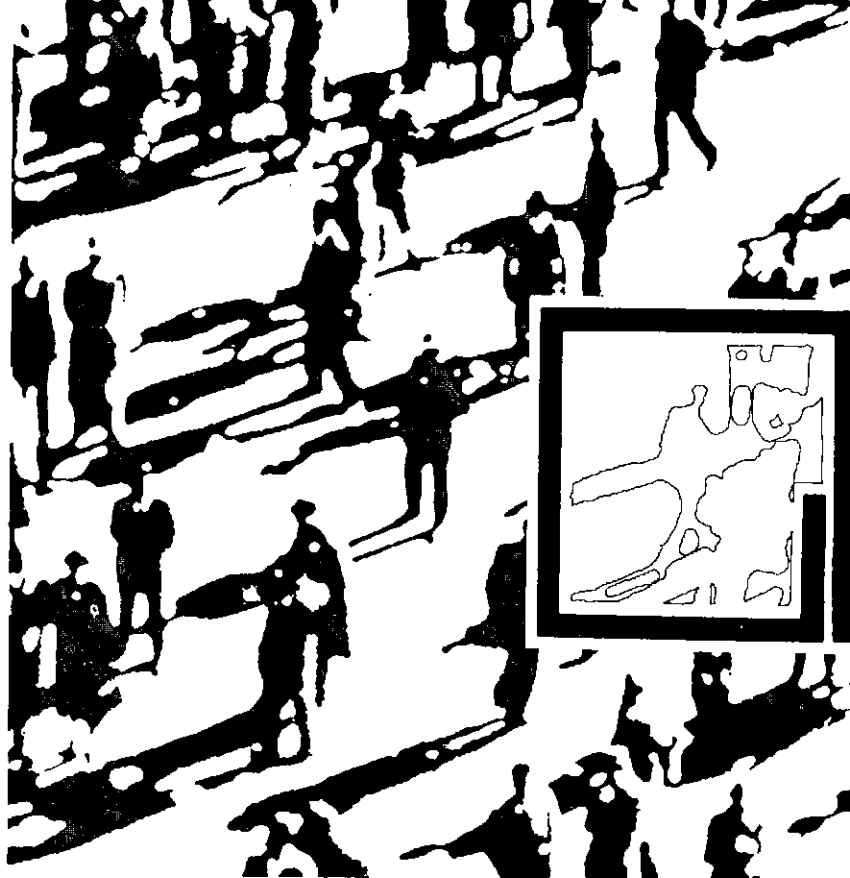


n یا n-۱

● دکتر عین‌الله پاشا
دانشگاه تربیت‌معلم



براساس انتظار یا اطمینان بود و حکم‌ها قطعی نبودند. بنابراین امکان هر چیزی هست، یعنی نمرات ممکن است به میانگین خیلی نزدیک باشند و یا برعکس، ممکن است خیلی دور باشند. در مبحث پراکندگی می‌خواهیم معیاری برای اندازه‌گیری پراکندگی و دور از هم بودن مقادیر به‌دست آوریم. ساده‌ترین و مقدماتی‌ترین معیاری که در این زمینه مطرح شده است دامنه تغییرات است.

دامنه تغییرات: فرض کنیم $x_1, \dots, x_n$ مقادیر مورد نظر باشند و همچنین فرض کنید $a = \text{Min}\{x_1, \dots, x_n\}$ و $b = \text{Max}\{x_1, \dots, x_n\}$ در این صورت $R = b - a$ را دامنه تغییرات می‌نامند و همان‌طوری که از ساختمان آن برمی‌آید، بزرگی آن نشان از اختلاف زیاد مقادیر و کوچکی آن نشان از اختلاف کم مقادیر دارد، درحالت خاص اگر $R = 0$ ، آنگاه تمام مقادیر با هم برابرند و پراکنده نیستند. اشکالی که دامنه تغییرات دارد این است که اصطلاحاً می‌گویند: نسبت به تمام مقادیر حساس نیست. یعنی در مقابل حتی تغییرات بزرگ در مقادیر، ممکن است دامنه تغییرات، تغییر نکند. به دو مثال زیر که دارای دامنه تغییرات برابر هستند توجه کنید:

۱، ۲، ۳، ۴، ۵، ۶، ۷، ۸، ۹، ۱۰

۱، ۴، ۵، ۵، ۵، ۶، ۶، ۶، ۱۰

مبحث واریانس، ابهامایی را در ذهنها به‌وجود آورده است. اغلب سؤال می‌شود که از کدام دستور زیر باید برای محاسبه واریانس استفاده شود:

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

ما در این مقاله سعی می‌کنیم به این سؤال پاسخ دهیم. این مقاله شامل دو بخش است، در بخش اول پارامترهای پراکندگی به‌ویژه واریانس بررسی می‌شوند و بخش دوم به موضوع برآورد می‌پردازد.

بخش اول: پراکندگی

فرض بر این است که پارامترهای مرکزی (میانگین) مد و نما به‌ویژه میانگین برای خواننده آشناست، با این همه، یادآوری می‌کنیم که این پارامترها محل تجمع داده‌ها را مشخص می‌کنند، وقتی که گفته می‌شود معدل دانش‌آموزی برابر $14/75$ است این انتظار به‌وجود می‌آید که نمرات این دانش‌آموز حول و حوش $14/75$ است. به‌ندرت ممکن است برای این دانش‌آموز نمره ۱۰ یا ۲۰ تصور کرد. هر نمره‌ای که به میانگین نزدیکتر باشد با اطمینان بیشتری می‌تواند جزء نمرات کسب شده به‌وسیله این دانش‌آموز باشد. همان‌طوری که ملاحظه کردید، این نتیجه‌گیریها

یعنی واریانس برابر میانگین مجذور انحراف از میانگین است. این دستور نشان داده است که نه تنها نقاط ضعف دامنه تغییرات و انحراف از میانگین را ندارد، بلکه نتایج و مفاهیم عمیقی را می‌توان براساس آن پی‌ریزی کرد و از این‌رو به‌عنوان مناسبترین پارامتر برای اندازه‌گیری پراکندگی شناخته شد.

پس تا اینجا معلوم شد که واریانس برابر میانگین مجذور انحرافات از میانگین است و دستور آن چنین است:

$$(\text{میانگین})^2 = \frac{1}{n} \sum (x_i - \text{میانگین})^2 = \text{واریانس}$$

مثال:

واریانس مقادیر زیر را به‌دست آورید:

$$1, 8, 6, 3, 2$$

حل: ابتدا میانگین را حساب می‌کنیم

$$\text{میانگین} = \frac{1+8+6+3+2}{5} = \frac{20}{5} = 4$$

بنابراین

$$\text{واریانس} = \frac{(1-4)^2 + (8-4)^2 + (6-4)^2 + (3-4)^2 + (2-4)^2}{5}$$

$$= \frac{9+16+4+1+4}{5}$$

$$= \frac{34}{5} = 6.8$$

بخش دوم: برآورد

معمولاً در یک مسأله آماری می‌خواهیم پارامتری را در یک جامعه آماری براساس یک نمونه، یعنی بخش - تقریباً - کوچکی از جامعه که با شرایط خاصی انتخاب می‌شود، تعیین کنیم، از آنجایی که ما نمی‌توانیم و یا نمی‌خواهیم از اطلاعات تمام جامعه استفاده کنیم، در تعیین این پارامتر با اطلاعات ناقصی که از نمونه به‌دست آورده‌ایم سر و کار داریم، با اطلاعات ناقص هم نمی‌توان دقیقاً پارامتر مورد مطالعه را به‌دست آورد. آنچه که از طریق نمونه حاصل می‌شود یک تقریب از پارامتر مورد نظر است که آن را برآورد پارامتر می‌نامند.

مثلاً می‌خواهیم در یک کلاس ۴۰ نفری بدانیم میانگین نمره درس ریاضی دانش‌آموزان چه قدر است. اگر تمام نمرات را در نظر بگیریم و آنها را با هم جمع و بر ۴۰ تقسیم کنیم، عدد حاصل، دقیقاً همان چیزی است که به‌دنبال آن هستیم ولی حالا فرض کنید که ۱۰ نفر از دانش‌آموزان کلاس را به تصادف

در دسته دوم بجز ۱ و ۱۰ سایر مقادیر به هم خیلی نزدیک هستند و پراکندگی در مقادیر دسته اول محسوس‌تر است و حال آن‌که دامنه تغییرات این ویژگی را نشان نمی‌دهد و از نظر این پارامتر، هر دو دسته مقادیر به یک اندازه پراکنده هستند. برای رفع این ضعف، باید معیار مورد استفاده به گونه‌ای باشد تا مقادیر - مستقیماً - در محاسبه آن دخیل باشند. از این‌رو معیار زیر را که شاید بهتر از دامنه تغییرات باشد معرفی کردند.

انحراف از میانگین: در این معیار، اختلاف مقادیر از میانگین در نظر گرفته می‌شود.

می‌دانیم که همواره $\sum_{i=1}^n (x_i - \text{میانگین}) = 0$ (ثابت کنید)، یعنی مجموع این اختلاف‌ها اطلاعی درباره پراکندگی مقادیر نمی‌دهد. برای رهایی از این مشکل فاصله مقادیر از میانگین، یعنی $|x_i - \text{میانگین}|$ به‌عنوان پارامتری برای پراکندگی مطرح شد. اشکال این دستور آن است که با افزایش تعداد مقادیر افزایش می‌یابد (جمله‌های مثبت جدیدی اضافه می‌شود). از این‌رو برای از بین بردن اثر تعداد مقادیر این عبارت را بر n تقسیم می‌کنند که به این صورت دستور انحراف از میانگین به‌دست می‌آید:

$$\frac{1}{n} \sum_{i=1}^n |x_i - \text{میانگین}|$$

هر جایی که مجموع مقادیر به تعداد آنها تقسیم می‌شود، میانگین آن مقادیر حساب می‌شود. پس در عبارت بالا، میانگین مقادیر $|x_1 - \text{میانگین}|, |x_2 - \text{میانگین}|, \dots, |x_n - \text{میانگین}|$ حساب شده است. این دستور عیب دامنه تغییرات را ندارد و پارامتر خوبی است، ولی اشکال آن در وجود قدر مطلق است. قدر مطلق با وجود سادگی، تابع سر به راهی نیست و کار کردن با آن مشکلاتی دارد و به همین دلیل قضایا و نتایج پیشرفته و خوبی درباره دستورهای حاصل از آن نمی‌توان استخراج کرد. برای رفع این مشکل چاره زیر اندیشیده شد.

واریانس: قدر مطلق در دستور انحراف از میانگین صرفاً برای این در کار آمد تا جمله‌های دخیل در $\sum$ را نامنفی کند تا از برابر صفر شدن مجموع آنها جلوگیری کند. می‌دانیم که تنها قدر مطلق نیست که می‌تواند این کار را انجام بدهد، با توان ۲ رساندن نیز می‌توان به این مقصود رسید. بنابراین دستور معیار اندازه‌گیری پراکندگی به‌صورت زیر اصلاح شد و آن را واریانس نامیدند.

$$\frac{1}{n} \sum (x_i - \text{میانگین})^2$$

اختلاف خواهیم داشت، همچنین ملاحظه می‌شود که این میانگین مثبت است. یعنی برآورد ما در مجموع بیش از میانگین (پارامتر مورد نظر) بوده است.

ب. این بار برای اینکه جلوی برآورد اضافی را بگیریم، ممکن است برآورد را به صورت زیر انجام دهیم:

$$\text{برآورد میانگین کلاس} = \text{Min}\{x_1, x_2\}$$

جدول زیر را داریم:

نمونه	برآورد	خطای برآورد
۱۴,۱۴	۱۴	-۲
۱۴,۱۶	۱۴	-۲
۱۴,۱۸	۱۴	-۲
۱۶,۱۴	۱۴	-۲
۱۶,۱۶	۱۶	۰
۱۶,۱۸	۱۶	۰
۱۸,۱۴	۱۴	-۲
۱۸,۱۶	۱۶	۰
۱۸,۱۸	۱۸	۲

میانگین خطا عبارت است از:

$$-\frac{8}{9} = -0.88$$

که در این حالت نیز تفاوت به اندازه همان مقدار قبلی است. منتها منفی، یعنی برآوردهای ما در این روش در مجموع، پارامتر را کمتر از آنچه که هست برآورد می‌کند.

ج. ممکن است برآورد میانگین کلاس را به صورت زیر محاسبه کنیم:

$$\text{میانگین مقادیر حاصل از نمونه} = \frac{x_1 + x_2}{2} = \text{برآورد میانگین کلاس}$$

در این مورد نیز می‌توانیم جدول زیر را تشکیل و میانگین خطای برآورد را حساب کنیم:

نمونه	برآورد	خطای برآورد
۱۴,۱۴	۱۴	-۲
۱۴,۱۶	۱۵	-۱
۱۴,۱۸	۱۶	۰
۱۶,۱۴	۱۵	-۱
۱۶,۱۶	۱۶	۰
۱۶,۱۸	۱۷	۱
۱۸,۱۴	۱۶	۰
۱۸,۱۶	۱۷	۱
۱۸,۱۸	۱۸	۲

در این حالت، میانگین خطای برآورد برابر صفر است. یعنی

انتخاب کرده‌ایم و براساس نمره این ۱۰ دانش‌آموز می‌خواهیم بدانیم میانگین نمره کل کلاس در درس ریاضی (حدوداً) چه قدر است؟ چون اطلاعات ما از کلاس کامل نیست آنچه که براساس این ۱۰ نمره به دست خواهد آمد مقداری تقریبی است. تمام سعی و کوشش در آمار و احتمال آن است که محاسبه روی این ۱۰ نمره به قسمی انجام شود که تقریب خوبی از میانگین نمرات کلاس به دست آید. آنچه از طریق نمونه برای پارامتر جامعه به دست می‌آید، برآورد آن پارامتر می‌نامند.

مثال: برای سهولت و درک بهتر، فرض می‌کنیم جامعه ما شامل ۳ دانش‌آموز با نمرات ۱۴, ۱۸, ۱۶ باشد. می‌خواهیم براساس یک نمونه ۲ تایی، میانگین نمره این ۳ دانش‌آموز را برآورد کنیم.

حل: به خاطر کوچکی جامعه می‌توانیم میانگین را حساب کنیم:

$$\text{میانگین کلاس} = \frac{۱۶+۱۸+۱۴}{۳} = \frac{۴۸}{۳} = ۱۶$$

ممکن است ما راههای مختلفی را برای به دست آوردن برآورد در نظر بگیریم، برخی از آنها عبارتند از:

الف. اگر x_1 و x_2 مقادیر حاصل از نمونه باشند، میانگین کلاس را به صورت ماکزیم این دو نمره برآورد می‌کنیم:

$$\text{برآورد میانگین کلاس} = \text{Max}\{x_1, x_2\}$$

حال اگر به دفعات مختلف از این کلاس یک نمونه دوتایی (با جایگذاری) انتخاب کنیم، برآوردهای زیر حاصل می‌شوند:

نمونه	برآورد	خطای برآورد
۱۴,۱۴	۱۴	-۲
۱۴,۱۶	۱۶	۰
۱۴,۱۸	۱۸	۲
۱۶,۱۴	۱۶	۰
۱۶,۱۶	۱۶	۰
۱۶,۱۸	۱۸	۲
۱۸,۱۴	۱۸	۲
۱۸,۱۶	۱۸	۲
۱۸,۱۸	۱۸	۲

ملاحظه کنید که در این روش در اکثر مواقع ما به اندازه دو واحد از میانگین جامعه فاصله داریم، میانگین خطای برآورد در اینجا عبارت است از:

$$\frac{-۲+۰+۲+۰+۰+۲+۲+۲+۲}{۹} = \frac{۸}{۹} = 0.88$$

پس در این روش، به طور متوسط، به اندازه ۰/۸۸ با میانگین

$$\text{میانگین خطای برآورد} = \frac{8+2-16+2+8+2-16+2+8}{9} = 0$$

بنابراین همان طوری که در این مثال دیده شد، در حالت کلی هم با استفاده از مطالب نظری احتمال ثابت می شود، برآورد واریانس جامعه با استفاده از دستور

$$\frac{1}{n-1} \sum (x_i - \text{میانگین})^2$$

نالاریب است. بنابراین دستور مناسبی برای برآورد واریانس (نه محاسبه واریانس) است.

در خاتمه: بنابر آنچه که گذشت، اگر بخواهیم واریانس را حساب کنیم از n استفاده می کنیم و اگر بخواهیم واریانس را به طور نالاریب براساس یک نمونه n تایی برآورد کنیم، از $n-1$ استفاده می کنیم. و اما آخرین نکته این که: اگر خواننده نکته سنج دقت کرده باشد، در سراسر این مقاله از به کار بردن اصطلاح داده که همان کمیت های حاصل از نمونه است، خودداری شده است و همچنین هیچگاه از نمادی برای میانگین و یا واریانس استفاده نشده است. اگر از نمادهای مناسب استفاده می شد یا اصلاً ابهام n یا $n-1$ در کار نمی آمد و یا این ابهام هم به ابهام قبلی اضافه می شد که کی از نماد μ و کی از نماد $\bar{x}$ برای میانگین استفاده کنیم.

جدول زیر، ضمن معرفی نمادها دستور محاسبه و ویژگی آنها را دربر دارد.

دستور محاسبه برآورد	ویژگی برآورد	برآورد پارامتر	نماد	پارامتر
$\bar{x} = \frac{1}{n} \sum x_i$	نالاریب	$\bar{x}$	μ	میانگین جامعه
$S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$	نالاریب	S^2	σ^2	واریانس جامعه

برای توضیح بیشتر، اضافه می شود که μ نمادی است برای میانگین جامعه و برای محاسبه آن باید تمام مقادیر جامعه را با هم جمع و بر تعداد افراد جامعه تقسیم کنیم. σ^2 نمادی است برای واریانس در جامعه که برای محاسبه آن باید میانگین انحرافات از میانگین (یعنی μ) را حساب کرد و لذا دستور آن به صورت $\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$ است. $\bar{x}$ نمادی است برای میانگین مقادیر حاصل از نمونه (یعنی $\bar{x}$ میانگین داده ها است) و S^2 نمادی است برای برآورد نالاریب σ^2 با دستور $\frac{1}{n-1} \sum (x_i - \bar{x})^2$. توجه کنید که $\frac{1}{n} \sum (x_i - \bar{x})^2$ برآورد نالاریب σ^2 نیست، بلکه واریانس داده ها را حساب می کند.

این روش برآورد به گونه ای است که میانگین خطای این روش نه مثبت است و نه منفی. برآوردی که این ویژگی را داشته باشد، از اهمیت بالایی در مباحث آماری برخوردار است و اغلب کارشناسان آمار سعی می کنند برای پارامترها روشی برای محاسبه برآورد بیابند که دارای این خاصیت باشد. به علت اهمیت این ویژگی به آن عنوان مستقلی داده اند که در زیر آن را تعریف می کنیم.

برآورد نالاریب: برآورد یک پارامتر را نالاریب می گوئیم؛ هرگاه میانگین خطای برآورد، برابر ۰ باشد.
برآوردهای (الف) و (ب) نالاریب نیستند، ولی برآورد (ج) نالاریب است.

حال به مسأله اصلی خود یعنی واریانس بازمی گردیم. فرض می کنیم واریانس جامعه معلوم نباشد و در نتیجه بخواهیم براساس یک نمونه آن را برآورد کنیم و با مقدمه ای که گذشت به دنبال برآورد نالاریب برای واریانس هستیم. یعنی می خواهیم براساس مقادیر $x_1, \dots, x_n$ واریانس جامعه را به گونه ای برآورد کنیم که میانگین خطای برآورد، برابر ۰ باشد. در مباحث نیمه پیشرفته آمار، می توان ثابت کرد که یک برآورد نالاریب برای واریانس جامعه از دستور زیر حساب می شود.

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \text{میانگین})^2$$

مثال: در مثال قبل، سعی می کنیم میانگین خطای برآورد را که از دستور بالا حساب می شود، به دست آوریم و مطابق انتظار نشان دهیم که برابر صفر است.

حل: قبلاً واریانس جامعه را حساب می کنیم:

$$\text{واریانس} = \frac{(14-16)^2 + (16-16)^2 + (18-16)^2}{3} = \frac{4+4}{3} = \frac{8}{3} = 2.66$$

نمونه	میانگین	برآورد واریانس	خطای واریانس
۱۴، ۱۴	۱۴	۰	$\frac{8}{3}$
۱۴، ۱۶	۱۵	۲	$\frac{2}{3}$
۱۴، ۱۸	۱۶	۸	$-\frac{16}{3}$
۱۶، ۱۴	۱۵	۲	$\frac{2}{3}$
۱۶، ۱۶	۱۶	۰	$\frac{8}{3}$
۱۶، ۱۸	۱۷	۲	$\frac{2}{3}$
۱۸، ۱۴	۱۶	۸	$-\frac{16}{3}$
۱۸، ۱۶	۱۷	۲	$\frac{2}{3}$
۱۸، ۱۸	۱۸	۰	$\frac{8}{3}$

$$\leftrightarrow \frac{8}{3} - 2 = \frac{8-6}{3} = \frac{2}{3}$$

$$\leftrightarrow \frac{8}{3} - 8 = -\frac{16}{3}$$