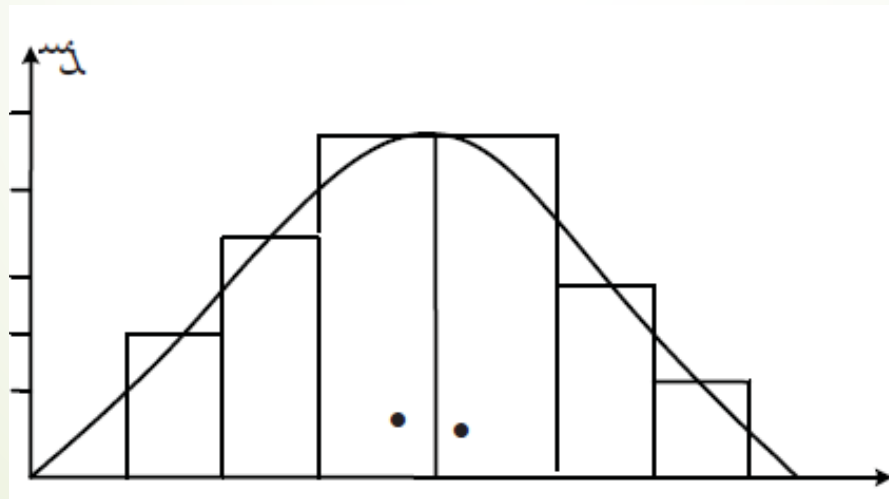
نمودار فراوانی تجمعی برای داده های پیوسته

- این نمودار روی کرانها رسم می شود.
- نمودار فراوانی تجمعی (نمودار توزیع فراوانی) مثال قبل را رسم کنید.



نمودار فراوانی/نسبی (چگالی) داده ها

هستوگرام (نمودار ستونی) و چندبر فراوانی: این نمودار کاملاً مشابه نمودار میله ای می باشد با این تفاوت که محور Xها نمایش دهنده رده های جدول فراوانی است و به ازای هر رده مستطیلی که عرض آن یک برابر طول طبقه و ارتفاع آن معادل فراوانی (نسبی) آن رده است رسم می شود.



مثال

- مثال 5: جدول فراوانی داده های زیر را کامل کنید، هیستوگرام، چندبر فراوانی، نمودار فراوانی نسبی تجمعی و نمودار دایره ای داده ها را رسم کنید.
- داده ها 12 11 10 9 8 7 6 4 3
- فراوانی 1 3 4 3 7 4 6 1 1
- تعداد طبقات را 5 در نظر بگیرید.
- تذکر: برای طبقه بندی داده های پیوسته، کران پایین طبقه اول را برابر کمترین مقدار داده ها یا عددی کوچکتر از آن در نظر می گیریم. مثلاً اگر $\min(X)=5.6587$ باشد کران پایین طبقه اول می تواند همین مقدار یا 5.65 باشد.
- واحد تقریب 1، 2.5 تا 12.5، $w=2$

پارامترهای مرکزی مرکزی Central tendency parameters

عموماً داده ها در جامعه آماری یک نوع تجمع و فشردگی حول یک مقدار خاص از صفت مورد مطالعه را بوجود می آورند که این مقدار خاص به عنوان یک پارامتر مرکزی معرفی می شود. مهمترین پارامترهای (شاخصهای) مرکزی عبارتند از:

میانگین Mean،

میان Median

مد (نما) Mode

میانگین Mean: میانگین بر چند نوع است که معروفترین آنها میانگین حسابی، هندسی، وزنی، همساز (هارمونیک) و میانگین درجه دوم می باشد.

پارامترهای مرکزی مرکزی Central tendency parameters

- میانگین Mean: میانگین بر چند نوع است که معروفترین آنها:
- میانگین حسابی Arithmetic mean،
- میانگین هندسی Geometric Mean،
- میانگین Weighted mean،
- میانگین همساز (هارمونیک) Harmonic mean،
- میانگین درجه دوم Quadratic mean می باشد.
- (1) میانگین حسابی :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^K f_i x_i$$

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N}$$

که در آن k تعداد طبقات است.

پارامترهای گرایش مرکزی Central tendency parameters

(1) میانگین وزنی: مثال 2

$$\bar{X} = \frac{\sum_{i=1}^k w_i X_i}{w} ; \quad w = \sum_{i=1}^k w_i \quad (2)$$

(3) معدل درسی شما در ترم جاری.

(4) 15, 16, 18, 14 داده‌ها

(5) 3, 3, 2, 4 فراوانی

(6) میانگین هندسی: برای داده‌های (مثبت) درصدی، نرخ تورم،
نرخ سود و نرخ زیان و نرخ بازدهی سبد سرمایه گذاری ...

(7) مثال 1

$$\bar{X}_G = (\prod_{i=1}^k X_i^{f_i})^{1/n} \quad \text{یا} \quad \bar{X}_G = (\prod_{i=1}^n X_i)^{1/n} \quad 2, 2, 2, 4, 4, 5, 5 \quad (8)$$

که در آن $\sum f_i = n$

پارامترهای مرکزی Central parameters

میانگین همساز: برای داده هایی مثل سرعت (کیلومتر بر ساعت)، نسبت ها و نرخ ها مورد استفاده قرار می گیرد.

$$\bar{X}_H = \frac{n}{\sum_{i=1}^k \frac{1}{X_i f_i}} \quad \text{یا} \quad \bar{X}_H = \frac{n}{\sum_{i=1}^n \frac{1}{X_i}}$$

میانگین ریشه ای رتبه دو:

وقتی علامت داده ها مهم نیست و اندازه کلی آن مورد توجه است از این میانگین استفاده می شود. جریان متناوب، محاسبه انحراف معیار، ارزیابی خطاها در مدل سازی، محاسبه انرژی سیگنال یا شدت نوسان

$$\bar{X}_{R2} = \left(\frac{\sum_{i=1}^n X_i^2}{n} \right)^{1/2} \quad \text{یا} \quad \bar{X}_{R2} = \left(\frac{\sum_{i=1}^k f_i X_i^2}{n} \right)^{1/2}$$

$$\text{Harmonic} \leq \text{Geometric} \leq \text{Arithmetic} \leq \text{Quadratic}$$

ارایه جدول مقایسه میانگینها.

تمرینات تحویلی رابطه میانگی های فوق.

میانه (برای داده های گسسته):

میانه: با مرتب کردن داده ها به صورت غیر نزولی، عدد m را بعنوان میانه آنها در نظر می گیریم اگر تقریباً نیمی از داده ها سمت چپ آن و نیمی دیگر در سمت راست آن قرار داشته باشند. عددی است که در وسط داده ها قرار دارد.

1) برای داده های گسسته با ترتیب نزولی:

(اگر n زوج باشد)

$$m = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

مثال 6: 10, 2, 5, 7, 8, 9, 4, 3

اگر n فرد باشد عددی را که در وسط داده ها قرار گیرد میانه است.

مثال: 7, 2, 3, 8, 9, 1, 6

تذکر: میانه در برابر داده های پرت (دور افتاده) مقاوم است. مثال:

2) میانه برای داده های پیوسته و طبقه ای:

$$m = L_m + \frac{\left(\frac{N}{2} - g_{m-1}\right) W}{f_m}$$

L_m : حد پایین رده میانه دار

N : تعداد داده ها

g_{m-1} : فراوانی تجمعی رده قبل از رده میانه دار

f_m : فراوانی رده میانه دار

W : طول رده

مثال 7: میانگین، میانه داده های صفحه 27 را محاسبه کنید.

(3) چندکها برای داده های گسسته :

فرض کنید n داده داریم . $x_1, \dots, x_n$

اول آنها را مرتب می کنیم .

$$y_1 \leq \dots \leq y_n$$

حال اگر Q_p چندک مرتبه p باشد و $P(n+1)$ برابر عدد صحیح مانند r باشد آنگاه y_r را برای چندک مرتبه p می گیریم یعنی $Q_p = y_r$. در غیر اینصورت قسمت صحیح آنرا r و قسمت اعشاری آنرا w فرض می کنیم و

$$Q_p = (1 - w)y_r + wy_{r+1}$$

مثال : $2, 6, 4, 6, 8, 9, 5 \rightarrow 2, 4, 5, 6, 6, 8, 9$

$$P=0.25, \quad p(n+1)=0.25*8=2 \quad Q_p = 4$$

اگر به داده ها عدد 10 را اضافه کنیم : $2, 4, 5, 6, 6, 8, 9, 10$

$$P=0.25, \quad p(n+1)=0.25*9=2.25$$

$$Q_p = (1 - 0.25)4 + 0.25 * 5 = 4.25$$

تذکر : اگر $p = 0.25, 0.5, 0.75$ یا $p = 0.1, 0.2, \dots, 0.9$ یا $p = 0.01, 0.02, \dots, 0.99$

Q_p را بترتیب چارک، دهک، صدک مرتبه p نامند .

➡ **مد یا نما (Mode):** داده‌ای که فراوانی آن از فراوانی سایر داده‌ها بیشتر باشد مد یا نما گفته می‌شود و با M نمایش داده می‌شود.

(1) روش محاسبه مد برای داده‌های گسسته:

برای داده‌های ۵ و ۴ و ۴ و ۲ و ۲ و ۲ و ۱ و ۱ چون $M=2$

اما برای داده‌های ۴ و ۴ و ۴ و ۲ و ۲ و ۱ و ۱ و ۱ فراوانی ۱ و ۴ مساوی، از همه بیشتر می‌باشد ولی مجاور یکدیگر نیستند پس دو مقدار برای مد خواهیم داشت که عبارتند

$$M_1 = 1 \quad M_2 = 4$$

➡ **مد یا نما:** اگر دو داده کنار هم با بیشترین فراوانی باشد، میانگین حسابی دو داده کنار هم را میانه در نظر می‌گیریم. چون برای داده های ۴ و ۴ و ۲ و ۲ و ۲ و ۱ و ۱ و ۱، فراوانی ۱ و ۲ مساوی می‌باشد و چون مجاور یکدیگر هستند مقدار مد برابر است با

$$M = \frac{1+2}{2} = \frac{3}{2}$$

(2) روش محاسبه مد برای داده های پیوسته:

- L_M : کران پایین رده نمایی دار
- D_1 : اختلاف فراوانی های نسبی رده نمایی و رده قبل از آن
- D_2 : اختلاف فراوانی های نسبی رده نمایی و رده بعد از آن
- ω : طول رده
- مثال 8: صفحه 12

$$M = L_M + \left(\frac{D_1}{D_1 + D_2} \right) \omega$$

(2) روش محاسبه مد برای داده های پیوسته (طبقه بندی شده) :

- L_M : کران پایین رده نمایی دار
- D_1 : اختلاف فراوانی های نسبی رده نمایی و رده قبل از آن
- D_2 : اختلاف فراوانی های نسبی رده نمایی و رده بعد از آن
- ω : طول رده
- مثال 9 : مد = 19.556

	FX	Xi	Freq	CFreq	Relative	Crelative
•	1 (5.4, 10.3]	7.85	4	4	0.08	0.08
•	2 (10.3,15.2]	12.75	7	11	0.14	0.22
•	3 (15.2,20.1]	17.65	15	26	0.30	0.52
•	4 (20.1,25.0]	22.55	14	40	0.28	0.80
•	5 (25.0,29.9]	27.45	9	49	0.18	0.98
•	6 (29.9,34.8]	32.35	1	50	0.02	1.00

$$M = L_M + \left(\frac{D_1}{D_1 + D_2} \right) \omega$$

- تمرین 1: میانگین داده های فوق و بیمه را محاسبه کنید.

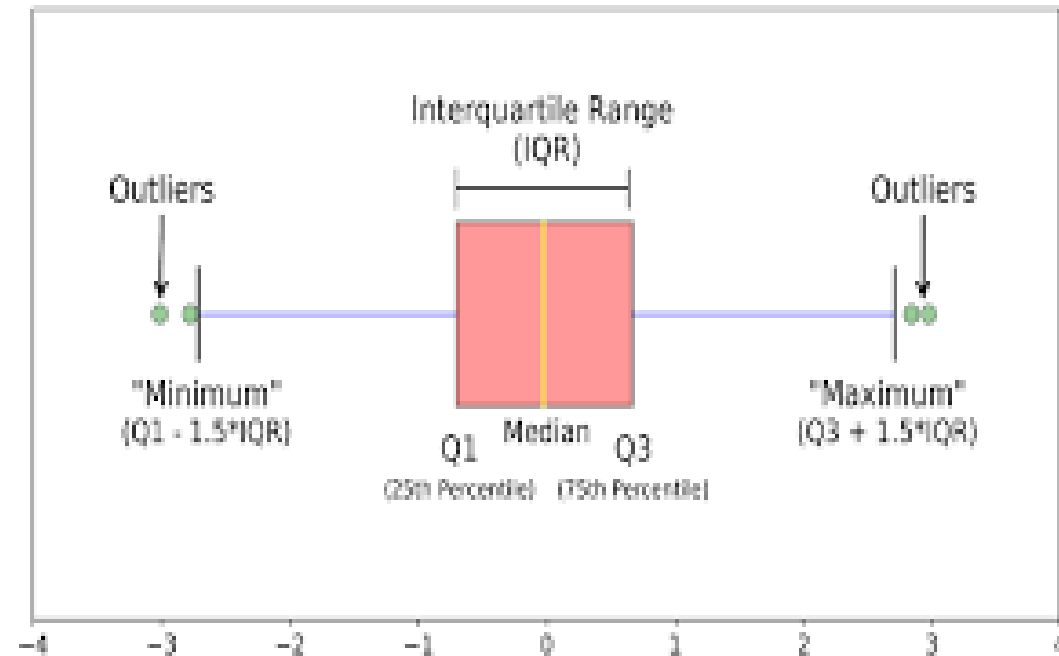
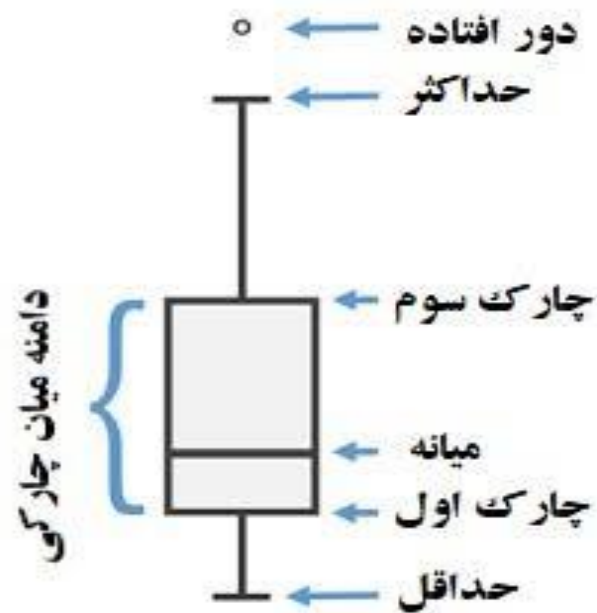
(2) $IQR = Q3 - Q1$ دامنه میان چارکی:

(3) دارای خواص زیر است:

(4) برای تشخیص داده های پرت استفاده می شود.

(5) مقاوم در برابر داده های پرت (Outliers) است.

(6) برای محاسبه باکس پلات (Box Plot) استفاده می شود.



پارامترهای (شاخص های) پراکندگی Dispersion parameters (Measures)

➤ **دامنه (تغیرات) داده ها:** حاصل تفاضل بزرگترین داده از کوچکترین داده را R یا دامنه داده ها می گویند.

$$R = X_{(N)} - X_{(1)}$$

➤ که در آن داده ها به فرم $X_{(1)} < X_{(2)} < \dots < X_{(N)}$ با فرض اینکه مرتب شده اند.

➤ **میانگین انحرافات:** فاصله داده x_i از میانگین داده ها $\bar{x}$ که به صورت $|x_i - \bar{x}|$ محاسبه می شود را انحراف از میانگین می نامند. اگر مجموع این مقادیر را بر تعداد داده ها تقسیم کنیم، میانگین انحرافات از میانگین بدست خواهد آمد.

$$D = \frac{\sum_{i=1}^k f_i |X_i - \bar{X}|}{N}; \quad D = \frac{\sum_{i=1}^N |X_i - \bar{X}|}{N} \quad \rightarrow$$

➤ **واریانس و انحراف استاندارد:** میانگین مجذور انحرافات را واریانس می نامیم و با نماد s^2 یا s_x^2 نمایش می دهیم. جذر واریانس را انحراف معیار

$$s_b^2 = \frac{1}{N} \sum_{i=1}^K f_i (x_i - \bar{x})^2$$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^K f_i (x_i - \bar{x})^2 \quad \text{نامند.}$$

➤ واریانس برای داده های خام:

44

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

➤ ثابت کنید تساوی فوق برقرار است.

➤ توجه داشته باشید که:

$$\Rightarrow H = \frac{\sum_{i=1}^N (X_i - \bar{X})}{N} = 0 = \sum_{i=1}^N (X_i - \bar{X}) = \sum_{i=1}^N X_i - n\bar{X}$$

➤ مثال:

➤ 1,2,3,3,3,4,5: mean= M=m=3 توزیع فراوانی متقارن

➤ 1,1,1,2,2,3,10 : mean(X)=2.86>Median=2> Mode=1 توزیع فراوانی کشیده

➤ بر است

➤ 0,7,8,8,9,9,9: mean=7.14<median=8<Mode=9 توزیع فراوانی کشیده به چپ

➤ واریانس و انحرافات و قدر مطلق آن را برای داده های فوق حساب کنید.

ضرب تغییرات:

45

در محاسبه میزان پراکندگی داده‌ها همواره با داده‌هایی سروکار داریم که با مقیاس‌های مختلفی اندازه‌گیری شده‌اند.

بنابراین برای مقایسه میزان پراکندگی داده‌های بدست‌آمده از دو جامعه آماری که با مقیاس‌های مختلفی اندازه‌گیری شده‌اند استفاده از واریانس مناسب نیست. در نتیجه در این موارد از ضرب تغییرات استفاده می‌شود که به واحد اندازه‌گیری بستگی ندارد و به صورت زیر تعریف می‌شود:

$$CV = \frac{S}{\bar{X}}$$

برای سه گروه داده فوق ضرب تغییرات را محاسبه کنید.

- $Cv=0.40, sd=1.19, var=\frac{10}{7}=1.43$
- $Cv=1.049, sd=2.97, var=8.84$
- $Cv=0.42, sd=2.998, var=8.98$

تغییر مقیاس و مبدا داده ها:

(1) تغییرات مقیاس:

$$x_1, x_2, \dots, x_n \rightarrow ax_1, ax_2, \dots, ax_n$$

$$\Rightarrow \bar{X} = \frac{1}{N} \sum_{i=1}^n x_i \rightarrow \bar{X}_a = \frac{1}{N} \sum_{i=1}^n ax_i = a \left(\frac{1}{N} \sum_{i=1}^n x_i \right) = a \bar{X} \Rightarrow \bar{X}_a = a \bar{X}$$

(1) اثبات شود

(2) تغییر مبدا و مقیاس داده ها: در صورتی که به تمام داده ها مقدار b را اضافه یا کم کنیم، و همه داده ها را در عدد a ضرب کنیم می توان نشان داد که مقادیر $\bar{y}$ و s_y^2 جدید از روابط زیر محاسبه می شوند. اثبات شود

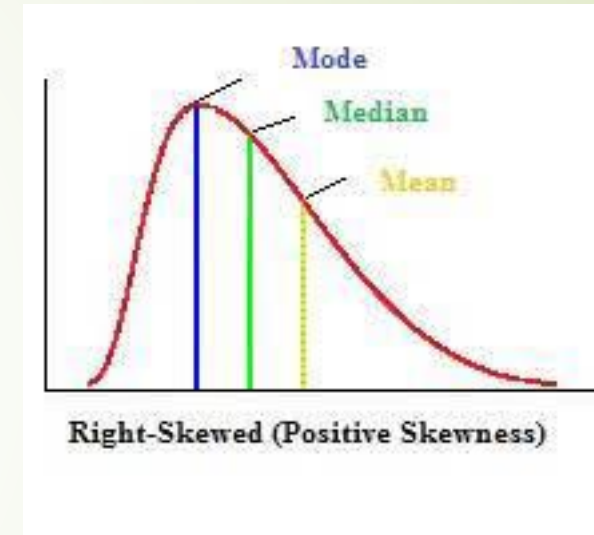
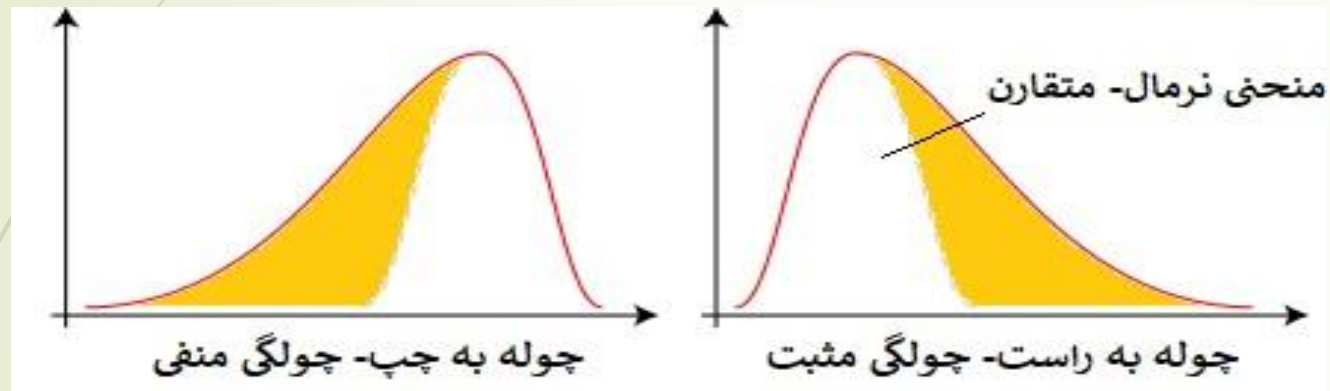
$$x_1, x_2, \dots, x_n; \bar{x}, s_x^2 \rightarrow y_1, y_2, \dots, y_n; y_i = ax_i + b$$

$$\bar{y} = a\bar{x} + b$$

$$s_y^2 = a^2 s_x^2$$

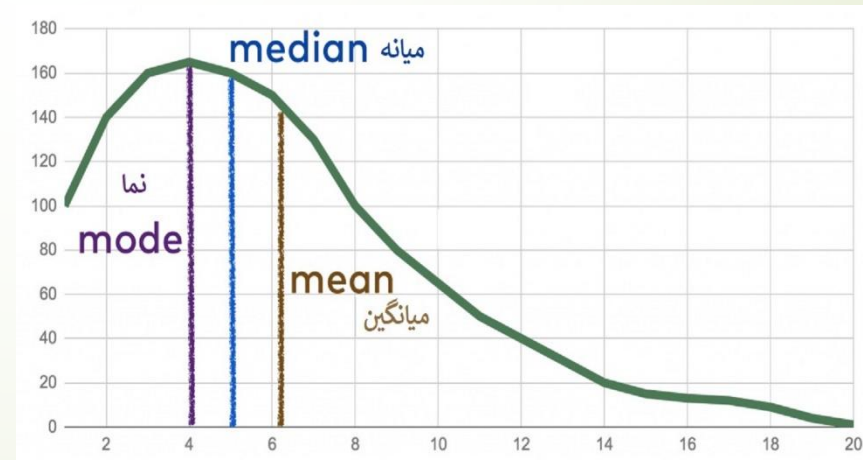
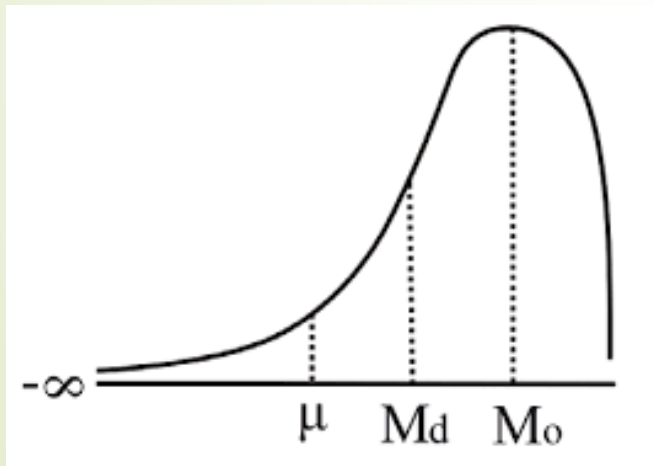
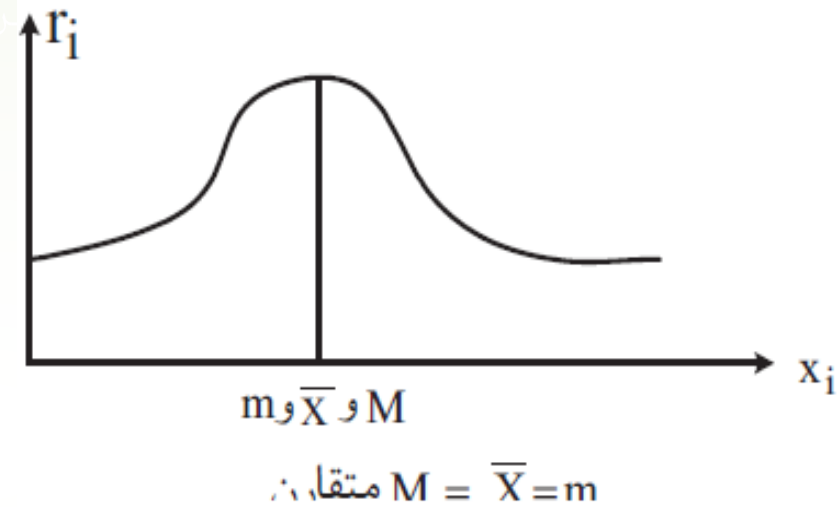
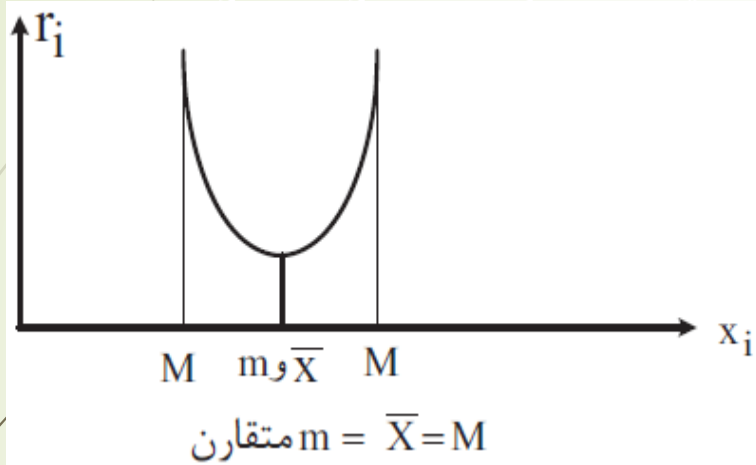
عدم تقارن یا چولگی (کشیدگی چپ یا راست) توزیع فراوانی داده‌ها

چگونگی قرار گرفتن میانگین و میانه و مد را در اشکال زیر مشاهده می‌شود.



چولگی (کشیدگی چپ یا راست) توزیع فراوانی داده‌ها Left/Right Skewness

نمودار و میانه و مد را در اشکال



محاسبه میزان عدم تقارن منحنی (ضریب چولگی پیرسون):

$$\bar{X} - M \cong 3(\bar{X} - m)$$

اگر این رابطه برقرار باشد نشان می دهد که منحنی تقریباً مقارن است.

➤ $sk_p^1 = \frac{\bar{X} - M}{s}$

➤ $sk_p^2 = 3 \frac{\bar{X} - m}{s}$

➤ مثبت بودن هردو ضریب فوق نشان می دهد که منحنی به راست کشیده شده و منفی بودن ضرایب نشان می دهد که منحنی به چپ کشیده شده است.

➤ بدست آوردن ضریب چاولگی از مفهوم گشتاور مرتبه K ام حول میانگین (گشتاور مرکزی و غیر مرکزی مرتبه K ام):

گشتاور مرکزی برای داده های خام: $\mu_k = \frac{\sum_{i=1}^N (X_i - \bar{X})^k}{N}$

گشتاور مرکزی برای داده های طبقه بندی: $\mu_k = \frac{\sum_{i=1}^p f_i (X_i - \bar{X})^k}{N}$

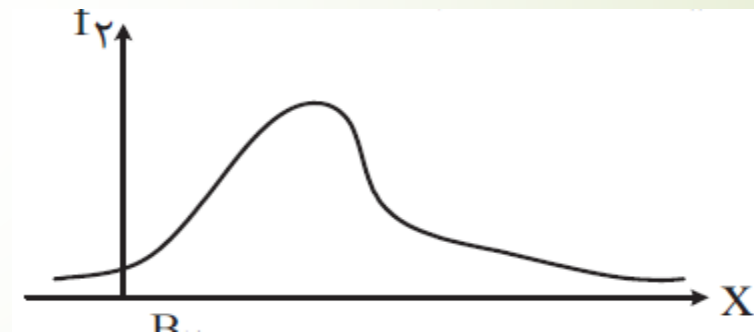
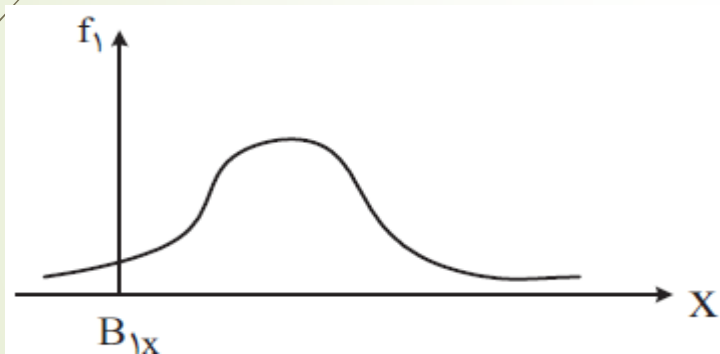
گشتاور غیرمرکزی برای داده های خام: $\mu_k' = \frac{\sum_{i=1}^p f_i X_i^k}{N}$

گشتاور غیرمرکزی برای داده های طبقه بندی: $\mu_k' = \frac{\sum_{i=1}^N X_i^k}{N}$

➡ ضریب چاولگی که با B_1 :

هرچه B_1 مقدار بزرگتری (کوچکتری) داشته باشد نشان دهنده افزایش عدم تقارن می باشد .

$$B_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\delta^3}$$



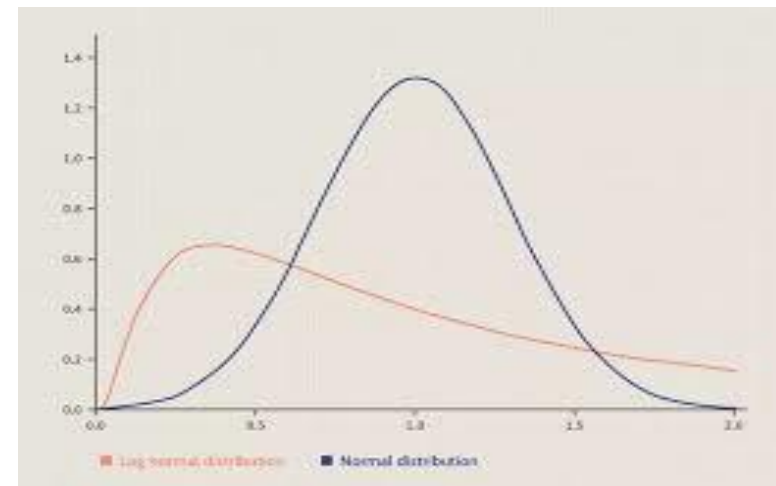
$$B_{1x} < B_{1y}$$

➤ ضریب کشیدگی (ارتفاع) منحنی فراوانی: نشان دهنده میزان کشیدگی منحنی می باشد و به صورت زیر تعریف می شود.

$$B_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

➤ کشیدگی توزیع نرمال برابر 3 است.

دو منحنی زیر مقارن اما ضریب کشیدگی در آنها متفاوت است.



نرمال (استاندارد) سازی Standardize

از یک جامعه آماری یک نمونه n تایی مانند $x_1, x_1, \dots, x_n$ به صورت تصادفی انتخاب می‌کنیم بطوریکه میانگین واریانس نمونه‌ها به ترتیب $\bar{x}$ و s_x^2 می‌باشد با توجه به تغییر مبدا و تغییر مقیاس مقدار هر نمونه از میانگین داده‌ها کم می‌کنیم و حاصل را بر s_x تقسیم می‌کنیم بنابراین داده‌های

$$y_1 = \frac{x_1 - \bar{x}}{s_x}, y_2 = \frac{x_2 - \bar{x}}{s_x}, \dots, y_n = \frac{x_n - \bar{x}}{s_x}$$

را خواهیم داشت. مثال عددی داده شود.

خواص داده های نرمال استاندارد:

1- مستقل از واحد است.

$$\bar{y} = 0, \text{var}(y) = 1$$

3- برای مقایسه دو جامعه بکار می‌رود.

4- هرچه بزرگتر باشد بهتر است.

مثالهای عددی از داده های تبدیل یافته

➤ مثال 13:

X: 8 7 4 4 3 7 3 9 5 7 10 2 7 7 10 12 9 7 3 10 ➤

7 8 8 2 9 7 10 5 5 8 ➤

Mean(x)=6.776667 sd(x)= 2.64814 ➤

Var(x)=7.012644 ➤

Y=2+3x ➤

Y: 26 23 14 14 11 23 11 29 17 23 32 8 23 23 32 38 29 23 11 32 23 ➤

26 26 8 29 23 32 17 17 26

Mean(y)=2+3mean(x)=22.3 , sd(y)=3*sd(x)=7.944419 ➤

Var(y)=9var(x)=63.11379 ➤

مثالهای عددی از داده های استاندارد سازی

مثال 14: ➡

X: 8 7 4 4 3 7 3 9 5 7 10 2 7 7 10 12 9 7 3 10 ➡
7 8 8 2 9 7 10 5 5 8

Mean(x)=6.776667 sd(x)= 2.64814, Var(x)=7.012644

$Y = (x - \text{mean}(x)) / \text{sd}(x)$

Y: -0.69 -0.83 -1.25 -1.25 -1.38 -0.83 -1.38 -0.55 -1.11 -0.83 -0.41 -
1.52

-0.83 -0.83 -0.41 -0.14 -0.55 -0.83 -1.38 -0.41 -0.83 -0.69 -0.69 -1.52

-0.55 -0.83 -0.41 -1.11 -1.11 -0.69

sd(y)= 0.9996253, var(y)=0.9992507

Mean(y)=0.001

مفاهیم داده و منابع آن

- در دنیای امروز، داده‌ها نقش اساسی در تصمیم‌گیری‌های علمی، اقتصادی و اجتماعی دارند .
- داده Data چیست؟ اطلاعات یا مشاهدات ...
- شناخت منابع داده Source of data ،
- داده‌های ثبتی و انواع آن register data ،
- داده‌های میدانی Filed Data
- مه داده‌ها Metadata
- کلان‌داده‌ها Big Data
- برای پژوهشگران تحلیل‌گران و دانشجویان ضروری است .
- در این گزارش، چهار مفهوم کلیدی مرتبط با داده‌ها معرفی و در صورت امکان با مثال تشریح می‌شوند.

مفاهیم داده و منابع آن

- در دنیای امروز، داده‌ها (اطلاعات) نقش اساسی در تصمیم‌گیری‌های علمی، اقتصادی و اجتماعی دارند .
- شناخت منابع داده،
- انواع داده‌های ثبتی،
- فراداده (متاداده)،
- مه داده‌ها
- کلان‌داده‌ها
- برای پژوهشگران، تحلیل‌گران و دانشجویان ضروری است .
- در این زمینه، چهار مفهوم کلیدی مرتبط با داده‌ها معرفی و در صورت امکان با مثال تشریح می‌شوند.

منابع داده‌ها (Data Sources)

- منابع داده‌ها محل یا منبعی هستند که داده‌ها از آن جمع‌آوری می‌شوند. این منابع می‌توانند اولیه یا ثانویه باشند..
- • منابع اولیه: داده‌هایی که پژوهشگر مستقیماً جمع‌آوری می‌کند مثلاً پرسشنامه، مشاهده، آزمایش.
- • منابع ثانویه: داده‌هایی که پیش‌تر (در جای دیگری) گردآوری شده‌اند مانند داده‌های سرشماری یا بانک‌های اطلاعاتی.
- مثال: مصاحبه با معلمان منبع اولیه و داده‌های رسمی وزارت آموزش و پرورش منبع ثانویه.

داده‌های ثبتی (Administrative or Register Data)

➤ داده‌های ثبتی در جریان فعالیت‌های اداری و سازمانی ایجاد می‌شوند و هدف اولیه از جمع‌آوری آن‌ها انجام کارهای اجرایی است، اما می‌توانند برای پژوهش و تحلیل نیز استفاده شوند.

➤ ویژگی‌ها: پیوسته، گسترده و دقیق.

➤ مثال: اطلاعات ثبت احوال درباره تولد و مرگ، داده‌های مالیاتی، سوابق تحصیلی دانش‌آموزان، دانشجویان، تراکنش بانک‌ها.

مه داده (فراداده یا متاداده) (Metadata)

داده‌هایی (اطلاعات) درباره‌ی داده است که معنا، ساختار، منشأ، زمان/مکان گردآوری، روش تولید، مسئول/مالک، نسخه، مجوز استفاده، قالب فایل، گُدهای مقادیر گمشده و... را مشخص می‌کنند.

فراداده می‌تواند هنگام گردآوری، پردازش یا بارگذاری به‌صورت خودکار یا دستی تولید شود. فراداده برای شفافیت، کشف‌پذیری، یکپارچه‌سازی و استفاده‌ی مجدد از داده‌ها در هر مقیاسی (از خردداده تا کلان‌داده) ضروری است.

مثال: در فایل اکسل نمرات، فراداده شامل نام و نوع ستون‌ها (عددی/متنی)، واحد اندازه‌گیری، تاریخ و شیوه‌ی گردآوری، تعریف هر متغیر، دامنه‌ی معتبر، گُذگذاری مقادیر گمشده، نسخه‌ی فایل و مجوز استفاده است.

مثال: داده‌های شامل قد، وزن و جنسیت خود داده‌ها باشند، توضیح اینکه قد برحسب سانتیمتر و یا از پرونده‌های پزشکی سال 1400 استخراج شده‌اند فراداده هستند.

کلان داده‌ها (Big Data Sources)

- ▶ کلان داده‌ها مجموعه‌هایی بسیار بزرگ، متنوع و پویا از داده‌ها هستند که با سرعت بالا تولید می‌شوند و نیازمند روش‌های تحلیل پیشرفته اند.
- ▶ ویژگی‌ها: حجم زیاد (Volume)، سرعت بالا (Velocity)، تنوع زیاد (Variety)، و عدم قطعیت (Veracity).
- ▶ مثال: داده‌های شبکه‌های اجتماعی، اطلاعات موقعیت مکانی تلفن همراه، و داده‌های حسگرهای اینترنت اشیا (IoT).
- ▶ به لحاظ حجم: کلان داده > مه داده > فراداده
- ▶ متاداده می تواند هم برای شفافیت داده های کوچک و بزرگ استفاده شود.

جدول مقایسه‌ای مفاهیم

مثال	ویژگی‌ها	تعریف	مفهوم
پرسشنامه یا داده‌های سرشماری	اولیه یا ثانویه بودن	محل گردآوری داده‌ها، اولیه یا ثانویه	منبع داده‌ها
داده‌های ثبت احوال یا مالیاتی	پیوسته و دقیق	داده‌های تولیدشده در فرآیندهای سازمانی	داده‌های ثبتی
نام ستون، تاریخ، واحد اندازه‌گیری	توصیف‌کننده‌ی معنا و ساختار داده	داده (اطلاعات) درباره داده‌ها	مه داده‌ها
داده‌های شبکه‌های اجتماعی	حجم زیاد، سرعت بالا، تنوع زیاد	مجموعه داده‌های عظیم و متنوع دیجیتال	منابع کلان‌داده‌ها