فصل هفتم

شاخص‌های پراکندگی

فصل ۷ــ جلسه اوّل

مفاهیم: پراکندگی، دامنه تغییرات، چارک بالا، چارک پایین

مهارت‌ها: حذف مناسب داده‌های نامربوط

اهداف:

- ارزشیابی شاخص‌های مرکزی با استفاده از شاخص‌های پراکندگی
- استخراج اطلاعاتی دربارهٔ توزیع داده‌ها با استفاده از شاخص‌های پراکندگی
- مقایسهٔ شاخص‌های مرکزی و شاخص‌های پراکندگی با شاخص‌های هندسی

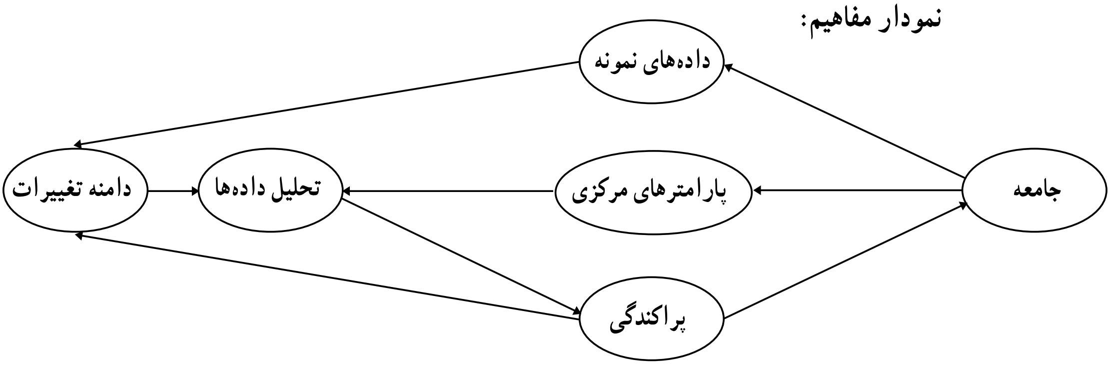
تأکید درس: کاربرد تحلیل داده‌ها با استفاده از شاخص‌های مرکزی بدون در دست داشتن شاخص‌های پراکندگی عملاً ممکن نیست. مثال معدل را در کلاس به بحث بگذارید و از دانش‌آموزان روشی بخواهید که پراکندگی داده‌ها از معدل را نیز مدنظر قرار دهند و سعی کنند برای این پراکندگی شاخصی معرفی کنند. دامنه تغییرات به‌عنوان ابتدایی‌ترین شاخص پراکندگی اهمیت دارد. دانش‌آموزان باید سعی کنند دامنه تغییرات را در تحلیل پراکندگی به‌کار ببرند. شاخص‌های پراکندگی به نوعی شهود هندسی در مورد ارزش شاخص‌های مرکزی به‌دست می‌دهند. این که تغییرات داده‌ها چگونه دامنه تغییرات و سایر شاخص‌های پراکندگی را تحت تأثیرقرار می‌دهد مورد تأکید است. حذف داده‌های نامربوط یا نامناسب باید با توجه به مسئله مورد تحقیق صورت بگیرد. هرگز داده‌ها را بدون تأکید بر صورت مسئله حذف ننمایید. دانش‌آموزان باید توانایی شاخص‌های مرکزی و پراکندگی در تحلیل داده‌ها را با شاخص‌های هندسی و نمودارها مقایسه کنند. بدون این مقایسه این که هر کدام چرا اهمیت دارند آشکار نمی‌شود.

بحث: تأثیر تغییرات داده‌ها بر دامنه تغییرات را به بحث بگذارید. یک مسئله واقعی انتخاب کنید و داده‌های آن را با روش‌های متفاوتی حذف کنید و از دانش‌آموزان بخواهید نتایج حاصل را تحلیل کنند و در مورد قابل قبول بودن روش‌های حذف متفاوت بحث کنند. از دانش‌آموزان بخواهید برای اندازه‌گیری پراکندگی شاخص‌هایی عددی بسازند. حذف چارک بالا و پایین از داده‌ها را در کلاس به بحث بگذارید : آیا فکر می‌کنید روش‌هایی کلی از حذف داده که وابسته به مسئله مورد تحقیق

باشد وجود دارد؟ دانش‌آموزان باید بدانند که انحراف از تصمیم‌گیری مناسب هم نتیجه جمع‌آوری داده غیرمناسب است و هم نتیجه تحلیل نامناسب.

مباحث ریاضی مربوط: نمودار جعبه‌ای

نمودار مفاهیم:



فصل ۷ ــ جلسه دوّم

مفاهیم: واریانس، انحراف معیار، پراکندگی

مهارت‌ها: محاسبه واریانس و انحراف معیار، تحلیل پراکندگی داده‌ها، واحد واریانس و انحراف میانگین و مقایسهٔ آن‌ها

اهداف:

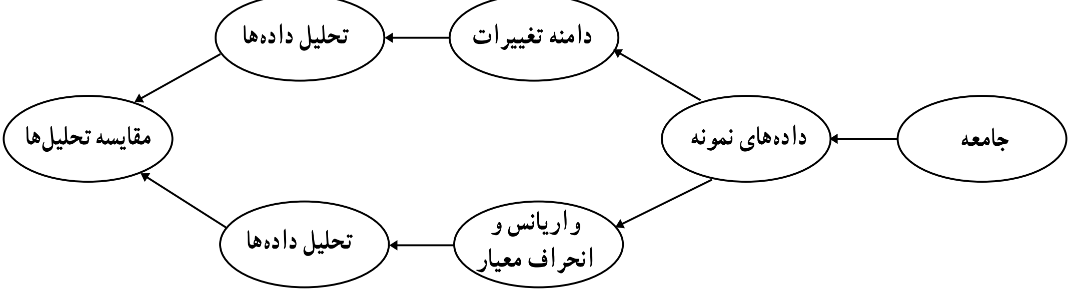
- درک تأثیر تغییرات داده‌ها بر شاخص‌های پراکندگی
- کاربرد شاخص‌های پراکندگی در زندگی روزمره
- کاربرد شاخص‌های پراکندگی در ارزشیابی

تأکید درس: این که چه تغییراتی در داده‌ها بر واریانس تأثیر می‌گذارد و چه تغییراتی تأثیر نمی‌گذارد مورد تأکید است.

بحث: این که آیا واریانس یک شاخص پراکندگی طبیعی است را در کلاس به بحث بگذارید. واحد واریانس و واحد انحراف معیار را با واحد میانگین مقایسه کنید و با توجه به این مقایسه، مناسب بودن آن‌ها به عنوان شاخص پراکندگی را نقد کنید. آیا این شاخص را می‌توان به سادگی در زندگی روزمره به‌کار برد؟ واریانس و انحراف معیار را برای نمرات دروس مختلف حساب کنید و در مورد پراکندگی نمرات از میانگین نمرات تحلیل ارائه دهید. نمودار جعبه‌ای را با واریانس و انحراف معیار مقایسه کنید. کدام یک اطلاعات بیشتری در مورد پراکندگی می‌دهند؟ در نمودار جعبه‌ای میانه را با میانگین جایگزین کنید و نمودار دیگری بسازید. این نمودار جدید را با واریانس و انحراف معیار مقایسه کنید. کدام یک اطلاعات بیشتری در مورد پراکندگی داده‌ها می‌دهند؟

مباحث ریاضی مربوط و پراکندگی: جامعه آماری، میانگین

نمودار مفاهیم:



فصل ۷ ــ جلسه سوم

مفاهیم: ضریب تغییرات، میانگین، پراکندگی در منحنی نرمال

مهارت‌ها: محاسبه ضریب تغییرات، تحلیل شاخص‌های پراکندگی

اهداف:

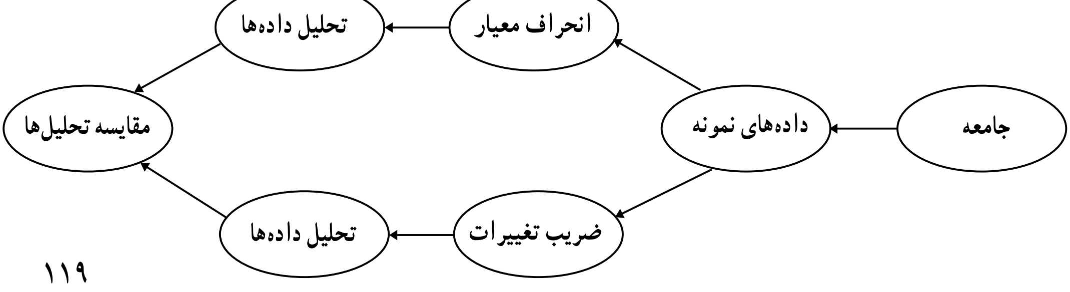
- تحلیل داده‌ها توسط شاخص‌های مرکزی و شاخص‌های پراکندگی
- پراکندگی داده‌ها مستقل از واحد اندازه‌گیری قابل بررسی است.
- نقد کاربرد میانگین در زندگی روزمرّه با استفاده از شاخص‌های پراکندگی

تأکید درس: این که چه تغییراتی در داده‌ها بر ضریب تغییرات تأثیر می‌گذارد و چه تغییراتی تأثیر نمی‌گذارد مورد تأکید است.

بحث: انحراف معیار را در حالتی که پراکندگی داده‌ها از توزیع نرمال پیروی می‌کند در نظر بگیرید. در این حالت انحراف معیار اطلاعات نسبتاً خوبی در مورد پراکندگی می‌دهد. پراکندگی در منحنی نرمال را چگونه می‌توان در زندگی روزمره به‌کار برد؟ این سؤال را در کلاس به بحث بگذارید. اگر بدانیم که منحنی از توزیع نرمال پیروی می‌کند آیا می‌توانید از روی داده‌ها میانگین و انحراف معیار را محاسبه کنید؟ بدون اینکه از فرمول‌های محاسبهٔ آن‌ها استفاده کنید؟

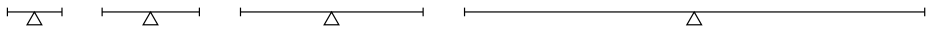
مباحث ریاضی مربوط: منحنی نرمال، ضریب تغییرات، شاخص‌های پراکندگی

نمودار مفاهیم:



شاخص‌های پراکندگی

در فصل پارامترهای مرکزی گفتیم، میانگین مانند مرکز ثقل عمل می‌کند. یعنی اگر جسم را در نقطه مرکز ثقل آن روی تکیه‌گاهی قرار دهیم، جسم به صورت متعادل باقی می‌ماند. حال می‌خواهیم بدانیم تعادل این جسم چقدر پایدار است. در شکل‌های زیر



(الف) (ب) (ج) (د)

مسلماً هر چقدر از چپ به سمت راست یا برعکس حرکت کنیم تعادل جسم ناپایدارتر خواهد بود. ضربه کوچکی که به یکی از دو سر جسم در شکل (د) وارد شود ممکن است آن را به راحتی از حالت تعادل خارج کند و حال آنکه در شکل (الف) ممکن است تعادل جسم به‌هم نخورد. علت این پابرجایی تعادل در شکل‌های بالا را به‌خوبی می‌توانید از نظر فیزیکی توضیح دهید. عدم پایداری در شکل‌های (ج) و (د) بـه علت آن است که در ایـن دو حالت نقاط بسیار دور از هم وجـود دارد ولی در شکل الف و ب نقاط به‌هم نزدیک‌اند. در کتاب مثال رفتن از خانه به مدرسه را آوردیم که در آن دانش‌آموز براساس میانگین تصمیم می‌گیرد کی از خانه خارج شود ولی هر چقدر داده‌ها از میانگین دورتر باشند اطمینان به‌موقع رسیدن کمتر خواهد بود. پارامتری که میزان پایداری تعادل و یا اطمینانی که به تصمیمی که گرفته‌ایم اندازه می‌گیرد پراکندگی است.

با این مقدمه ممکن است این‌طور تصور شود که پراکندگی چیز خوبی نیست. درست است که در این مثال‌ها پراکندگی زیاد سبب عدم تعادل می‌شود، همان‌طوری که پراکندگی زیاد در درآمدها نشانه‌ای از بیماری اقتصاد است. امّا در برخی از موارد، پراکندگی زیاد یک معیار خوب خواهد بود. مثلاً در یک آزمون ورودی اگر آزمون به قسمی باشد که تمام افراد نمرات ۱۹، ۲۰ و یا برعکس نمرات ۳،۲ و یا حتی ۱۴، ۱۵ بگیرند، این آزمون نتوانسته است هدف از برگزاری آن را برآورده کند. در آزمون‌ها می‌خواهند بهترین‌ها را انتخاب کنند، لذا باید نمرات متفاوت باشد تا بتوانند داوطلبین را براساس آن‌ها مرتب کنند. اگر نمرات بسیار به هم نزدیک باشد این مقایسه به‌خوبی انجام نمی‌شود. آزمونی خوب است که واریانس آن زیاد باشد، که در این صورت می‌گویند این آزمون افراد را خوب سنجیده است. شما از دانش‌آموزان بخواهید واریانس نمرات درس‌های متفاوت را حساب کنند، یعنی مثلاً نمرات درس آمار و مدلسازی را در نظر گرفته، واریانس آن را حساب کنند. به همین ترتیب واریانس نمرات درس فارسی را هم حساب کنند و الی آخر (مثلاً در پنج درس). با مقایسه واریانس‌های این نمرات متوجه خواهید شد که کدام درس بهتر می‌تواند قابلیت‌های دانش‌آموزان را بسنجد. معمولاً در مطالعات آماری، متغیری که پراکندگی آن کم باشد به‌عنوان ثابت در نظر گرفته می‌شود و از

مطالعه حذف می‌شود. مثلاً اگر شما متغیرهای نمره، قد و سن دانش‌آموزان کلاس را بخواهید با هم مطالعه کنید، از آن‌جایی که سن آن‌ها پراکندگی خیلی کمی خواهد داشت، اطلاع زیادی دربارهٔ این جامعه به شما نخواهد داد و لذا می‌توانید آن را حذف کنید.

همان‌طوری که در کتاب درسی اشاره شد دامنهٔ تغییرات و واریانس دو شاخص مهم در اندازه‌گیری پراکندگی در جامعه هستند که واریانس به علت درگیر شدن با تمام داده‌ها و هم‌چنین وجود قضایا و روابط خوب ریاضی برای آن از پارامترهای مهم پراکندگی است.

علت استفاده از انحراف معیار نیز ذکر شد. در کتاب درسی بیان شد که همواره

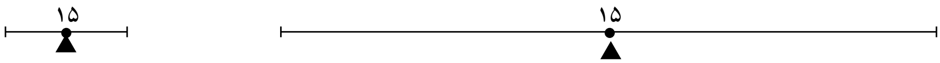
$$\sum(X_i - \overline{X})^۲ \lneqq \sum (X_i - a)^۲$$

شاخص دیگری کـه در بعضی از کتاب‌ها به عنوان شاخص پراکندگی ذکر می‌شود به صورت زیر است.

$$\frac{۱}{n}\sum|X_i - \overline{X}|$$

این شاخص هم شاخص خوبی است ولی به علت وجود قدرمطلق محاسبهٔ آن آسان نیست و به‌علاوه خواص خوب واریانس را ندارد. به این علت کمتر از آن استفاده می‌شود و ما هم در کتاب به آن اشاره نکردیم.

معمولاً درگزارش‌های آماری میانگین را ذکر می‌کنند و این با توجه به توضیحاتی که گذشت نمی‌تواند جامعه را برای ما ترسیم کند. مثلاً اگر بگویند میانگین ۱۵ است آن‌گاه نمی‌دانیم با کدام حالت زیر سر و کار داریم.



بنابراین لازم است شاخص پراکندگی هم در کنار شاخص مرکزی جهت ترسیم بهتر جامعه بیان شود.

معمولاً در مطالعات اجتماعی، اقتصادی دو جامعه (یا دو کشور) را ممکن است براساس درآمد سرانه مقایسه کنند. مثلاً بگویند درآمد سرانه کارکنان ایران و یکی از کشورهای همسایه برابر است. این مقایسه دلیل بر یکسان بودن وضعیت کارکنان در این دو کشور نیست.

اولاً درکشوری که شاخص قیمت‌های آن پایین است، کارکنان از رفاه بیشتری برخوردارند ثانیاً حتی اگر شاخص قیمت‌ها هم یکی باشد، رفاه و سلامت اقتصادی در کشوری بیشتر است که پراکندگی درآمد در آن کم باشد. آنچه که از نظر اقتصادی بر افراد یک جامعه فشار می‌آورد زیاد بودن فاصله درآمدهاست نه زیاد بودن درآمدها. پس در این قبیل موارد لازم است واریانس را نیز مدنظر قرار داد. از بهترین شرایط در این مسأله استفاده از ضریب تغییرات است. ممکن است میانگین درآمد در دو جامعه یکی نباشد که در این صورت مقایسه واریانس‌ها هم کمک بزرگی نخواهد کرد. امّا ضریب

تغییرات به‌خوبی می‌تواند جامعه‌ای را که از سلامت اقتصادی بیشتری برخوردار است انتخاب کند.

برای محاسبه واریانس از دستورهای زیر استفاده می‌شود :

(استفاده از داده‌های خام)

$$\frac{۱}{n}\sum(X_i-\overline{X})^۲$$

$$\frac{۱}{n}\sum X_i^۲-(\overline{X})^۲$$

(در جدول فراوانی)

$$\frac{۱}{n}\sum f_i(X_i-\overline{X})^۲ \quad \text{و} \quad \overline{X}=\frac{۱}{n}\sum f_iX_i$$

$$\frac{۱}{n}\sum f_iX_i^۲-(\overline{X})^۲$$

ممکن است در برخی از کتاب‌ها دیده باشید که برای واریانس دستور زیر را به کار برده‌اند.

$$\frac{۱}{n-۱}\sum(X_i-\overline{X})^۲$$

آنچه که در عبارت بالا آمده است محاسبه واریانس نیست بلکه *برآورد نااریب* واریانس است که ما در کتاب درسی ذکری از این اصطلاحات به میان نیاورده‌ایم.

پاسخ تمرینات فصل

پاسخ تمرینات صفحهٔ ۱۴۷

۱ــ

$$\overline{X}_A \approx ۴۸/۰۳۲$$

$$\overline{X}_B \approx ۴۸/۹۹$$

الف : کارخانه B شکلات بیشتری را می‌فروشد. از شاخص میانگین استفاده شده است.

ب : کارخانهٔ B یکنواخت‌تر عمل نموده است.

$$R_۱ = ۴۸/۳۲-۴۷/۸۴=۰/۴۸$$

$$R_۲ = ۴۹/۱۶-۴۸/۸۴=۰/۳۲$$

۲ــ تمام داده‌ها برابر و یکسان هستند.

۳ــ با تغییر واحد اندازه‌گیری، دامنه تغییرات در واحد ضرب می‌شود (لازم به تذکر است که در این‌جا دقت اندازه‌گیری موردنظر نمی‌باشد). در مثال اگر داده‌ها برحسب متر اندازه‌گیری شده و سپس با سانتی‌متر، دامنه تغییرات در ۱۰۰ ضرب می‌شود.

۴ــ اگر به تمام داده‌ها مقدار ثابتی افزوده شود، دامنه تغییرات تغییر نمی‌کند.

پاسخ سؤالات صفحات ۱۵۶ــ۱۵۴

۱ــ الف : طول ۱۰ گیاه بین ۵۰ تا ۶۰ سانتی‌متر است.

ب : طول ۲۵ گیاه بین ۷۰ تا ۹۰ سانتی‌متر است.

ج :

$$\overline{X}=\frac{(۲\times۱۵)+(۵\times۲۵)+(۷\times۳۵)+(۹\times۴۵)+(۱۰\times۵۵)+(۱۵\times۶۵)+(۱۵\times۷۵)+(۱۰\times۸۵)+(۹\times۹۵)}{۸۲}\approx ۶۲/۹۳$$

$$\sqrt{\sigma^۲}=\sqrt{\frac{۲\times(۱۵-۶۲/۹۳)^۲+(۵)\times(۲۵-۶۲/۹۳)^۲+۷\times(۳۵-۶۲/۹۳)^۲+۹\times(۴۵-۶۲/۹۳)^۲+۱۰\times(۵۵-۶۲/۹۳)^۲+۱۵\times(۶۵-۶۲/۹۳)^۲+۱۵\times(۷۵-۶۲/۹۳)^۲+۱۰\times(۸۵-۶۲/۹۳)^۲+۹\times(۹۵-۶۲/۹۳)^۲}{۸۲}}$$

$$\approx ۲۱/۲۸$$

د : متقارن نبودن شکل نمایان‌گر این مطلب است که میانه و میانگین برابر نیستند.

۲ــ الف

$$\overline{X}=\frac{۳۸\times۲+۳۹\times۲+۴۰\times۲+۴۱\times۲}{۸}=۳۹/۵$$

$$\sqrt{\sigma^۲}=\sqrt{\frac{۲(۳۸-۳۹/۵)^۲+۲(۳۹-۳۹/۵)^۲+۲(۴۰-۳۹/۵)^۲+۲(۴۱-۳۹/۵)^۲}{۸}}$$

$$\approx ۳/۱۶$$

$$\overline{X}=\frac{۳۸+۳۸/۴+۳۹+۳۹/۴+۳۹/۸+۴۰/۲+۴۰/۹+۴۱/۲}{۸}$$

$$=\frac{۳۱۶/۹}{۸}=۳۹/۶۱۲۵$$

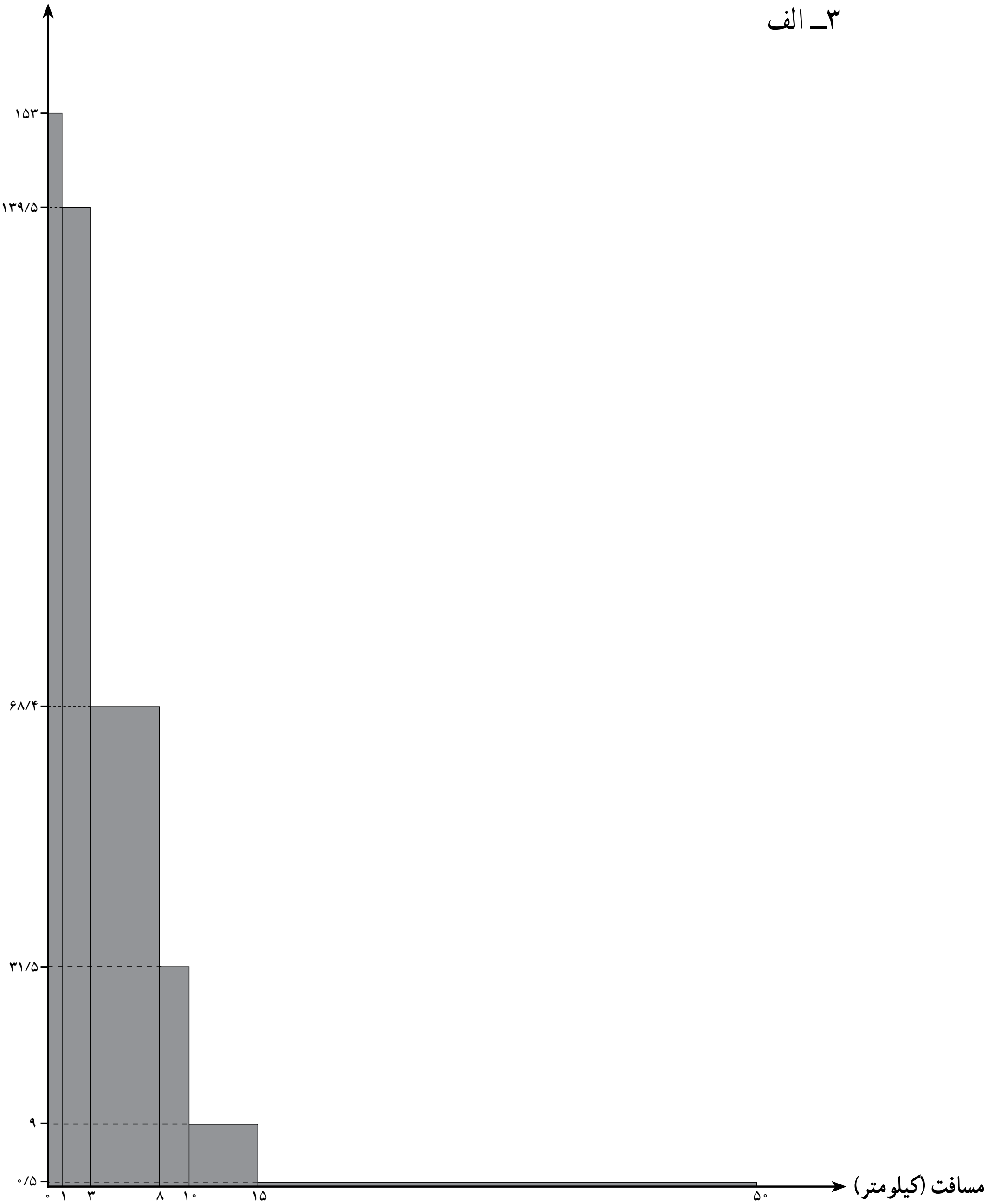
$$\sqrt{\sigma^۲}=\sqrt{\frac{(۳۸-\overline{X})^۲+(۳۸/۴-\overline{X})^۲+(۳۹-\overline{X})^۲+(۳۹/۴-\overline{X})^۲+(۳۹/۸-\overline{X})^۲+(۴۰/۲-\overline{X})^۲+(۴۰/۹-\overline{X})^۲+(۴۱/۲-\overline{X})^۲}{۸}}$$

$$\approx ۳/۰۱$$

ب : میانگین کمتر شده و انحراف معیار افزایش پیدا نمود.

ج : به‌طورکلی بله

۳ــ الف



$$۹۰۰ \times \frac{۱۷}{۱۰۰} = ۱۵۳$$

$$۹۰۰ \times \frac{۳۱}{۱۰۰} = ۲۷۹$$

$$۹۰۰ \times \frac{۳۸}{۱۰۰} = ۳۴۲$$

$$۹۰۰ \times \frac{۷}{۱۰۰} = ۶۳$$

$$۹۰۰ \times \frac{۵}{۱۰۰} = ۴۵$$

$$۹۰۰ \times \frac{۲}{۱۰۰} = ۱۸$$

ب :

$$\overline{X} = \frac{(۱۵۳ \times ۰/۵) + (۲۷۹ \times ۲) + (۳۴۲ \times ۵/۵) + (۶۳ \times ۹) + (۴۵ \times ۱۲/۵) + (۱۸ \times ۳۲/۵)}{۹۰۰}$$

$$= ۴/۷$$

ج : زنان

۴ــ الف

$$\overline{X}_۱ = \frac{۲+۳+۴+۵+۶}{۵} = ۴$$

$$\overline{X}_۲ = ۲ + \overline{X}_۱ = ۲ + ۴ = ۶$$

$$\sigma_۱^۲ = \frac{(۲-۴)^۲ + (۳-۴)^۲ + (۴-۴)^۲ + (۵-۴)^۲ + (۶-۴)^۲}{۵} = \frac{۴+۱+۰+۱+۴}{۵}$$

$$= \frac{۱۰}{۵} = ۲$$

$$\sigma_۲^۲ = \frac{(۴-۶)^۲ + (۵-۶)^۲ + (۶-۶)^۲ + (۷-۶)^۲ + (۸-۶)^۲}{۵} = \frac{۴+۱+۰+۱+۴}{۵}$$

$$= \frac{۱۰}{۵} = ۲$$

ب : اگر به هر یک از داده‌های گروه اول ۲ واحد افزوده شود، داده‌های گروه دوم به‌دست می‌آیند.

ج : واریانس تغییری ننموده است.

د : با افزودن عددی ثابت به تمام داده‌ها در یک گروه، میانگین داده‌ها نیز به همان اندازه افزایش می‌یابد. در نتیجه انحراف هر یک از داده‌ها از میانگین نسبت به حالت اولیه تفاوتی نخواهد داشت پس مجموع مجذور انحرافات از میانگین نیز تغییری نخواهد کرد. یعنی واریانس بدون تغییر خواهد بود.

۵ ــ الف

$$\overline{X}_۱ = \frac{۱+۲+۳+۴+۵}{۵} = ۳$$

$$\overline{X}_۲ = ۳(\overline{X}_۱) = ۳(۳) = ۹$$

$$\sigma_۱^۲ = \frac{(۱-۳)^۲+(۲-۳)^۲+(۳-۳)^۲+(۴-۳)^۲+(۵-۳)^۲}{۵} = \frac{۴+۱+۰+۱+۴}{۵} = ۲$$

$$\sigma_۲^۲ = \frac{(۳-۹)^۲+(۶-۹)^۲+(۹-۹)^۲+(۱۲-۹)^۲+(۱۵-۹)^۲}{۵} = \frac{۳۶+۹+۰+۹+۳۶}{۵} = ۱۸$$

ب : اگر هر یک از داده‌های اول را سه برابر کنیم، داده‌های گروه دوم به‌دست می‌آیند.

ج : با سه برابر کردن داده‌های گروه اول، واریانس گروه دوم $۳^۲ = ۹$ برابر شد.

د : با سه برابر کردن داده‌ها، میانگین گروه دوم سه برابر میانگین گروه اول شد. بدین ترتیب انحراف هر یک از داده‌ها از میانگین نیز سه برابر شد. پس مجذور انحراف هر یک از داده‌ها، ۹ برابر انحراف داده‌ها از میانگین در گروه اول خواهد شد. پس واریانس ۹ برابر می‌شود.

۶ ــ اگر $x_۱, x_۲, \ldots, x_n$ مفروض باشند :

$$\overline{X}_۱ = \frac{x_۱ + x_۲ + \ldots + x_n}{n}$$

حال به هر یک از این داده‌ها عدد ثابت a افزوده شود :

$$\overline{X}_۲ = \frac{(x_۱ + a) + (x_۲ + a) + \ldots + (x_n + a)}{n} = a + \overline{X}_۱$$

پس

$$\sigma_۲^۲ = \frac{\left(x_۱ + \cancel{a} - (\cancel{a} + \overline{X}_۱)\right)^۲ + \left(x_۲ + \cancel{a} - (\cancel{a} + \overline{X}_۱)\right)^۲ + \ldots + \left(\cancel{a} + x_n - (\cancel{a} + \overline{X}_۱)\right)^۲}{n}$$

$$= \frac{(x_۱ - \overline{X}_۱)^۲ + (x_۲ - \overline{X}_۱)^۲ + \ldots + (x_n - \overline{X}_۱)^۲}{n} = \sigma_۱^۲$$

۷ــ اگر $x_۱, x_۲, \ldots, x_n$ مفروض باشند :

$$\overline{X}_۱ = \frac{x_۱, x_۲, \ldots, x_n}{n}$$

حال اگر هر یک از این داده‌ها را در عددی ثابت چون a ضرب کنیم، خواهیم داشت :

$$\overline{X}_۲ = \frac{ax_۱ + ax_۲ + \ldots + ax_n}{n} = a\left(\frac{x_۱ + x_۲ + \ldots + x_n}{n}\right) = a\overline{X}_۱$$

پس :

$$\sigma_2^2=\frac{(ax_1-a\overline{X}_1)^2+(ax_2-a\overline{X}_1)^2+\ldots+(ax_n-a\overline{X}_1)^2}{n}$$

$$=\frac{\left(a(x_1-\overline{X}_1)\right)^2+\left(a(x_2-\overline{X}_1)\right)^2+\ldots+\left(a(x_n-\overline{X}_1)\right)^2}{n}$$

$$=a^2\left(\frac{(x_1-\overline{X}_1)^2+(x_2-\overline{X}_1)^2+\ldots+(x_n-\overline{X}_1)^2}{n}\right)=a^2\sigma_1^2$$

۸ ــ اگر $x_1,x_2,\ldots,x_n$ مفروض باشند و هر یک از این داده‌ها را در عددی ثابت ضرب کرده و با عددی ثابت جمع کنیم داریم : $ax_1+b, ax_2+b,\ldots,ax_n+b$

پس : $\overline{X}_2=a\overline{X}_1+b$

پس :

$$\sigma_2^2=\frac{\left(ax_1+\cancel{b}-(a\overline{X}_1+\cancel{b})\right)^2+(ax_2+\cancel{b}-(a\overline{X}_1+\cancel{b}))^2+\ldots+(ax_n+\cancel{b}-(a\overline{X}_1+\cancel{b}))^2}{n}$$

$$=\frac{\left(a(x_1-\overline{X}_1)\right)^2+\left(a(x_2-\overline{X}_1)^2\right)+\ldots+\left(a(x_n-\overline{X}_1)^2\right)}{n}$$

$$=a^2\left(\frac{(x_1-\overline{X}_1)^2+(x_2-\overline{X}_1)^2+\ldots+(x_n-\overline{X}_1)^2}{n}\right)$$

$$=a^2\sigma_1^2$$

۹ــ در تمرین قبل دیدیم که : $\sigma_{a_{x+b}}^2=a^2\sigma_x^2$

پس : $\sigma_{a_{x+b}}=\sqrt{a^2\sigma_x^2}=|a|\sigma_x$

۱۰ــ با توجه به شکل داریم :

تعداد فرزندان	تعداد زنان
۰	۵
۱	۱۰
۲	۶
۳	۳
۴	۱

الف : $\overline{X}=\frac{(0\times5)+(1\times10)+(2\times6)+(3\times3)+(4\times1)}{25}$

$$=\frac{۰+۱۰+۱۲+۹+۴}{۲۵}=\frac{۳۵}{۲۵}=\frac{۷}{۵}=۱/۴$$

ب :

$$\sqrt{\sigma_۲^۲}=\sqrt{\frac{۵(۰-۱/۴)^۲+۱۰(۱-۱/۴)^۲+۶(۲-۱/۴)^۲+۳(۳-۱/۴)^۲+۱(۴-۱/۴)^۲}{۲۵}}$$

$$\approx ۱/۰۶$$

ج : با توجه به این که متوسط تعداد فرزندان گروه دوم بیشتر از گروه اول است می‌توان گفت که مجموع تعداد فرزندان گروه دوم از گروه اول بیشتر است و از بیشتر بودن انحراف معیار در گروه دوم نسبت به گروه اول متوجه می‌شویم که تعداد فرزندان زنان در گروه دوم از پراکندگی بیشتری برخوردار بوده (در این مورد بیشتر بحث کنید.) ولی در گروه اول تعداد فرزندان این پراکندگی نسبت به گروه دوم کمتر مشهود است.

پاسخ تمرینات صفحات ۱۵۹ و ۱۶۰

۱ــ $\overline{X}=۴$ و $\sigma=۶ \Rightarrow CV=\frac{۶}{۴}=۱/۵$

۲ــ فرض مسئله نشان می‌دهد که هر یک از داده‌های شما با ۱۰ جمع شده‌اند. پس

$$\overline{X}_۲=۱۰+\overline{X}_۱\ ,\ \sigma_۲^۲=\sigma_۱^۲ \Rightarrow$$

$$CV_۲=\frac{\sigma_۱}{۱۰+\overline{X}_۱}$$

یعنی ضریب تغییرات کم‌تر می‌شود.

۳ــ هر یک از داده‌ها (نمره‌ها) در ۰/۲ ضرب شده‌اند. پس :

$$\overline{X}_۲=۰/۲\overline{X}_۱ \text{ و } \sigma_۲^۲=(۰/۲)^۲\sigma_۱^۲ \Rightarrow \sigma_۲=|۰/۲|\sigma_۱=۰/۲\sigma_۱$$

$$CV_۲=\frac{۰/۲\sigma_۱}{۰/۲\overline{X}_۱}=\frac{\sigma_۱}{\overline{X}_۱}=CV_۱$$

ضریب تغییرات تغییری نمی‌کند

۴ــ الف ــ پراکندگی بیشتری در بازار اول مشهود است.

ب ــ $R_۱=۱۳-۸=۵$ و $R_۲=۱۳-۸=۵$

ج ــ خیر

د ــ بازاری که قیمت کالا در آن پراکندگی کم‌تری دارد.

هـ ــ خیر

و ــ ضریب تغییرات :

$$\overline{X}_۱ = \frac{۸+۱۳+۹+۱۲+۱۰+۱۱+۱۲+۹+۱۰+۱۱+۱۰}{۱۱} =$$

$$= \frac{۱۱۵}{۱۱} \approx ۱۰/۴۶$$

$$\overline{X}_۲ = \frac{۱۰+۱۳+۸+۱۰+۹+۱۱+۱۰+۱۰+۱۱+۹+۱۰}{۱۱} = \frac{۱۱۱}{۱۱} \approx ۱۰/۱$$

$$\sigma_۱ = \sqrt{\frac{(۸-\overline{X}_۱)^۲ + (۱۳-\overline{X}_۱)^۲ + (۹-\overline{X}_۱)^۲ + (۱۲-\overline{X}_۱)^۲ + (۱۰-\overline{X}_۱)^۲ + (۱۱-\overline{X}_۱)^۲}{}}$$

$$\ldots\sqrt{\frac{+(۱۲-\overline{X}_۱)^۲ + (۹-\overline{X}_۱)^۲ + (۱۰-X_۱)^۲ + (۱۱-X_۱)^۲ + (۱۰-\overline{X}_۱)^۲}{۱۱}}$$

$$\approx \sqrt{\frac{۲۲/۶۸}{۱۱}} \approx ۱/۴۳$$

$$\sigma_۲ = \sqrt{\frac{(۱۰-\overline{X}_۲)^۲ + (۱۳-\overline{X}_۲)^۲ + (۸-\overline{X}_۲)^۲ + (۱۰-\overline{X}_۲)^۲ + (۹-\overline{X}_۲)^۲ + (۱۱-\overline{X}_۲)^۲}{}}$$

$$\ldots\sqrt{\frac{+(۱۰-\overline{X}_۲)^۲ + (۱۰-\overline{X}_۲)^۲ + (۱۱-\overline{X}_۲)^۲ + (۹-\overline{X}_۲)^۲ + (۱۰-\overline{X}_۲)^۲}{۱۱}}$$

$$\approx \sqrt{۱/۵۴} \approx ۱/۲۴$$

۵ ــ $\overline{X} - CV = (۱/۵)(۰/۵) = ۰/۷۵$ ☜ : استخر دوم

الف : استخری که انحراف معیار آن ۰ است مطمئن‌تر است. درست است که ۱/۴۵ از ۱/۵ کمتر است ولی می‌توان مطمئن بود که وارد قسمت‌هایی با عمق بیشتر نمی‌شود. (و نیز با پرشی می‌تواند روی آب بیاید).

ب : استخر سوم با انحراف معیار ۱/۰۵.

زیرا در این استخر قسمت‌هایی وجود دارد که عمق آن خیلی بیشتر از ۱/۵ متر است. با فرض آن که از بیرون استخر نمی‌توان عمق قسمت‌های مختلف آن را دانست، فرد مبتدی در این استخر با

خطر بیش‌تری روبرو است.

۶ ــ ضریب تغییرات برای دو سال ۶۵ و ۷۰ محاسبه و سپس مقایسه کنید.

۷ــ الف : یکی از سال‌ها را در نظر گرفته و ۱۰ مرکز به تصادف انتخاب کنید. پاسخ‌ها متفاوت است.

ــ متغیر مراکز استان‌ها، این متغیر کیفی اسمی است.

ــ متغیر میزان بارندگی، این متغیر کمّی پیوسته است.

ــ انحراف معیار

ــ میلی‌متر $\frac{۱۱۸۶+۱۲۹۸+۱۶۸۵+۱۳۱۹+۱۲۱۰+۱۰۱۵+۱۲۹۷}{۷(۱۲)} \approx ۱۰۷/۳$

متوسط بارندگی ماهانه در رشت.

ــ میزان بارندگی در سال‌های مختلف در رشت، کمی پیوسته.

فصل هشتم

همبستگی و رگرسیون

همانطور که در مقدمه کتاب آمده است تدریس فصل هشتم اختیاری است. در صورت کشش کلاس به‌خصوص در کلاس‌های ریاضی از معلمین تقاضا داریم این فصل را تدریس نمایند و ما را از نظرات خود مطلع سازند.

مفاهیم: نمودار پراکنش، ضریب همبستگی، خط رگرسیون

مهارت‌ها: رسم و تحلیل خط رگرسیون

اهداف:

- به‌وجود آوردن این احساس که روش‌های آماری ابزاری قدرتمند برای تصمیم‌گیری هستند.
- توجه به علم آمار به عنوان موضوعی که همراه با سایر مواد درسی ریاضی و آمیخته با آنان کارآمد است.
- پیش‌بینی رفتار متغیرها توسط خط رگرسیون
- مقایسه تغییرات متغیرها توسط خط رگرسیون

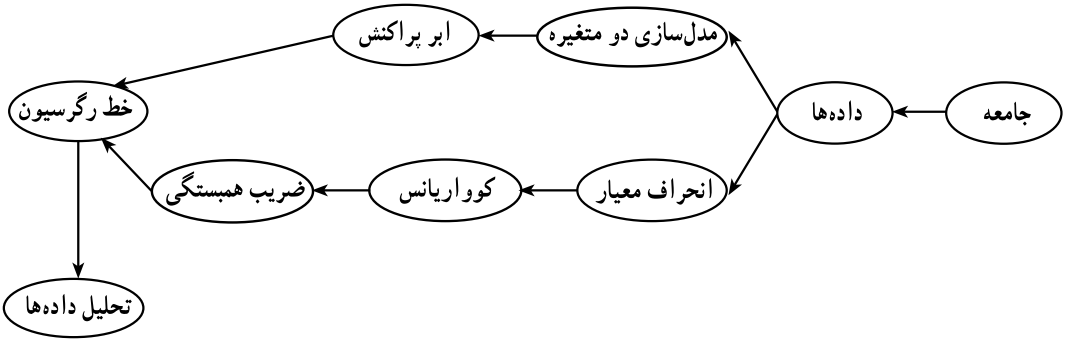
تأکید درس: جمع‌آوری، نمایش و کار با داده‌ها در زندگی ما از اهمیت روزمره برخوردارند. علوم طبیعی و اجتماعی بدون جمع‌آوری و خلاصه سازی و تحلیل داده‌ها در بسیاری از حیطه‌ها نمی‌توانند حضور پیدا کنند. دانش‌آموزان باید تشویق شوند روش‌های آماری را در دیگر رشته‌های درسی نیز به‌کار ببرند. دانش‌آموزان باید فرق بین درست و نادرست منطقی را با نتایج بررسی‌های آماری که کیفی هستند تشخیص دهند. پس ارائه مثال‌های نقض نمی‌تواند مانع تفکر آماری شود. بلکه روش‌های آماری یک پل میانی بین دقت ریاضی و طبیعت اطراف ما ایجاد می‌کند. در این فصل تأکید بر این است که دانش‌آموزان یاد بگیرند داده‌های عددی را هوشمندانه و مؤثر تحلیل کنند.

بحث: از دانش‌آموزان بخواهید در مورد این که چه خطی بهترین تقریب است تصمیم بگیرند و روشی ریاضی برای رسم بهترین خط ارائه دهند. این سؤال را مطرح کنید که آیا همهٔ داده‌ها را می‌توان به طرز معنی‌داری با یک خط تقریب زد و رفتار آن‌ها را پیش‌بینی نمود؟ توجه کنید که خط رگرسیون که بدون توجه به ابر پراکنش حساب شده باشد اعتبار چندانی ندارد چون ممکن است داده‌های پرت

محل خط رگرسیون را به مقدار قابل ملاحظه‌ای جابه‌جا کنند. دو مثال خوب برای رگرسیون ارتباط مدل ماشین و میزان کارکرد آن، ارتباط تعداد صفحات کتاب و قیمت آن هستند که در مورد اینکه ارتباط معنی‌داری وجود دارد یا خیر بسیار آموزنده است.

مباحث ریاضی مربوط: متغیرهای وابسته، ضریب تغییرات

نمودار مفاهیم:



همبستگی و رگرسیون

در فصل‌های قبل جامعه‌هایی مورد مطالعه قرار گرفتند که در آن‌ها فقط یک متغیر وجود داشت. مثلاً می‌خواستیم قد دانش‌آموزان کلاس را بررسی کنیم و یا درجه حرارت را در مکان مشخصی در روزهای متفاوت مطالعه کنیم. اگر در برخی از پروژه‌ها به بیش از دو متغیر پرداخته شد، هر یک از متغیرها را جداگانه و تک تک مورد بررسی قرار دادیم و کمتر از ارتباط و تأثیر این دو متغیر بر روی یکدیگر صحبت به میان آوردیم. در این فصل جامعه‌هایی مطالعه می‌شوند که در آن‌ها دو متغیر تصادفی کمّی وجود دارد که عموماً پیوسته‌اند. با تکرار آنچه که در فصل‌های قبل دیدیم هر یک از متغیرها به طور جداگانه تحلیل می‌شوند، امّا می‌خواهیم بدانیم که اگر مقدار یکی از متغیرها رابدانیم آیا می‌توانیم به مقدار متغیر دیگر را دست بیابیم؟ به عبارت دیگر آیا می‌توان یکی از متغیرها را به صورت تابعی ریاضی از متغیر دیگر بیان کرد یا خیر؟ شاید برای بسیاری از متغیرها بتوانیم بلافاصله قضاوت کنیم که چنین امری غیرممکن است. مثلاً اگر مطالعه قد و وزن مطرح باشد با معلوم بودن قد فردی نمی‌توان به‌طور دقیق گفت که وزن او چقدر است زیرا افرادی هستند که اندازه وزن آن‌ها متفاوت است ولی همگی یک قد دارند. درباره دو نمره درس آمار و مدلسازی و درس ادبیات فارسی هم همین گفته صحیح است، زیرا در بین دانش‌آموزان می‌توان مثال‌هایی یافت که نمرهٔ آمار آن‌ها برابر باشد ولی نمرات ادبیات فارسی متفاوت داشته باشند. با گسترش این استدلال می‌توان دریافت که شاید هیچ‌گاه نتوان یک رابطهٔ ریاضی بین دو متغیر برقرار کرد. امّا آنچه که در آمار مطرح است این است که

اگر مقدار یکی از متغیرها معلوم باشد آیا می‌توان *میانگین* مقادیر متغیر دوم را به‌دست آورد؟ مثلاً میانگین قد افراد ۶۰ کیلویی چقدر است؟ میانگین نمره درس ادبیات فارسی دانش‌آموزانی که نمره درس آمار و مدلسازی آن‌ها ۱۴ است چقدر است؟

یعنی می‌خواهیم رابطه‌ای بین مقادیر یک متغیر و میانگین مقادیر متغیر دیگر برقرار کنیم، این رابطه را رگرسیون متغیر دوم نسبت به متغیر اول می‌گوییم. همان‌طوری که از این تعریف برمی‌آید رابطهٔ رگرسیونی یک رابطهٔ دو طرفه نیست. ساده‌ترین رابطه‌ای که ممکن است بین دو متغیر برقرار کرد، رابطهٔ خطی است.

برای پیدا کردن رابطهٔ خطی از روش کم‌ترین مربعات استفاده شده است که ذیلاً محاسبات مربوط به ضریب a و مقدار ثابت b را در رابطه $y = ax + b$ بیان می‌کنیم.

در روش کم‌ترین مربعات این طور استدلال می‌کنیم که برای مقدار x_i مقدار y_i در نمونه مشاهده شده است امّا می‌خواهیم y_i را بر داریم و به‌جای آن $ax_i + b$ را از مدل قرار دهیم. پس مربع خطایی که مرتکب شده‌ایم عبارت است از :

$$(ax_i + b - y_i)^2$$

پس مجموع خطاها در تمام نقاط عبارت است از :

$$\sum(ax_i + b - y_i)^2$$

حال می‌خواهیم a و b را چنان بیابیم تا عبارت بالا می‌نیمم شود.

عبارت بالا تابعی از دو متغیر a و b است. برای می‌نیمم کردن آن لازم است مشتقات جزئی این عبارت برحسب a و b برابر صفر شوند :

مشتق نسبت به a $\Rightarrow \sum 2x_i(ax_i + b - y_i) = 0$

مشتق نسبت به b $\Rightarrow \sum 2(ax_i + b - y_i) = 0$

از معادله دوم با بسط $\sum$ خواهیم داشت :

$$a\sum x_i + nb - \sum y_i = 0$$

حال اگر طرفین را بر n تقسیم کرده و جملات را مرتب کنیم خواهیم داشت.

$$\overline{Y} = a\overline{X} + b$$

از این برابری نتیجه می‌شود که خط رگرسیون از نقطه $(\overline{X}, \overline{Y})$ می‌گذرد اگر a معلوم باشد از برابری بالا می‌توانیم b را حساب کنیم. برای محاسبهٔ a از این برابری b را به‌دست آورده در معادله مشتق نسبت به a قرار می‌دهیم :

$$b = \overline{Y} - a\overline{X}$$

$$
\begin{aligned}
۰ &= \sum x_i\left[ax_i + (\overline{Y} - a\overline{X}) - y_i\right] \\
&= a\sum x_i^{۲} + \overline{Y}\sum x_i - a\overline{X}\sum x_i - \sum x_i y_i \\
&= a\left[\sum x_i^{۲} - n\overline{X}^{۲}\right] + n\overline{X}\overline{Y} - \sum x_i y_i
\end{aligned}
$$

پس

$$
\begin{aligned}
a &= \frac{\sum x_i y_i - n\overline{X}\overline{Y}}{\sum x_i^{۲} - n\overline{X}^{۲}} \\
&= \frac{\frac{۱}{n}\sum x_i y_i - \overline{X}\overline{Y}}{\frac{۱}{n}\sum x_i^{۲} - \overline{X}^{۲}}
\end{aligned}
$$

مخرج کسر برابر واریانس x است صورت کسر هم ارز عبارتی است که آن را کوواریانس x و y می‌نامند و با نماد $COV(x,y)$ نشان داده و به صورت زیر تعریف می‌کنند.

$$
COV(X,Y) = \frac{۱}{n}\Sigma(x_i - \overline{X})(y_i - \overline{Y})
$$

بنابراین

$$
a = \frac{COV(x,y)}{\sigma_x^{۲}}
$$

و در نتیجه

$$
b = \overline{Y} - \frac{COV(x,y)}{\sigma_x^{۲}}.\overline{X}
$$

به سادگی با باز کردن Σ و ضرب جملات دیده می‌شود که :

$$
\begin{aligned}
&\frac{۱}{n}\Sigma(x_i - \overline{X})(y_i - \overline{Y}) = \\
&\frac{۱}{n}\Sigma(x_i y_i - x_i\overline{Y} - y_i\overline{X} + \overline{X}\overline{Y}) \\
&= \frac{۱}{n}\sum x_i y_i - \overline{Y}\frac{۱}{n}\sum x_i - \overline{X}\frac{۱}{n}\sum y_i + \overline{X}\overline{Y} \\
&= \frac{۱}{n}\sum x_i y_i - \overline{X}\overline{Y} - \overline{X}\overline{Y} + \overline{X}\overline{Y} \\
&= \frac{۱}{n}\sum x_i y_i - \overline{X}\overline{Y}.
\end{aligned}
$$

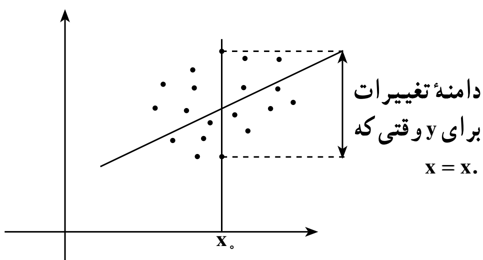
در مباحث آماری معمولاً مقدار هر متغیر را برابر میانگین آن در نظر می‌گیرند. ازاین‌رو با داشتن خط رگرسیون y نسبت به x، برای هر مقدار x می‌توانیم *مقدار تقریبی* y را از روی خط

رگرسیون $y = ax + b$ محاسبه کنیم.

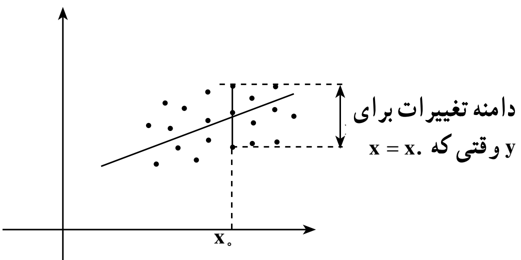
تذکر: برای هر تعداد نقطه در صفحه و هر شکلی که پراکنده شده باشند می‌توانیم خط رگرسیون را به‌دست آوریم ولی ممکن است این خط، مدل مناسبی برای داده نباشد، به عبارت دیگر ممکن است داده‌ها از یک مدل خطی پیروی نکنند و لذا خطی که از این طریق به‌دست می‌آید مفید نیست. اولین راهنما برای مناسب بودن خط به‌دست آمده از روش کمترین مربعات، بررسی نمودار پراکنش است، این نمودار به میزان قابل ملاحظه‌ای می‌تواند ما را در تشخیص مدل خطی هدایت کند. در مباحث پیشرفته آماری روش‌هایی است که می‌توان به کمک آن‌ها مناسب بودن مدل را تشخیص داد. ما در این کتاب فقط استفاده از نمودار پراکنش را مدنظر قرار داده‌ایم.

ضریب همبستگی

قبلاً گفتیم که معمولاً متغیرها از یک رابطه ریاضی پیروی نمی‌کنند و معمولاً داده‌ها روی نمودار یک منحنی مشخص واقع نمی‌شوند. به‌ویژه به ندرت ممکن است مثال‌هایی یافت که نقاط مربوط به داده‌ها روی یک خط باشند. آنچه که اتفاق می‌افتد آن است که نقاط داده‌ها در اطراف خط پراکنده شده‌اند. در متن کتاب سعی شده است که نشان دهیم داده‌ها چگونه ممکن است بر اثر عوامل تصادفی از مدل خطی منحرف شده باشند. حال هر چقدر نقاط در اطراف خط پراکنده‌تر باشند تبعیت آن‌ها از مدل خطی ضعیف‌تر می‌شود و بالعکس هرچقدر نقاط به خط نزدیکتر باشد این مدل قوی‌تر خواهد بود. اگر نقاط به خط نزدیکتر باشند پیروی آن‌ها از یکدیگر بیشتراست، به‌عبارت دیگر با معلوم شدن یکی، دیگری نمی‌تواند مقادیر پراکندهٔ دلخواه اختیار کند بلکه در یک فاصله کوتاهی نوسان خواهد کرد. ولی اگر نقاط در اطراف خط پراکندگی زیادی داشته باشند، معلوم بودن یک متغیر محدودیت زیادی در مقادیر متغیر دوم ایجاد نمی‌کند. شکل‌های زیر این نکته را نشان می‌دهند.



(ب)



(الف)

در شکل الف دامنهٔ تغییرات y با معلوم بودن x کم ولی در شکل ب دامنه تغییرات زیاد است.

حال می‌خواهیم میزان پیروی و تابعیت (خطی) متغیرها از یکدیگر را به صورت کمی اندازه‌گیری کنیم

یعنی می‌خواهیم تابعی از داده‌ها *بسازیم* که این میزان همبستگی و تابعیت متغیرها از یکدیگر را اندازه بگیرد.

برای این منظور اولاً دقت می‌کنیم که اگر داده‌ها را به صورتی که حاصل شده‌اند مورد استفاده قرار دهیم یعنی تابعی به صورت

$$f(x_1, y_1, \ldots, x_n, y_n)$$

در نظر بگیریم، آن‌گاه با تعیین مبدأ اندازه‌گیری داده‌ها تغییر کرده و در نتیجه مقدار تابع فوق تغییر خواهد کرد، و این بدان معنی است که همبستگی بین دو متغیر که اساساً باید یک صفت ذاتی جامعه باشد با تغییر مبدأ تغییر می‌کند که درست نیست. پس باید داده‌ها به قسمی مورد استفاده قرار گیرند که تأثیر مبدأ در آن‌ها از بین برود برای این منظور از داده‌ها به صورت زیر استفاده می‌کنیم.

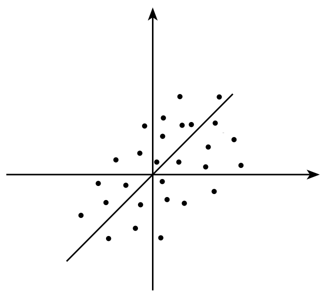
$$f(x_1 - \bar{x}, y_1 - \bar{y}, \ldots, x_n - \bar{x}, y_n - \bar{y})$$

حال اگر مبدأ تغییر کند مقدار عبارت فوق تغییر نخواهد کرد.

نکته دیگر آنکه اگر واحد اندازه‌گیری را عوض کنیم مقدار داده‌ها نیز تغییر خواهد کرد و لذا مجدداً مقدار تابع فوق تغییر خواهد کرد و این با ذاتی بودن همبستگی دو متغیر تناقض دارد بنابراین لازم است داده‌ها به صورتی مورد استفاده قرار گیرند که این مشکل پیش نیاید. برای رفع این مشکل داده‌ها را به صورت زیر مورد استفاده قرار می‌دهیم

$$f\left(\frac{x_1 - \bar{X}}{\sigma_x}, \frac{y_1 - \bar{Y}}{\sigma_y}, \ldots, \frac{x_n - \bar{X}}{\sigma_x}, \frac{y_n - \bar{Y}}{\sigma_y}\right)$$

حال فرض کنید داده‌ها از مدل خطی به شکل زیر پیروی کرده باشند.



در این حالت هر چقدر داده‌ها به خط نزدیکتر باشند، نقاط کم‌تری در ربع دوم و چهارم داریم و در عوض نقاط بیش‌تری در ربع اول و سوم خواهیم داشت از ویژگی‌های این دو ربع اخیر آن است که مختصات نقاط آن‌ها هم‌علامت است بنابراین حاصل‌ضرب مختصات آن‌ها مثبت است پس در این حالت ما در عبارت‌های زیر تعداد زیادی جمله مثبت و تعداد کمی جمله منفی داریم.

$$\frac{x_1 - \bar{X}}{\sigma_x} \cdot \frac{y_1 - \bar{Y}}{\sigma_y}, \ldots, \frac{x_n - \bar{X}}{\sigma_x} \cdot \frac{y_n - \bar{Y}}{\sigma_y}$$

پس مجموع آن‌ها باید بزرگ (و مثبت) باشد پس اگر مدل به صورت خط مذکور در بالا باشد، هر چقدر مدل به خط نزدیکتر باشد تعداد جمله‌های مثبت بیشتر و در نتیجه مجموع بالا بزرگتر خواهد بود. بنابراین یک معیار برای اندازه‌گیری همبستگی عبارت زیر است.

$$\Sigma\left(\frac{x_i-\overline{X}}{\sigma_x}\right)\left(\frac{y_i-\overline{Y}}{\sigma_y}\right)$$

امّا عبارت بالا یک مشکل دیگر دارد و آن اینکه اگر تعداد داده‌ها را اضافه کنیم به جملات مثبت افزوده می‌شود و در نتیجه عبارت بالا بزرگتر خواهد شد و این درست نیست که بگوییم همبستگی دو متغیر با افزایش اندازه نمونه بزرگتر شده است، پس برای اینکه تأثیر اندازه نمونه را از بین ببریم معیاری به صورت زیر تعریف می‌کنیم.

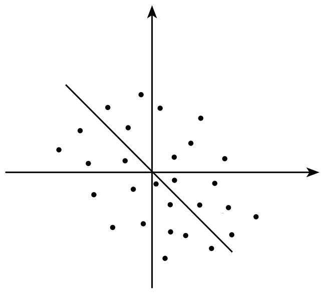
$$\frac{۱}{n}\Sigma\left(\frac{x_i-\overline{X}}{\sigma_x}\right)\left(\frac{y_i-\overline{Y}}{\sigma_y}\right)$$

این معیار را ضریب همبستگی x و y می‌نامند و با علامت $r_{x,y}$ نشان می‌دهند.

$$r_{x,y}=\frac{۱}{n}\Sigma\left(\frac{x_i-\overline{X}}{\sigma_x}\right)\left(\frac{y_i-\overline{Y}}{\sigma_y}\right)$$

$$=\frac{\frac{۱}{n}\Sigma(x_i-\overline{X})(y_i-\overline{Y})}{\sigma_x.\sigma_y}=\frac{COV(x,y)}{\sigma_x.\sigma_y}.$$

البته اگر مدل به صورت زیر باشد.



آن گاه تعداد جملات منفی در مجموع بیشتر خواهد شد و لذا هر چقدر مجموع مذکور کوچکتر (و منفی) باشد همبستگی بیشتر است.

از ساختن دستور فوق نتایج زیر نتیجه می‌شود.

۱ــ اگر $r>۰$ آن‌گاه x و y در یک جهت حرکت می‌کنند و بالعکس، یعنی اگر x زیاد شود، y نیز عموماً زیاد می‌شود.

۲ــ اگر $r < ۰$ ، آن‌گاه y و x در خلاف جهت حرکت می‌کنند و بالعکس

۳ــ اگر $r = ۰$ ، x و y از یکدیگر تبعیت نمی‌کنند و تغییرات آن‌ها مستقل از هم انجام می‌شود.

۴ــ اگر تمام نقاط روی خط باشند، آنگاه $r = \pm ۱$ (البته می‌شود ثابت کرد که اگر $r = \pm ۱$ آن‌گاه نقاط روی یک خط قرار دارند).

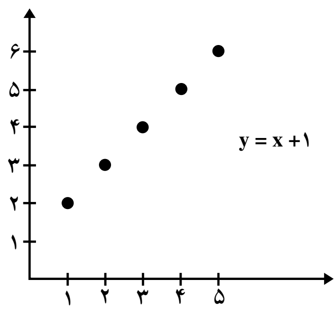
توضیح فعالیت‌ها

فعالیت صفحه ۱۶۸

هدف این است که وابستگی متغیرهای قد و وزن دانش‌آموزان بررسی شود. در طول درس نمودار پراکنش را برای داده‌های قد و وزن رسم نموده و تحلیل نمایید.

فعالیت صفحه ۱۷۱

نمودار پراکنش به شکل زیر است.



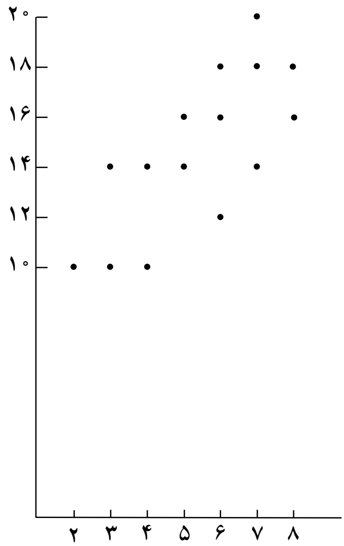
بعد از تغییر مؤلفه دوم بعضی مؤلفه‌های y بیشتر و بعضی کمتر می‌شود اما از آنجا که این تغییر تصادفی است هنوز داده‌ها اطراف خط $y = x + ۱$ پراکنده خواهند ماند. وابستگی y,x کمتر خواهد شد.

پروژه صفحه ۱۸۶

این پروژه توسط یک دانش‌آموز ۱۷ ساله انجام شده است. آن را در کلاس به نقد بگذارید.

پاسخ تمرینات صفحه ۱۷۲

۱ــ

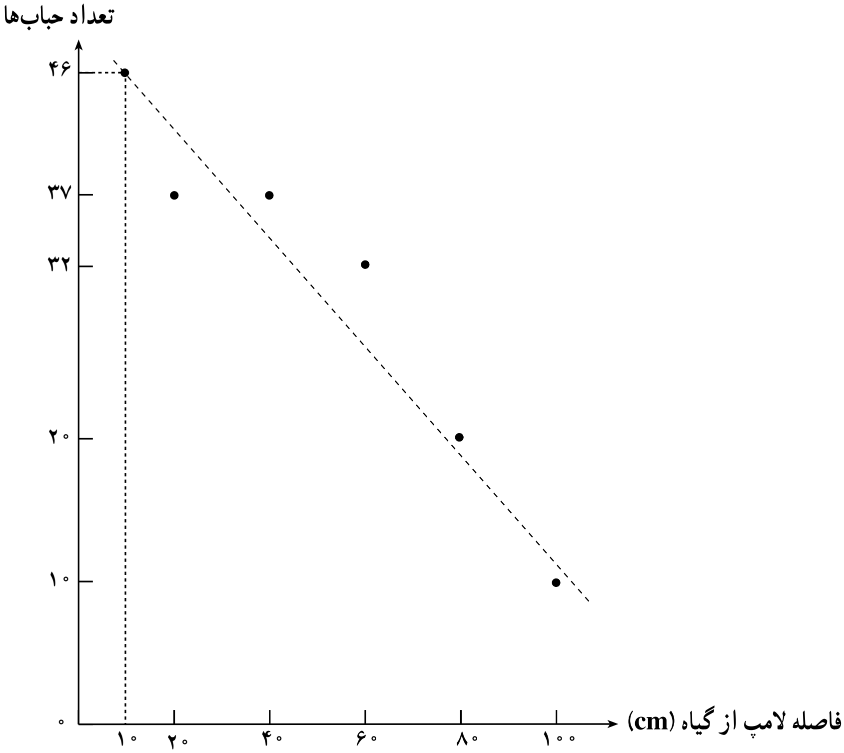


این نمودار نشان می‌دهد که می‌توان ارتباطی مستقیم بین تعداد ساعات مطالعه و نمرهٔ آزمون مشاهده کرد.

۲ــ ارتباط معکوس وجود دارد. بدین شکل که هرچه عمر اتومبیل افزایش یابد، قیمت آن عموماً کاهش پیدا می‌کند.

پاسخ تمرینات صفحات ۱۸۳ــ۱۸۱

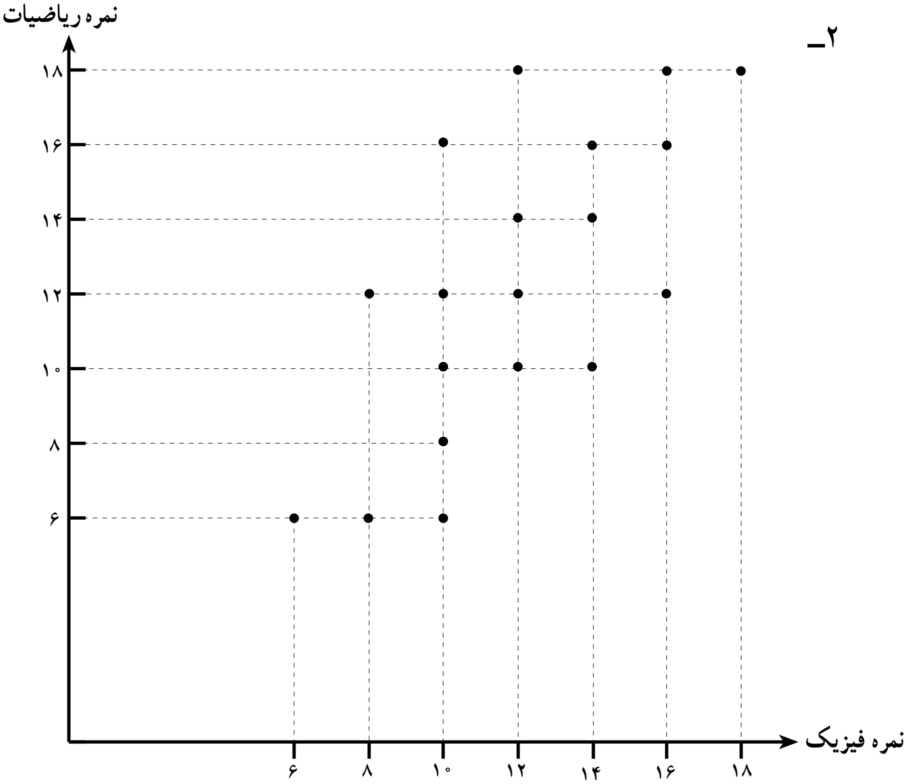
۱ــ الف : نمودار ابر پراکنش تعداد حباب‌ها و فاصله لامپ از گیاه



ب : تقریباً ۴۶ حباب

ج : همان‌طور که فاصله لامپ کم می‌شود، سرعت فتوسنتز زیاد می‌شود.

۲ـ



$$\overline{X}_{\text{فیزیک}}=\frac{۱۶+۱۰+۸+۱۸+۱۴+۱۰+۱۰+۶+۱۰+۱۴+۱۰+۱۶+۱۲+۱۶+۱۲+۱۲+۸+۱۲+۱۴+۱۶}{۲۰}=\frac{۲۴۴}{۲۰}=۱۲/۲$$

$$\overline{X}_{\text{ریاضی}}=\frac{۱۸+۸+۱۲+۱۶+۱۸+۶+۱۰+۶+۱۶+۱۰+۱۲+۱۲+۱۰+۱۶+۱۸+۱۲+۶+۱۴+۱۴+۱۲}{۲۰}=\frac{۲۴۶}{۲۰}=۱۲/۳$$

$$\sigma_{\text{فیزیک}}=\sqrt{\frac{(۱۶-۱۲/۲)^{۲}۴+۵(۱۰-۱۲/۲)^{۲}+۲(۸-۱۲/۲)^{۲}+(۱۸-۱۲/۲)^{۲}+۳(۱۴-۱۲/۲)^{۲}+(۶-۱۲/۲)^{۲}+۴(۱۲-۱۲/۲)^{۲}}{۲۰}}=\sqrt{۹/۹۶}\approx ۳/۱۶$$

$$\sigma_{\text{ریاضی}}=\sqrt{\frac{۳(۱۸-۱۲/۳)^{۲}+(۸-۱۲/۳)^{۲}+۵(۱۲-۱۲/۳)^{۲}+۳(۱۶-۱۲/۳)^{۲}+۳(۶-۱۲/۳)^{۲}+۳(۱۰-۱۲/۳)^{۲}+۲(۱۴-۱۲/۳)^{۲}}{۲۰}}=\sqrt{۱۰/۸۳}\approx ۳/۳$$