به نام خداوند بخشنده مهربان

روش‌های آماری و اقتصادسنجی

در تحلیل و مدل‌سازی داده‌های حمل و نقلی

محمد مهدی بشارتی

besharati@iut.ac.ir

فصل۲. آمار توصیفی

CHAPTER 2: DESCRIPTIVE STATISTICS

- ✓ **شاخص‌های عددی برای توصیف داده‌ها**

مدرس: محمدمهدی بشارتی

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

شاخص‌های عددی،

- ✓ نسبت به روش‌های ترسیمی، امکان بررسی نسبتا دقیق‌تر وضعیت داده‌ها را فراهم می‌آورند.
- ✓ برای انجام استنباط‌هایی درمورد پارامترهای جامعه ضروری هستند.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

شاخص‌های عددی

- شاخص‌های وضعیت نسبی داده‌ها
- شاخص‌های تمرکز داده‌ها
- شاخص‌های پراکندگی داده‌ها
- شاخص‌های میزان عدم‌تقارن داده‌ها
- شاخص‌های همبستگی

شاخص‌های عددی برای توصیف داده‌ها

❖ شاخص‌های وضعیت نسبی داده‌ها

اگر داده‌های بدست آمده از یک نمونه‌گیری را به ترتیب مقدار، از کوچک به بزرگ مرتب کنیم:

چندک:
Quantile

داده‌ای که بخش معینی(مثلا یک‌پنجم) از داده‌های مشاهده‌شده در نمونه، مقداری کمتر از آن دارند.
مانند: ***صدک‌ها***، ***دهک‌ها*** و ***چارک‌ها***

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های وضعیت نسبی داده‌ها

اگر داده‌های بدست آمده از یک نمونه‌گیری را به ترتیب مقدار، از کوچک به بزرگ مرتب کنیم:

صدک:
p^{th} *percentile*

داده‌ای که p درصد از داده‌های مشاهده‌شده در نمونه، مقداری کمتر از آن دارند.

سوال: ***یک مثال از استفاده از شاخص صدک در مباحث ترافیکی؟***

سرعت ۸۵٪

برای مثال، اگر صدک۸۵٪ سرعت خودروها در یک بزرگراه برابر با ۹۰کیلومتربرساعت بدست آید؛ بدین‌معناست که ۸۵٪ از سرعت‌های مشاهده شده، برابر یا کمتر از ۹۰کیلومتربرساعت بوده‌است.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های وضعیت نسبی داده‌ها

اگر داده‌های بدست آمده از یک نمونه‌گیری را به ترتیب مقدار، از کوچک به بزرگ مرتب کنیم:

چارک:
Quartile

صدک‌هایی هستند که داده‌های نمونه براساس مقادیر آنها به چهار گروه با تعداد داده مساوی تقسیم می‌شوند.

چارک اول (یا صدک ۲۵٪ یا چارک پایین) داده‌ای که یک‌چهارم داده‌ها مقداری کمتر از آن دارند.

چارک دوم (یا صدک ۵۰٪) که همان میانه داده‌هاست.

چارک سوم (یا صدک ۷۵٪ یا چارک بالا) داده‌ای که ۷۵٪ از داده‌ها، مقداری کمتر از آن دارند

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های تمرکز داده‌ها

اگر داده‌های بدست آمده از یک نمونه‌گیری را به ترتیب مقدار، از کوچک به بزرگ مرتب کنیم:

میانه:
median

داده‌ایست که در مرکز داده‌ها قرار می‌گیرد($\tilde{x} = x_{(n+1/2)}$)

درحالتی که تعداد داده‌ها زوج باشد، میانگین دو عدد میانی را بعنوان میانه درنظر می‌گیرند($\tilde{x}$ $= \frac{x_{(n/2)}+x_{(n/2+1)}}{2}$).

برای داده‌هایی که بصورت طبقه‌ای دسته‌بندی شده‌باشند، روش تعیین میانه کمی متفاوت است.

کتاب‌های آمار مقدماتی

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های تمرکز داده‌ها

میانگین:
mean

مقدار متوسط داده‌ها

$$E[x] = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

درصورتیکه کل جامعه را درنظر بگیریم، آنگاه میانگین نمونه($\bar{x}$) با میانگین جامعه(μ) جایگزین می‌شود

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

سوال: **آیا میانگین نمونه، یک متغیر تصادفی است؟ میانگین جامعه چطور؟**

✓ میانگین نمونه یک **متغیر تصادفی** است؛ اما میانگین جامعه، یک مقدار **ثابت** می‌باشد.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های تمرکز داده‌ها

مد(یا نما):
mode

برای مجموعه‌ای از داده‌ها، مد، مقداری‌است که بیشترین فراوانی را داشته‌باشد.

- ✓ برخلاف میانگین و میانه، در یک مجموعه از داده‌ها، ممکن است بیش از یک عدد بعنوان مد وجود داشته باشد.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ مقایسه ویژگی‌های سه شاخص میانگین، میانه، نما

- ✓ **میانگین:** رایج‌ترین شاخص مورد استفاده در میان شاخص‌های تمرکز داده می‌باشد.
- ✓ **میانه:** در مقایسه با میانگین، نسبت به داده‌های پرت حساسیت کمتری دارد.
- ✓ **مد:** برای تحلیل داده‌های گسسته (متغیرهای اسمی یا ترتیبی) شاخص مناسبی است.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های پراکندگی داده‌ها

دامنه میانی چارکها:
Interquartile range (IQR)

میزان کوچکی یا بزرگی این بازه می‌تواند یک دید اولیه نسبت به وضعیت پراکندگی داده‌ها به تحلیل‌گر ارائه‌کند.
در ادامه و در قسمت نمودار جعبه ای، بیشتر درمورد این شاخص بحث خواهیم کرد

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های پراکندگی داده‌ها

واریانس و انحراف معیار:

واریانس= متوسط مجموع مربعات انحراف تک‌تک مشاهدات، از میانگین مشاهدات

واریانس نمونه:

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

واریانس جامعه

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

انحراف معیار (انحراف استاندارد) = ریشه دوم واریـانس

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های پراکندگی داده‌ها

ضریب تغییرات:
Coefficient of Variation (CV)

یکی از معایب انحراف استاندارد آنست که بصورت مطلق بوده و مقیاس مقادیر داده‌ها را درنظر نمی‌گیرد.
در ضریب تغییرات، میزان پراکندگی بصورت نسبتی از میانگین ارائه می‌شود.

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

ضریب تغییرات واحد (بُعد) ندارد. به همین دلیل معیار مناسبی برای مقایسه داده‌های آماری است که واحدهای مختلفی دارند.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های پراکندگی داده‌ها

مثال: فرض کنید مقادیر میانگین و انحراف استاندارد سرعت خودروها در دو قطعه بزرگراهی بصورت زیر است:

قطعه ۱: $\bar{X}_1 = 84^{km/h}$; $S_1 = 20^{km/h}$; $CV_1 = 0.24$

قطعه ۲: $\bar{X}_2 = 103^{km/h}$; $S_2 = 20^{km/h}$; $CV_2 = 0.19$

بدین‌ترتیب می‌توان گفت با وجود برابر بودن انحراف استاندارد سرعت‌ها در هر دو قطعه، سرعت خودروها در قطعه اول نسبت به قطعه دوم پراکندگی نسبی بیشتری دارند.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان عدم‌تقارن داده‌ها

چولگی:
Skewness

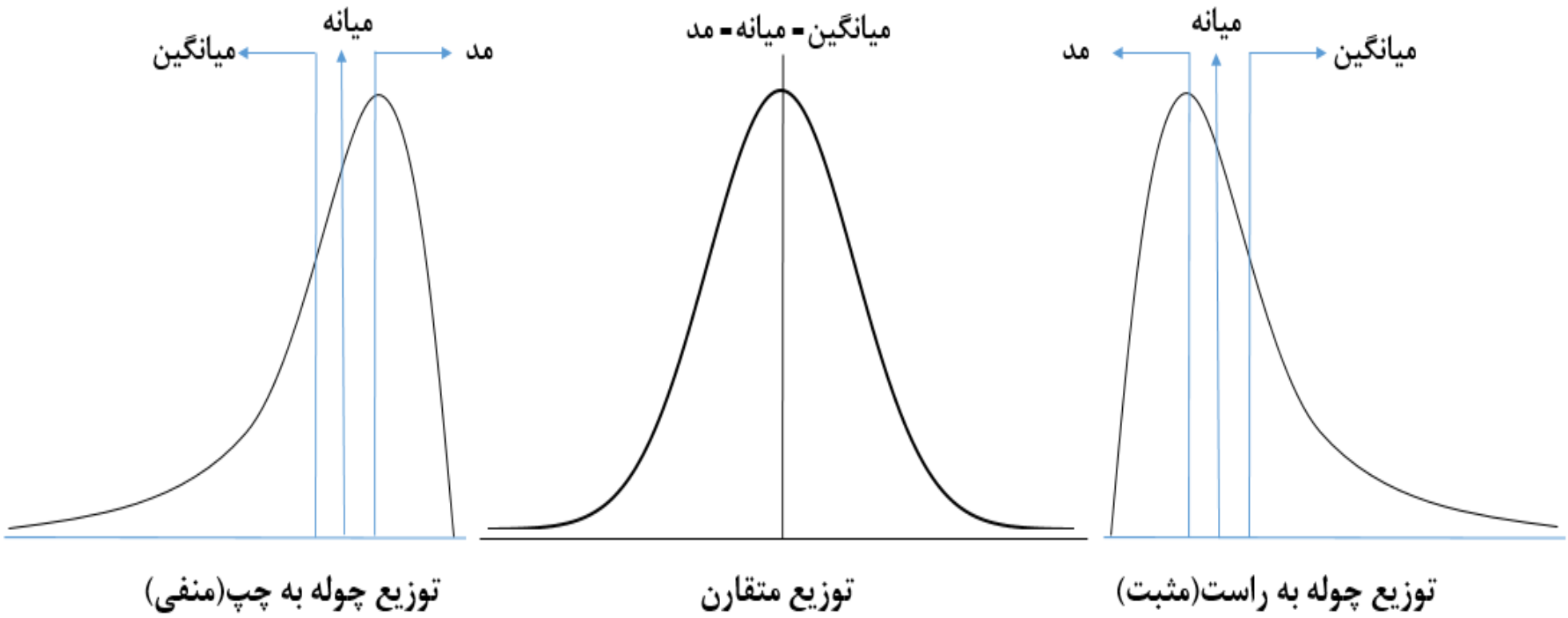
چولگی(یا گشتاور سوم حول میانگین)، شاخصی برای ارزیابی میزان نامتقارن‌بودن یک نمودار توزیع فراوانی **نسبت به میانگین** می‌باشد.

- **"چوله‌به‌راست" (یا دارای چولگی مثبت)** *(Skewed to the right (or positively skewed)*
- **متقارن**
- **"چوله‌به‌چپ" (یا دارای چولگی منفی)** *(Skewed to the left (or negatively skewed)*

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان عدم‌تقارن داده‌ها



مثال: توزیع نرمال

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان عدم‌تقارن داده‌ها

به عنوان یک قاعده سرانگشتی؛

- اگر مقدار چولگی بین -0.5 تا 0.5 باشد؛ می‌توان گفت توزیع داده‌ها نزدیک به قرینه است.
- اگر مقدار چولگی بیشتر از ۱ و یا کمتر از -1 باشد؛ میزان نامتقارن بودن توزیع داده‌ها بسیار بالاست.
- چنانچه مقدار ضرایب چولگی و کشیدگی در بازه (۱،۱-) باشد، می‌توان گفت توزیع داده‌ها "تاحدودی" نرمال است.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان عدم‌تقارن داده‌ها

کشیدگی:
Kurtosis

گشتاور چهارم حول میانگین

- شاخصی برای میزان تخت‌بودن یک توزیع فراوانی نسبت به یک توزیع نرمال با واریانس برابر می‌باشد.
- به عبارت دیگر، کشیدگی معیاری از بلندی (تیزی) منحنی در نقطه ماکزیمم است.
- میزان **تمرکز داده‌ها در اطراف میانگین** و همچنین **شدت دُم‌های توزیع** را نشان می‌دهد.

به بیان ساده، کشیدگی مشخص می‌کند که آیا توزیع:

- دارای **قله تیز** و **دُم‌های سنگین** است.
- یا **قله تخت** و **دُم‌های سبک** دارد.

مقدار کشیدگی برای یک توزیع نرمال =۳

نکته: در بسیاری از نرم‌افزارهای آماری (مانند R و SPSS) معمولاً از شاخصی به نام **Excess Kurtosis** استفاده می‌شود:

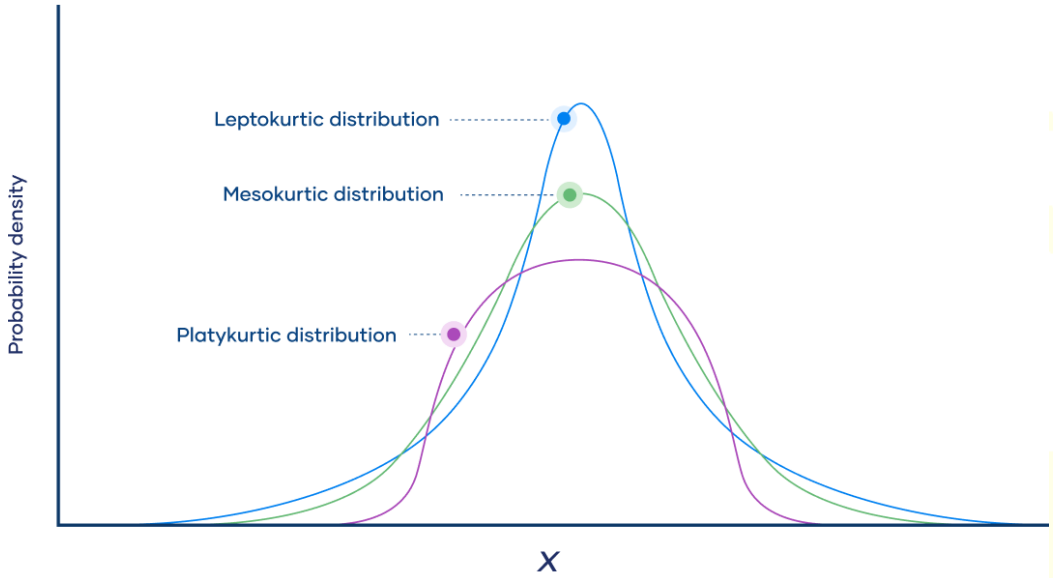
Excess Kurtosis=Kurtosis−3

در این حالت: مقدار کشیدگی توزیع نرمال = صفر

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان عدم‌تقارن داده‌ها



کشیدگی:
Kurtosis

Leptokurtic

✓ **(منحنی کشیده).** اگر مقدار ضریب کشیدگی برای یک توزیع فراوانی بیشتر از ۳ باشد؛ توزیع دارای **دُم‌های سنگین‌تر** و **تمرکز بیشتری در مرکز** نسبت به توزیع نرمال با واریانس برابر است (**یعنی احتمال وقوع مقادیر بسیار دور از میانگین بیشتر است**).

Platykurtic

✓ **(منحنی پَخ).** مقدار کمتر از ۳ برای کشیدگی بدین معناست که توزیع دارای **دُم‌های سبک‌تر** و **تمرکز مرکزی کمتر** نسبت به توزیع نرمال با واریانس برابر است (احتمال وقوع مقادیر حدی کمتر از حالت نرمال است).

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ شاخص‌های میزان همبستگی

همبستگی میان دو متغیر تصادفی، شاخصی از میزان **خطی بودن** رابطه میان آن دو می‌باشد.

ضریب همبستگی (ρ):
Correlation Coefficient

یکی از پارامترهای رایج برای ارزیابی میزان خطی بودن رابطه میان دو متغیر است.
مقادیر ممکن برای ضریب همبستگی در بازه [-1,1] قرار می‌گیرد.

- ✓ $\rho = 0$: عدم وجود رابطه **خطی** میان دو متغیر
- ✓ مقادیر مثبت ρ : رابطه خطی مستقیم
- ✓ مقادیر منفی ρ : رابطه خطی معکوس

❑ رایج‌ترین ضرایب همبستگی عبارتند از **ضریب همبستگی پیرسون** و **رتبه‌ای اسپیرمن**.

شاخص‌های عددی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

Statistics		
متوسط_سرعت		
N	Valid	1870
	Missing	0
Mean		31.02
Median		30.64
Mode		29
Std. Deviation		7.316
Variance		53.530
Skewness		0.215
Kurtosis		-0.124
Range		50
Minimum		6
Maximum		56
Percentiles	25	25.86
	50	30.64
	75	35.94
	85	38.68

→ excess kurtosis

مثال(خروجی نرم افزار): داده‌های مربوط به نمونه-گیری از سرعت متوسط خودروها در قطعه‌ای از یک خیابان

فصل۲. آمار توصیفی

CHAPTER 2: DESCRIPTIVE STATISTICS

- ✓ **روش‌های ترسیمی برای توصیف داده‌ها**

مدرس: محمدمهدی بشارتی

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

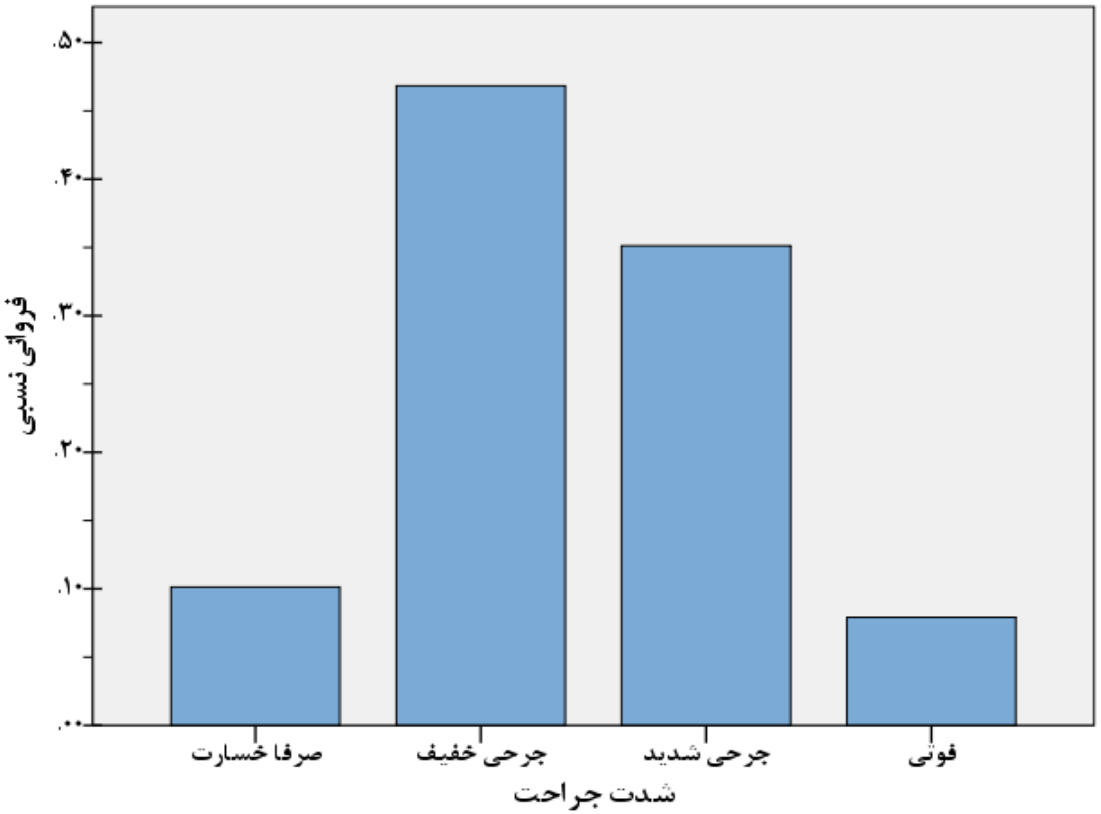
روش‌های ترسیمی می‌تواند برای اهداف زیر مورد استفاده قرار گیرد؛

- ✓ شناسایی روابط و روندها، داده‌های پرت
- ✓ بررسی وضعیت داده‌ها به‌صورت چشمی و کیفی

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

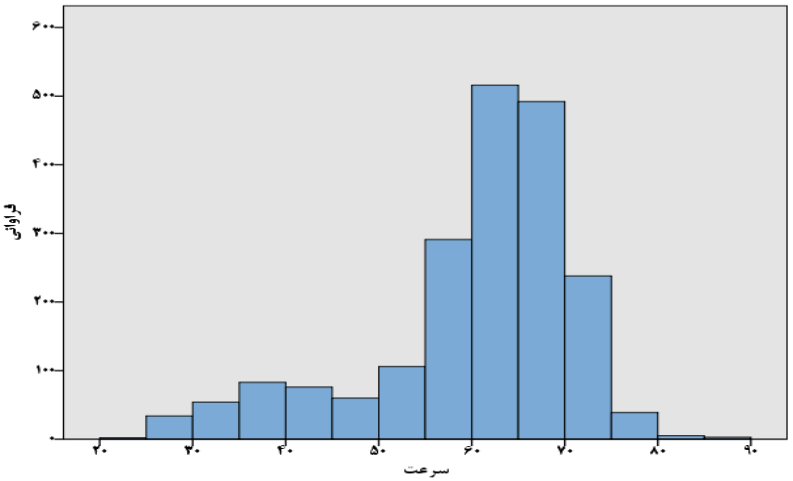
❖ نمودار میله‌ای (Bar Chart)



نموداری‌است که فراوانی نسبی سطوح مختلف یک "متغیر کیفی(اسمی یا ترتیبی)" را به‌صورت میله نمایش می‌دهد.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی



❖ نمودار مستطیلی یا هیستوگرام (Histogram)

نموداری است که فراوانی نسبی طبقات مختلف یک "متغیر کمیِ پیوسته" را نمایش می‌دهد.

تفاوت میان نمودار میله‌ای و مستطیلی

- ✓ برخلاف نمودار میله‌ای، میان ستون‌های مستطیلی **در نمودار مستطیلی**، فاصله‌ای وجود نداشته و این ستون‌ها به صورت به‌هم پیوسته رسم می‌شوند.
- ✓ **در نمودار میله‌ای**، هریک از ستون‌ها نماینده یک طبقه معین از یک **متغیر طبقه‌ای** می‌باشد. اما در **نمودار مستطیلی**، هرستون، نماینده بازه‌ای از مقادیر یک **متغیر کمی** می‌باشد.
- ✓ شاخص‌های چولگی و کشیدگی، تنها برای **نمودارهای مستطیلی** قابل بررسی خواهد.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ **نمودار مستطیلی یا هیستوگرام** (Histogram)

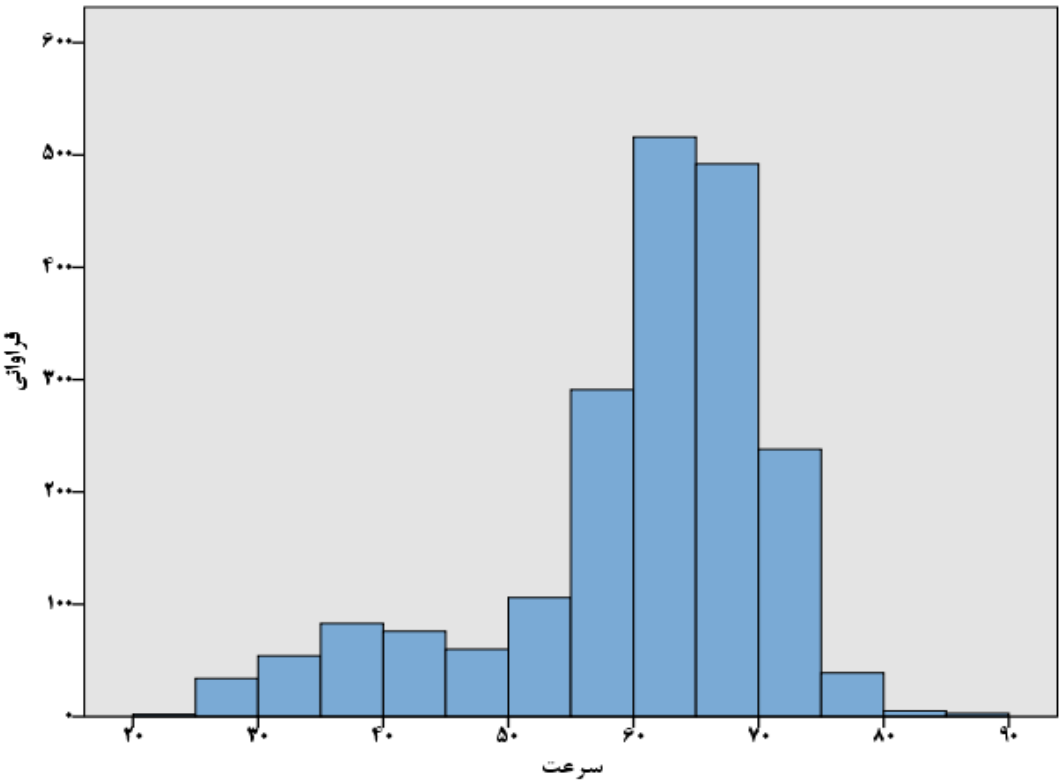
مهمترین موارد کاربرد نمودار هیستوگرام عبارتند از:

- نمایش فراوانی نسبی طبقات مختلف یک متغیر
- نمایش و بررسی چشمی عدم‌تقارن در داده‌ها
- استخراج چولگی و کشیدگی

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار مستطیلی یا هیستوگرام (Histogram)



مثال: داده‌های سرعت برداشت شده در یک قطعه از یک جاده دوخطه دوطرفه برون‌شهری

یافته:

همان‌طور که مشاهده می‌شود، توزیع داده‌ها چولگی منفی داشته و ***فراوانی سرعت‌های پایین، بیشتر از حدِ نرمال است.***

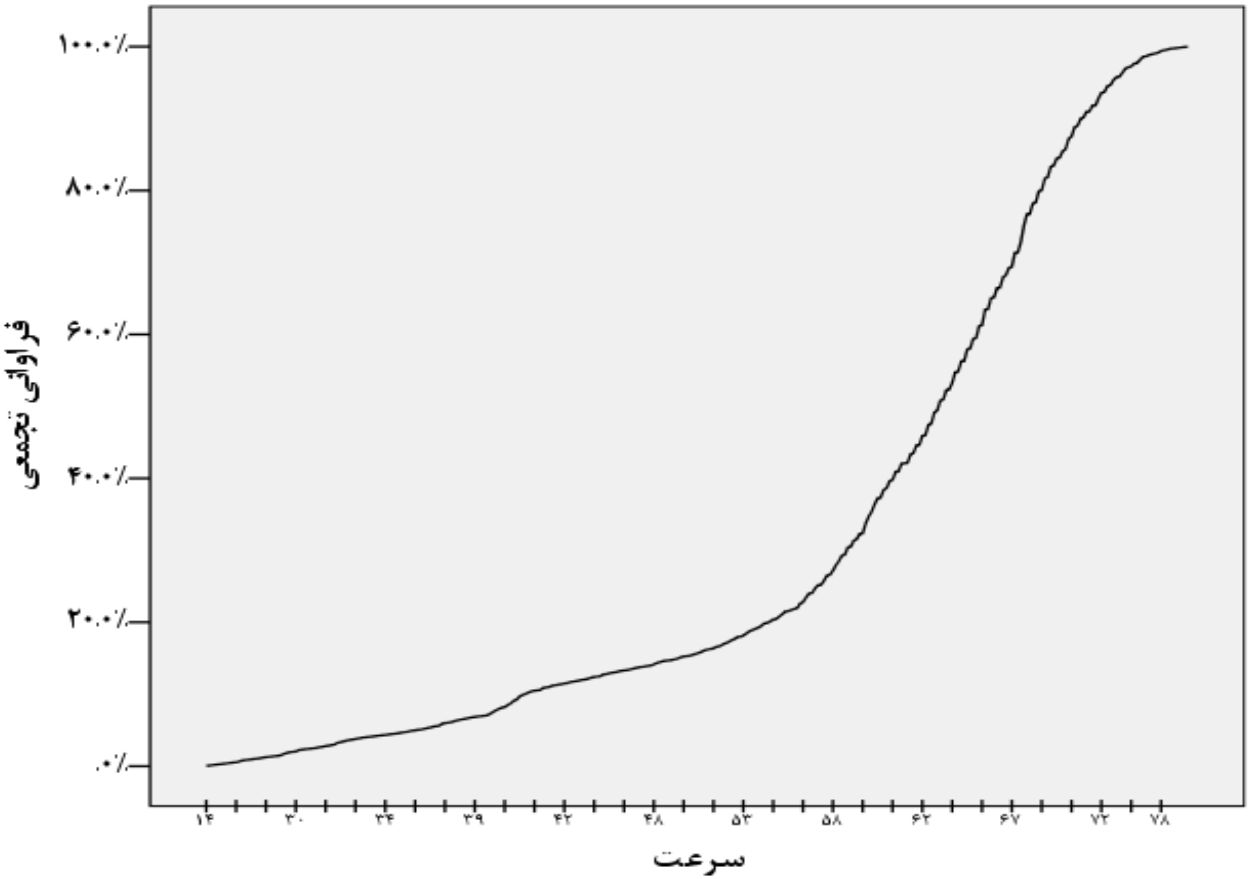
تفسیر:

پراکندگی زیاد سرعت‌ها می‌تواند موجب کاهش ایمنی ترافیک در این قطعه شود. به‌خصوص اینکه جاده مورد مطالعه دو خطه دوطرفه بوده و تردد برخی خودروها با سرعت‌های پایین می‌تواند موجب افزایش سبقت و تجاوز به خط مقابل شود.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار طاقدیس (Ogive)



طاقدیس‌ها در واقع، نمودارهای فراوانی تجمعی هستند.

محور عمودی این نمودار: درصد فراوانی تجمعی

به کمک این نمودارها می‌توان صدک‌ها را شناسایی کرد (مثلا سرعت ۸۵٪)

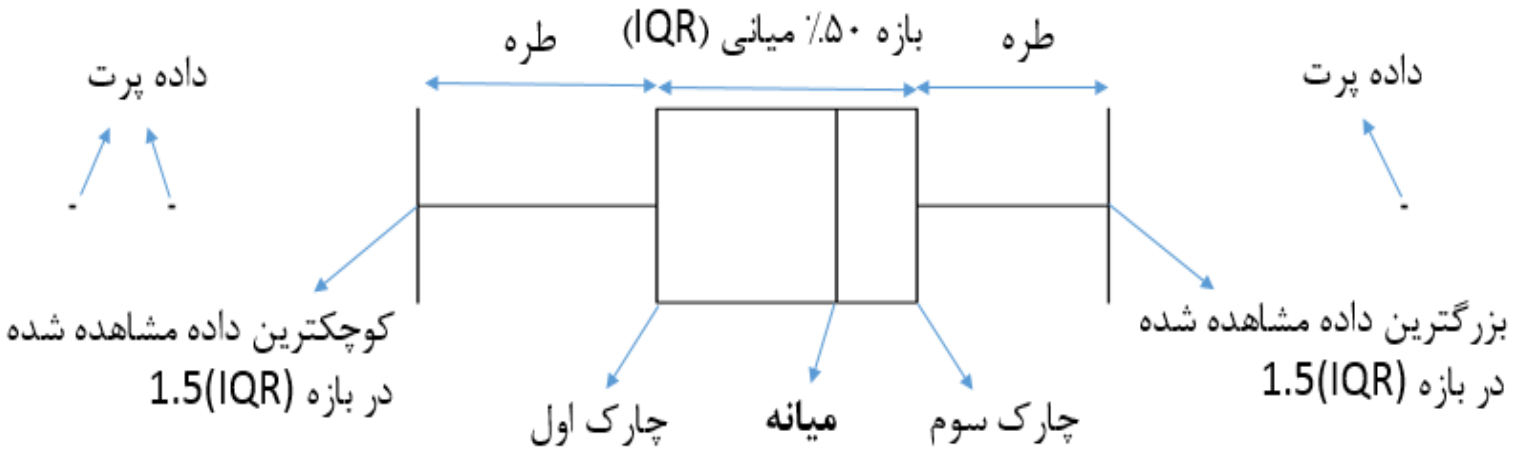
روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار جعبه‌ای (Box plot)

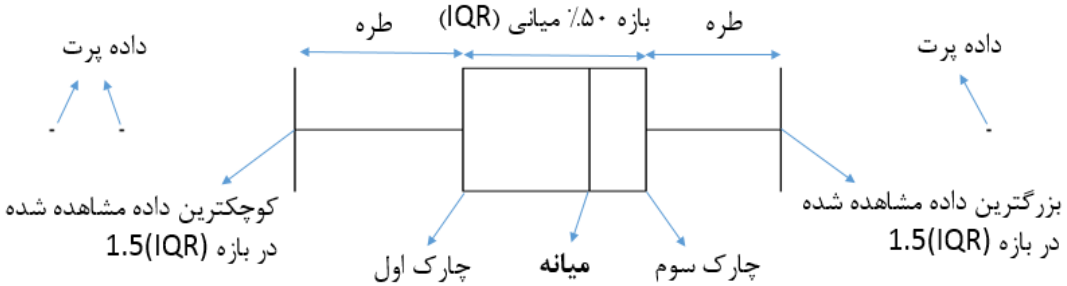
نمایش همزمان ۵ شاخص در مورد وضعیت توزیع داده‌ها:

1. بزرگ‌ترین داده مشاهده شده،
2. کوچک‌ترین داده مشاهده شده،
3. چارک اول،
4. میانه
5. چارک سوم



روش‌های ترسیمی برای توصیف داده‌ها

❖ نمودار جعبه‌ای (Box plot)



موارد استفاده از نمودار جعبه‌ای

1. **بررسی نحوه پراکندگی داده‌ها**

چگونه؟

به کمک دامنه میانی چارک‌ها (**IQR**) و نیز طول طره‌های دو طرف جعبه (هرچه جعبه کشیده‌تر باشد، بدین معناست که پراکندگی داده‌ها بیشتر است و بالعکس.)

2. **شناسایی وضعیت تمرکز داده‌ها**

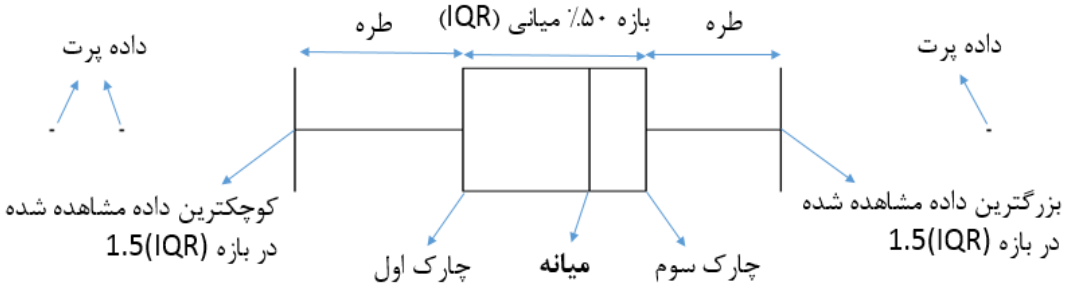
چگونه؟

به کمک خط میانه داده‌ها و میزان کشیده بودن/نبودن بازه **IQR**

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار جعبه‌ای (Box plot)



موارد استفاده از نمودار جعبه‌ای

3. مشاهده داده‌های پرت

چگونه؟

با تعریف بازه "۱.۵ برابر دامنه میانی چارک‌ها" (1.5×IQR) بعنوان محدوده قابل‌قبول، داده‌هایی که خارج از این بازه در دو طرف جعبه نمودار قرار می‌گیرند را به عنوان «داده پرت» بررسی می‌کنیم.

4. شناسایی وجود چولگی درون داده‌ها

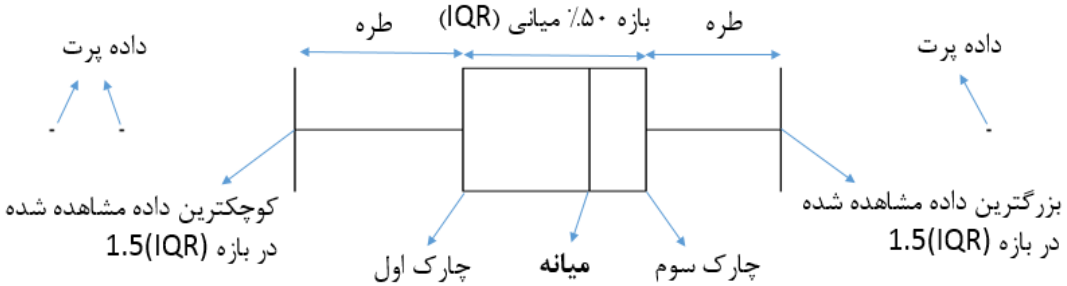
چگونه؟

با بررسی مکان میانه درون جعبه نمودار (اگر **طول قسمت سمت راست جعبه**، بیشتر از قسمت سمت چپ باشد، توزیع داده‌ها **چوله به راست** می‌باشد و برعکس)

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار جعبه‌ای (Box plot)



موارد استفاده از نمودار جعبه‌ای

5. **بررسی همزمان فراوانی مقادیر یک متغیر در چند جامعه مختلف**

چگونه؟

با کنار هم قرار دادن چند نمودار جعبه‌ای

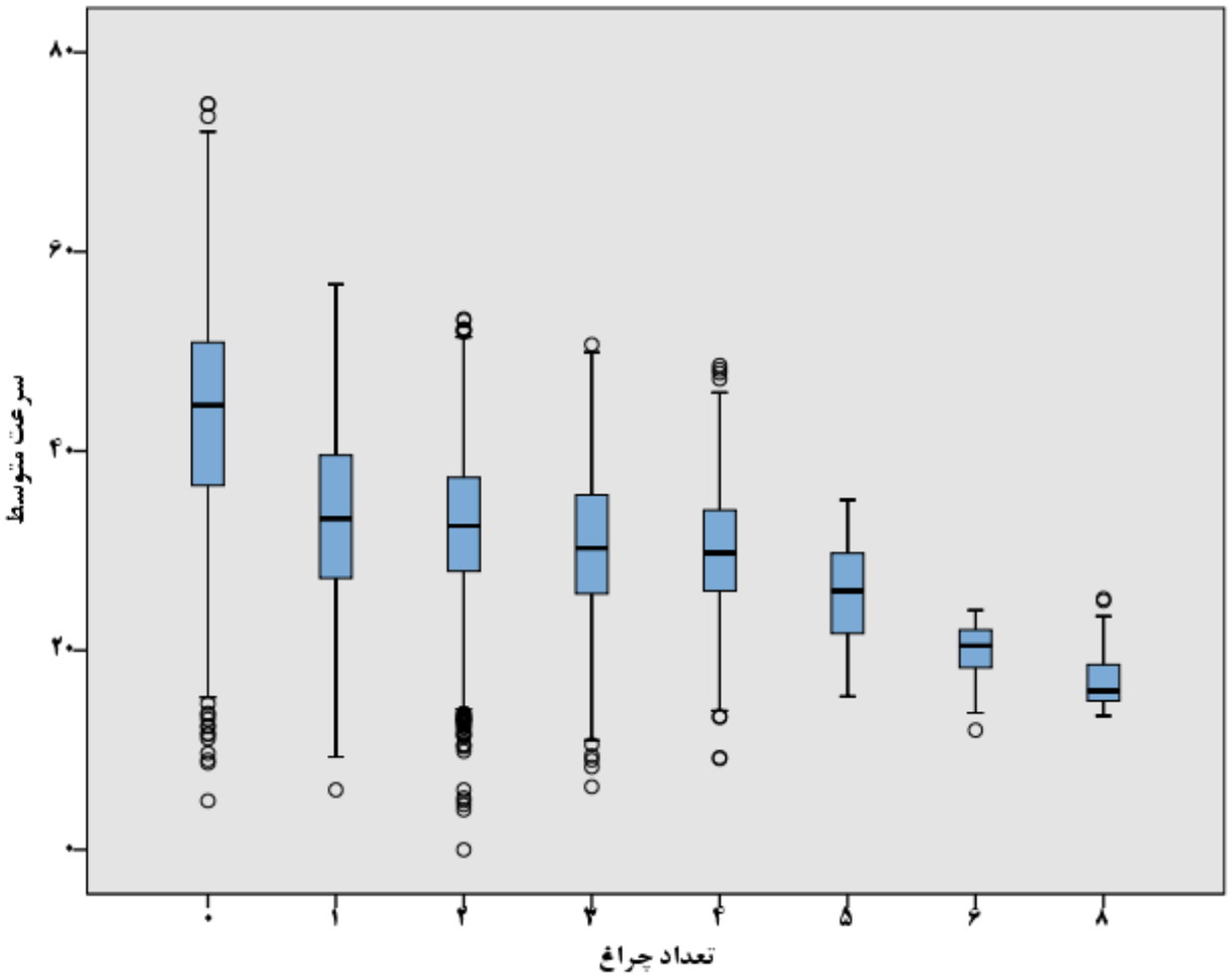
روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار جعبه‌ای (Box plot)

مثال:

چند خیابان با طول، حجم تردد و سایر شرایط مساوی را درنظر بگیرید که تنها تفاوت آن‌ها اینست که در هرکدام تعداد معینی از چراغ راهنمایی، جریان حرکت وسایل‌نقلیه را کنترل می‌کند. می‌خواهیم با بررسی همزمان نمودار جعبه‌ای سرعت‌های مشاهده‌شده در این خیابان‌ها، تاثیر تعداد چراغ‌های راهنمایی را بر توزیع سرعت متوسط وسایل‌نقلیه بررسی کنیم؟

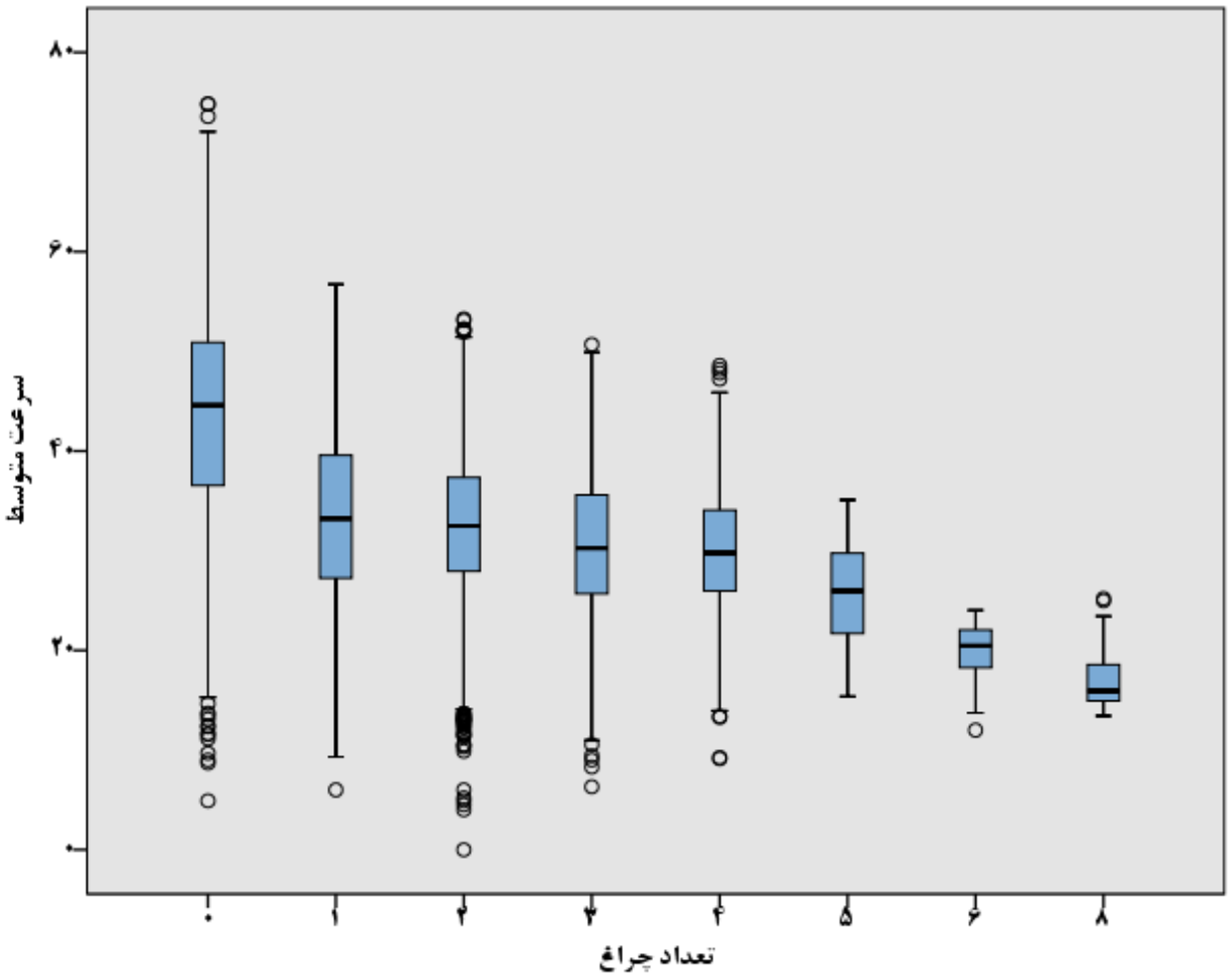


سوال: چه نکاتی از این نمودار، قابل درک است؟

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار جعبه‌ای (Box plot)



تفسیر نمودار:

- ✓ با افزایش تعداد چراغ‌های راهنمایی در یک کریدور، میانه سرعت خودروها، کاهش یافته است.
- ✓ کاهش میانه سرعت‌ها به ازای افزایش تعداد چراغ، بصورت خطی نبوده و نرخ افزایشی دارد.
- ✓ ازآنجا که با افزایش تعداد چراغ‌ها، طول جعبه‌ها و طره‌ها کاهش یافته، می‌توان نتیجه گرفت که با افزایش تعداد چراغ‌ها، واریانس سرعت‌ها کاهش می‌یابد. به‌طوری‌که به ازای ۸ چراغ‌راهنمایی، سرعت متوسط خودروها تفاوت زیادی با هم ندارند.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ داده های پرت (Outliers)

منظور از داده پرت، مشاهداتی است که به نوعی با سایر داده‌های درون مجموعه تحت مطالعه متفاوت بوده و ممکن‌است توجه تحلیل‌گر را به خود جلب کند.

- **توجه:** داده‌های پرت ممکن‌است **مهم‌ترین نقاط** در میان مجموعه داده‌های مورد مطالعه باشند!
  - ✓ لازم‌است پس از شناسایی و قبل از حذف آن‌ها از میان داده‌ها، مورد بررسی بیش‌تر و دقیق‌تر قرار گیرند.
  - ✓ با حذف داده‌های پرت، این خطر وجود دارد که تنها آن چیزی را مشاهده کنیم که از قبل انتظار مشاهده آن را داشته‌ایم.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ داده های پرت (Outliers)

❑ **وجود داده‌های پرت می‌تواند به دلایل زیر باشد؛**

1. وجود خطا در اندازه‌گیری یا ثبت داده‌ها،
2. تعلق نداشتن داده به جامعه مورد مطالعه (یا ناهمگنی جامعه)،
3. یک رویداد نادر در جامعه‌ای که به شدت چولگی دارد.

❑ **روش های جایگزین به جای حذف کردن داده‌های پرت؛**

- ✓ ممکن‌است توزیع واقعی داده‌ها، دارای چولگی بوده و با یک تبدیل (مانند تبدیل لگاریتمی) بتوان بدون حذف داده‌های پرت، توزیع داده‌ها را متقارن نمود.
- ✓ استفاده از روش‌هایی که نیازمند تقارن یا نرمال بودن توزیع داده‌ها نیستند (روش‌های ناپارامتری)

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار چندک–چندک (q-q)

یکی از انواع نمودارهای آماری برای بررسی میزان تطابق توزیع داده‌های نمونه با یک توزیع نظری خاص و یا توزیع داده‌های یک نمونه دیگر می‌باشد.

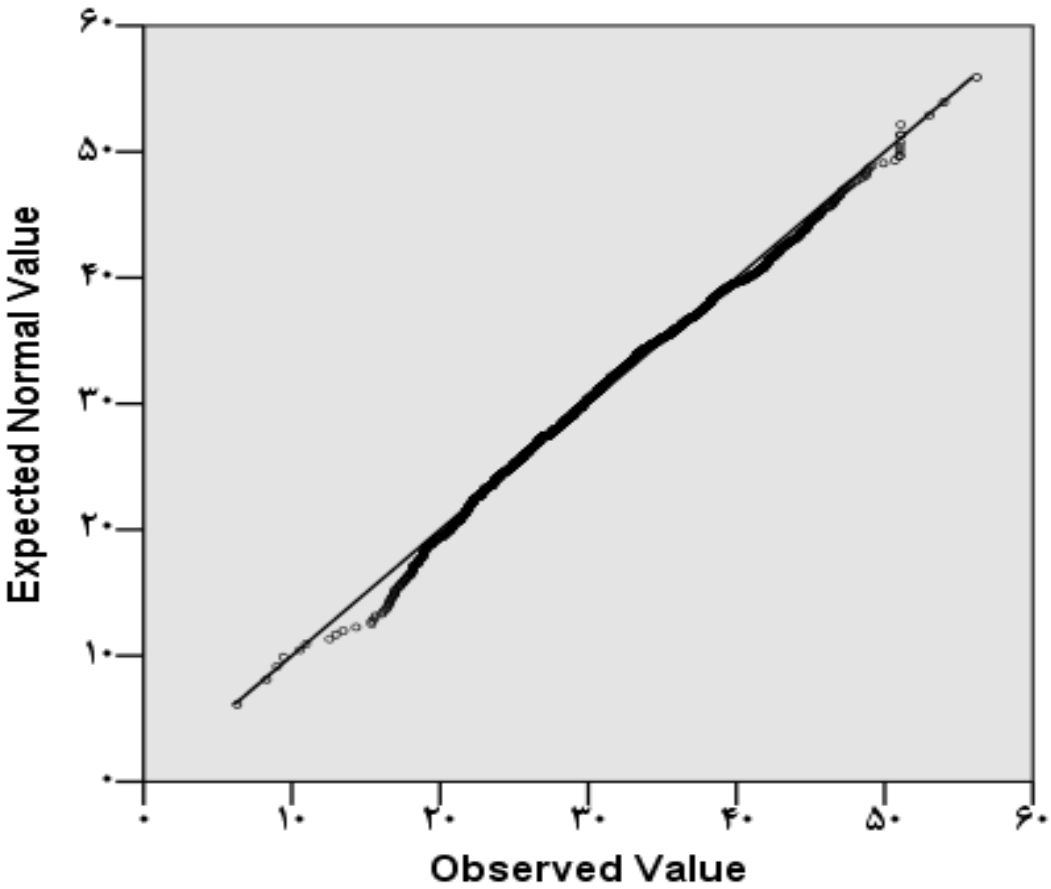
هرچه توزیع داده‌های مشاهده‌شده (مثلا داده‌های سرعت تردد) با توزیع موردنظر (مثلا توزیع نرمال)، انطباق بیشتری داشته‌باشد؛ نمودار رسم شده به یک خط مستقیم، نزدیک‌تر خواهد بود.

روش‌های ترسیمی برای توصیف داده‌ها

❖ نمودار چندک–چندک (q-q)

مثال:

به‌کمک نمودار چندک‌ها، می‌خواهیم بدانیم آیا داده‌های سرعت از توزیع نرمال پیروی می‌کنند یا خیر؟



نکته: هیچ مدلی به‌طور ۱۰۰درصد بر داده‌ها منطبق نخواهد بود.

در عمل، هنگامی‌که داده‌های نمونه به‌صورت مارپیچ به فاصله بسیار کمی حول خط مورب قرار بگیرند؛ می‌توان گفت که توزیع داده‌ها با توزیع موردنظر تطابق دارد.

روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار پراکنشی (Scatter plot)

مناسب‌ترین نمودار برای بررسی رابطه (خطی و غیرخطی) میان دو متغیر پیوسته

مثلا:

رابطه میان افزایش کرایه‌ها و تقاضای سفر اتوبوس بین‌شهری

برای رسم این نمودار، مقادیر یک متغیر در محور قائم و مقادیر متغیر دیگر در محور افقی قرار می‌گیرد.

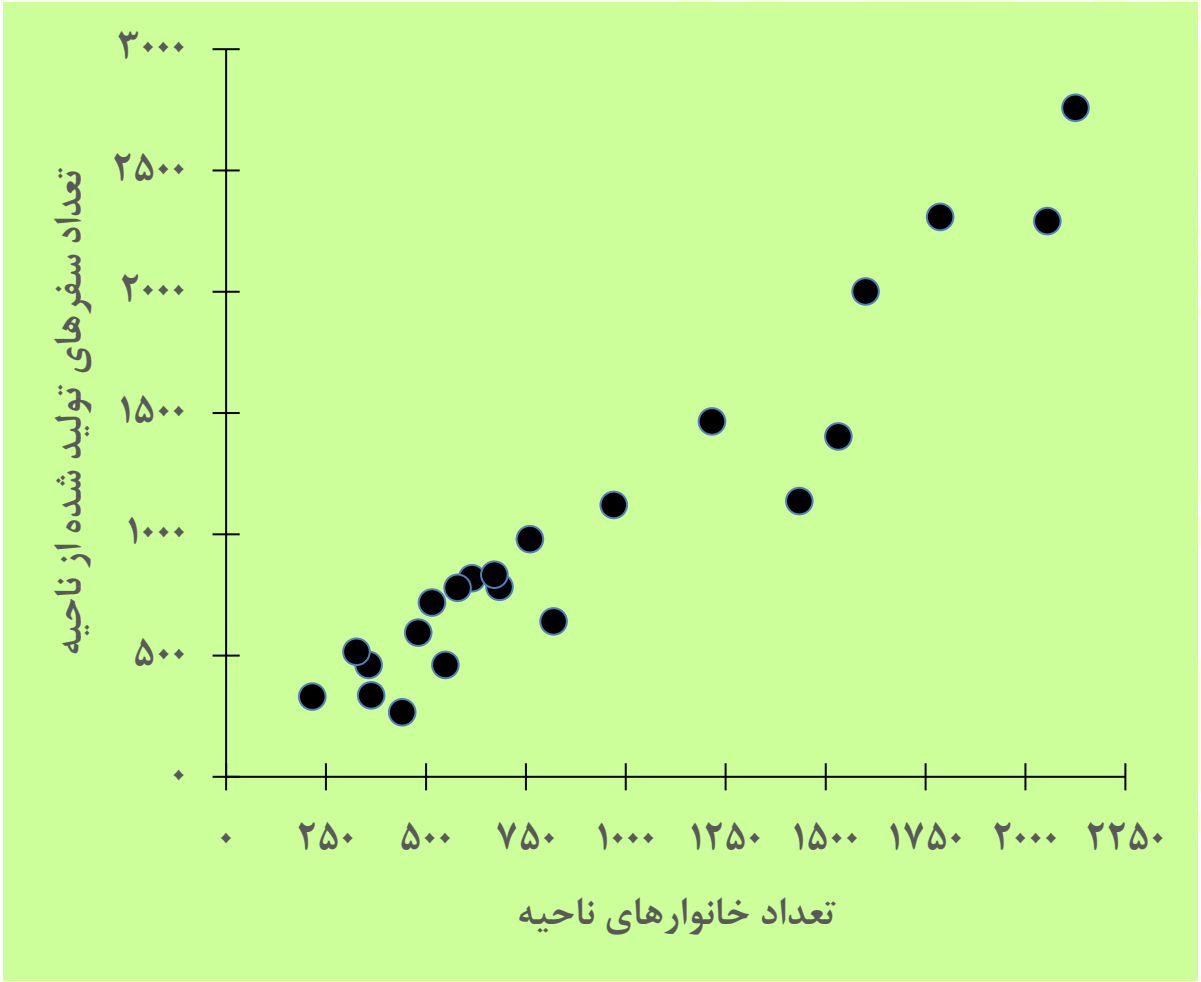
روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار پراکنشی (Scatter plot)

مثال:

تعداد سفرهای تولیدشده نواحی ترافیکی یک شهر را در محور عمودی و تعداد خانوارهای ساکن در هر ناحیه را در محور افقی

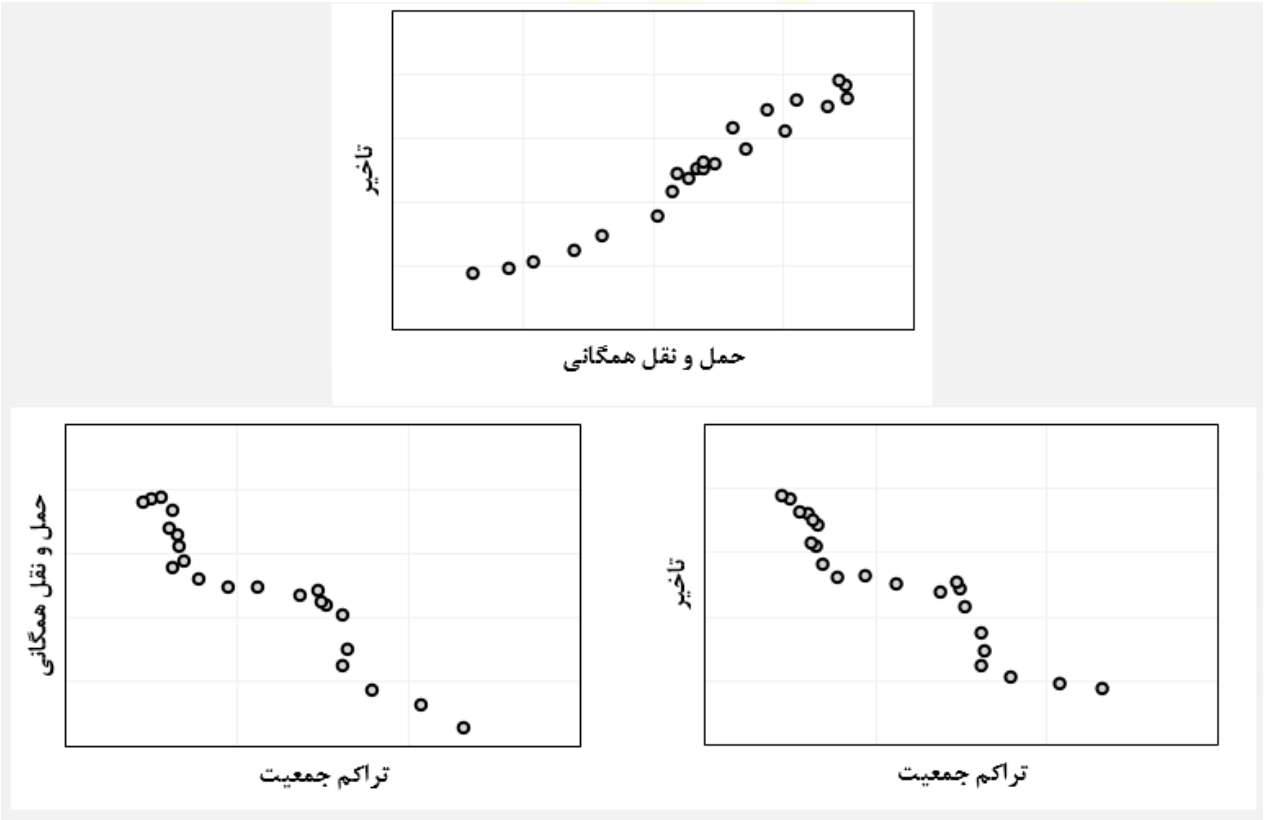


روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ نمودار پراکنشی (Scatter plot)

اگر بیش از دو متغیر پیوسته داشته‌باشیم و بخواهیم رابطه آن‌ها را با یکدیگر بررسی کنیم؛ می‌توانیم یک نمودار پراکنشی ماتریسی(Scatter plot matrix) ترسیم نماییم.



روش‌های ترسیمی برای توصیف داده‌ها

مدرس: محمدمهدی بشارتی

❖ **نمودار پراکنشی** (Scatter plot)

توجه: همبستگی میان مقادیر دو متغیر، لزوما به این معنا نیست که میان آنها رابطه علت‌ومعلولی وجود دارد.

سوال: تفاوت میان «رابطه همبستگی» با «رابطه علت و معلول» چیست؟