نگاهی به معیارهای میانی و پراکندگی

خسرو فضلی*

جواد بهبودیان†

چکیده

در این مقاله بر پایه یک تابع چند متغیری متقارن، تعریفی کلی برای معیار میانی، یا مرکزی، داده‌های $x_1, x_2, \dots, x_n$ ارائه می‌دهیم. همین کار را برای معیار پراکندگی این داده‌ها بر پایه دو تابع متقارن انجام می‌دهیم. برای روشن شدن موضوع، چند مثال ذکر می‌کنیم.

۱ پیشگفتار

داده‌های آماری را معمولاً در عددی به نام معیار میانی، یا مرکزی، مانند میانگین، میانه یا نما، خلاصه می‌کنند. این عدد در میان داده‌ها جا دارد و در حکم نماینده آنهاست. افزون بر این، عددی دیگر به نام معیار پراکندگی، مانند برد یا واریانس، را به کار می‌برند که میزان تغییر میان داده‌ها یا چگونگی پراکندگی آنها را نسبت به یک معیار میانی بیان می‌دارد. معیار میانی و معیار پراکندگی را می‌توان به راه‌های گوناگون با ویژگیهای مشخص تعریف کرد.

در این مقاله می‌خواهیم با یک نگاه بر پایه یک تابع چند متغیری متقارن، تمام معیارهای میانی را بیان داریم. همین کار را در مورد معیارهای پراکندگی بر پایه دو تابع متقارن انجام می‌دهیم. برای روشن شدن موضوع مثالهای گوناگون ارائه می‌دهیم.

۲ معیارهای میانی در یک نگاه

معیار میانی عددی است که داده‌های $x_1, x_2, \dots, x_n$ را بر پایه یک ویژگی مناسب خلاصه نماید. مثلاً $\bar{x}$ (میانگین حسابی) یک معیار میانی است، **زیرا اگر آن را به جای هر x_i بگذاریم مجموع داده‌ها تغییر نمی‌کند. به سخنی دیگر مجموع داده‌ها پایا می‌ماند. از اینرو $\bar{x}$ را یک معیار میانی با «پایایی مجموع» می‌نامیم. این ویژگی در جدول زیر نمایان است.**

داده‌ها	$x_1 x_2 \dots x_n$	$\sum_{i=1}^{n} x_i$
میانگین حسابی	$\bar{x}\ \bar{x} \dots \bar{x}$	$n\bar{x}$

اینک با فرض مثبت بودن x_iها یک معیار میانی با «پایایی حاصلضرب» می‌سازیم.

این معیار را با G نشان می‌دهیم. اگر G را به جای هر x_i بگذاریم باید حاصلضرب داده‌ها تغییر نکند، یعنی

$$\prod_{i=1}^{n} x_i = \prod_{i=1}^{n} G = G^m$$

بنابراین میانگین هندسی به صورت زیر به دست می‌آید:

$$G = \sqrt[n]{x_1 x_2 \dots x_n}.$$

این دو مثال نشان می‌دهند که ویژگی مطلوب را می‌توان به صورت یک تابع n متغیری متقارن از داده‌ها نشان داد. برای مثال اول این تابع به صورت

$$f(x_1, x_2, \dots, x_n) = x_1 + x_2 + \cdots + x_n$$

و برای مثال دوم به صورت

$$f(x_1, x_2, \dots, x_n) = x_1 x_2 \dots x_n$$

است. به طور کلی، یا تابع n متغیری را متقارن می‌گویند، هرگاه برای هر

*) خسرو فضلی گروه ریاضی دانشگاه کردستان †) دکتر جواد بهبودیان بخش آمار، دانشگاه شیراز

جایگشت دلخواه $x_{i1}, x_{i2}, \dots, x_{in}$ از $x_1, x_2, \dots, x_n$ داشته باشیم:

$$f(x_{i1}, x_{i2}, \dots, x_{in}) = f(x_1, x_2, \dots, x_n)$$

مثلاً برای x_1، x_2، x_3 تابع سه متغیری $x_1^2 + x_2^2 + x_3^2$ متقارن ولی تابع سه متغیری $x_1^2 + x_2^2 + x_3$ نامتقارن است. ویژگی مطلوب را با تابع متقارن بیان می‌دارند، زیرا معیار میانی نباید با جایگشت داده‌ها تغییر کند.

از آنچه در بالا گفتیم، می‌توان تعریف کلی زیر را برای معیار میانی ارائه داد.

تعریف معیار میانی. فرض کنید $A \subset R^n$، $f : A \longrightarrow R$، که در آن یک تابع n متغیری متقارن باشد. عدد x^* را یک معیار میانی با «پایایی f» می‌نامند، هرگاه برای هر n تایی $(x_1, x_2, \dots, x_n)$ در A داشته باشیم:

$$f(x_1, x_2, \dots, x_n) = f(x_1^*, x_2^*, \dots, x_n^*)$$

بدیهی است که n تایی $(x_1^*, x_2^*, \dots, x_n^*)$ هم باید در A باشد.

اینک به ذکر دو مثال دیگر می‌پردازیم.

مثال ۱. معیار میانی را برای داده‌های نامنفی با پایایی زیر بیابید:

$$f(x_1, x_2, \dots, x_n) = \sum_{i \neq j} x_i x_j$$

حل تابع داده شده متقارن است. از حل معادله

$$\sum_{i \neq j} x_i x_j = \sum_{i \neq j} x^* x^* = n(m-1) x^{*2}$$

برحسب x^* داریم:

$$x^* = \sqrt{\frac{\sum_{i \neq j} x_i x_j}{n(n-1)}}$$

این معیار را میانگین حاصلضربی می‌نامیم.

مثال ۲. داده‌ها را به صورت $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ مرتب می‌کنیم. معیار میانی را با پایایی زیر، در صورتی که n فرد باشد، بیابید:

$$f(x_1, x_2, \dots, x_n) = x_{(\frac{n+1}{2})}$$

حل . تابع داده شده متقارن است. از حل معادله

$$f(x_1, x_2, \dots, x_n) = f(x^*, x^*, \dots, x^*)$$

بر حسب x^* داریم:

$$x^* = x_{(\frac{n+1}{2})}$$

این معیار، میانه داده‌هاست. اگر n زوج باشد، تابع را به صورت زیر تعریف می‌کنیم:

$$f(x_1, x_2, \dots, x_n) = x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}$$

و در نتیجه داریم:

$$x^* = \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}]$$

در جدول زیر چند معیار میانی داده شده‌اند.

جدول معیارهای میانی

نام معیار	x^*	داده‌ها	تابع متقارن
میانگین حسابی	$\frac{1}{n}\sum_{i=1}^{n} x_i$	حقیقی	$x_1 + x_2 + \dots + x_n$
میانگین هندسی	$\sqrt{x_1 x_2 \dots x_n}$	مثبت	$x_1 x_2 \dots x_n$
میانگین توافقی	$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	مثبت	$\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{n}$
میانگین ریشه‌ای رتبه r	$\sqrt[r]{\frac{1}{n}\sum_{i=1}^{n} x_i^r}$	مثبت	$x_1^r + x_2^r + \dots + x_n^r$
میانگین نمایی	$\log \overline{e^x}$	حقیقی	$e^{x_1} + e^{x_2} + \dots + e^{x_n}$
میانه	$x_{(\frac{n+1}{2})}$	حقیقی n فرد	$x_{(\frac{n+1}{2})}$
میانگین حاصلضربی	$\sqrt{\frac{\sum_{i \neq j} x_i x_j}{n(n-1)}}$	مثبت	$\sqrt{\sum_{i \neq j} x_i x_j}$

لازم است بیفزاییم که در جدول بالا، پنج معیار نخست حالتهای ویژه‌از میانگین زیر به نام میانگین تعمیم یافته‌اند:

$$g(x_1, x_2, \dots, x_n) = h^{-1}(\frac{1}{n}\sum_{i=1}^{n} h(x_i))$$

در این رابطه تابع h یک تابع پیوسته یکنوای اکید با وارون h^{-1} و تابع g متقارن است. برای پنج معیار بالا تابع h به ترتیب عبارت است از x و $\log x$ و $\frac{1}{x}$ و x^r و e^x. ولی دو معیار آخر جدول را نمی‌توان از میانگین تعمیم یافته به دست آورد. بنابراین تعریف کلی بالا، معیارهای بیشتری را در بر می‌گیرد.

۳ معیارهای پراکندگی در یک نگاه

معیار پراکندگی عددی است که میزان تفاوت میان داده‌های $x_1, x_2, \dots, x_n$، $n > 1$، را نسبت به یک معیار میانی یا نسبت به یکدیگر بیان می‌دارد. برای تعیین چنین معیاری، نخست میزان دوری دو کمیت a و b را به وسیله یک تابع فاصله‌ای $d(a, b)$، که نامنفی و متقارن است، مانند $(a-b)^2$ یا $|a-b|^p$، $p > 0$، می‌سنجیم. اینک دو حالت را بررسی می‌کنیم:

الف ـ پراکندگی نسبت به یک معیار میانی

فرض کنید x^* یک معیار میانی با پایایی f برای داده‌ها باشد. میزان دوری داده‌ها نسبت به x^* بر پایه تابع فاصله‌ای $d(a, b)$ عبارت‌اند از

$$y_1 = d(x_1, x^*),\ y_2 = d(x_2, x^*), \dots, y_n = d(x_n, x^*).$$

حال معیار میانی این دوریها را با پایانی $g(y_1, y_2, \dots, y_n)$ می‌یابیم و آن را v^* می‌نامیم. بنابراین با داشتن معیار میانی x^*، یک معیار پراکندگی نسبت به x^* بر پایه تابع فاصله‌ای d و تابع n متغیری متقارن g به دست می‌آید.

مثال ۳ـ فرض کنید $\bar{x} = x^*$ معیار میانی برای $x_1, x_2, \dots, x_n$ باشد. بر پایه تابع فاصله‌ای $d(a,b) = (a-b)^2$ و تابع متقارن $g(t_1, t_2, \dots, t_n) = \sum_{i=1}^{n} t_i$ یک معیار پراکندگی نسبت به $\bar{x}$ به دست آورید.

حل ـ باید v^* را به گونه‌ای بیابیم تا داشته باشیم:

$$g(d(x_1,\bar{x}), d(x_2,\bar{x}), \dots, d(x_n,\bar{x})) = g(v^*, v^*, \dots, v^*)$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} v^* = nv^*$$

بنابراین واریانس داده‌ها به صورت زیر به دست می‌آید:

$$v^* = \frac{1}{n}\sum (x_i-\bar{x})^2$$

مثال ۴ـ فرض کنید $x^* = x_{(\frac{n+1}{2})} = m$ معیار میانی برای x_1،x_2،...،x_n با فرض فرد بودن n باشد. بر پایه $d(a,b) = |a-b|$ و $g(t_1, \dots, t_n) = \sum_{i=1}^{n} t_i^r$، $r > 0$، یک معیار پراکندگی نسبت به m بیابید.

حل ـ به آسانی به دست می‌آوریم:

$$v^* = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|^r.$$

ب ـ پراکندگی داده‌ها نسبت به یکدیگر

اگر دوری داده‌ها را نسبت به یکدیگر بر پایه تابع فاصله‌ای $d(a,b)$ در نظر بگیریم، $n(n-1)$ عدد به صورت $d(x_i, x_j)$، $i \neq j$، به دست می‌آید. حال بر پایه متقارن $g(t_1, t_2, \dots, t_{n(n-1)})$، معیار میانی این اعداد را می‌یابیم تا v^* یعنی معیار پراکندگی داده‌ها نسبت به یکدیگر به دست آید.

مثال ۵ـ بر پایهٔ $d(a,b) = (a-b)^2$ و $g(t_1, t_2, \dots, t_{n(n-1)}) = \sum_{i=1}^{n(n-1)} t_i$ یک معیار پراکندگی برای $x_1, x_2, \dots, x_n$ نسبت به یکدیگر بیابید.

حل ـ به آسانی داریم

$$v^* = \frac{1}{n(n-1)}\sum\sum_{i\neq j}(x_i - x_j)^2$$

مثال ۶ـ بر پایه $d(a,b) = |b-a|$ و $g(t_1, t_2, \dots, t_{n(n-1)}) = \max t_i$ یک معیار پراکندگی برای x_1،x_2،...،x_n نست به یکدیگر بیابید.

حل ـ v^* را باید به گونه‌ای بگیریم تا داشته باشیم:

$$g(|x_1 - x_2|, |x_1 - x_3|, \dots, |x_{n-1} - x_n|) = g(v^*, v^*, \dots, v^*)$$

اگر $x_{(1)}$ و $x_{(n)}$ به ترتیب کوچکترین و بزرگترین مقادیر داده باشند، از حل معادله بالا بر حسب v^*، برد داده‌ها یعنی

$$v^* = x_{(n)} - x_{(1)}$$

به دست می‌آید.

مراجع

[۱] جواد بهبودیان، آمار و احتمال مقدماتی، چاپ نهم، ۱۳۷۵، آستان قدس رضوی.

[2] Berger, R. L. and Casella, G. (1992), "Deriving Generalized Means as least Squares and Maximum Likelihood Estimates", The American Statistician, Vol. 46, No. 4, 279-282.