آمار توصیفی

آمار توصیفی به جمع آوری، تنظیم و خلاصه کردن داده ها گفته میشود. مسلما درک اطلاعات خام قبل از هرگونه آنالیز (سازماندهی و خلاصه کردن) بسیار مشکل و وقت گیر است. با خلاصه کردن داده ها خواننده می تواند با استفاده از جداول، نمودارها، شاخص های هر تحقیق به تصور ذهنی روشنی راجع به اطلاعات آن تحقیق دست پیدا کند. حال در اینجا به ذکر هر یک از آنها می پردازیم.

جدول

یکی از روشهای خلاصه کردن اطلاعات استفاده از جداول می باشد. با استفاده از جدول امکان گزارش نتایج بیشتر، در حجم کمتر فراهم می گردد.

❖ ***انواع جداول***

1. یک بعدی: که در آن تنها اطلاعات توصیفی یک متغیر بیان می گردد. به عنوان مثال در جدول شماره زیر تنها فراوانی افرادی که سیگار استعمال میکنند و آنهایی که مصرف نمی کنند معین می گردد.

سیگار	میکشد (+)	نمیکشد (−)
تعداد	...	...

2. دو بعدی: که در آن اطلاعات مربوط به دو متغیر بیان می گردد و از روی این اطلاعات می توان به وجود یا عدم وجود ارتباط بین دو متغیر پی برد در جدول زیراطلاعات مربوط به استعمال سیگار در افراد مورد پژوهش به تفکیک گروههای جنسی بیان گردیده است.

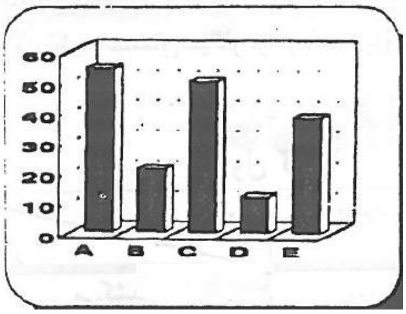
سیگار / جنس	مرد	زن
میکشد	....	....
نمیکشد	....	....

3. سه بعدی: که به بیان ارتباط بین ۳ متغیر می پردازد. در جدول زیر، وضعیت استعمال سیگار در افراد مورد پژوهش به تفکیک گروههای جنسی و سنی بیان گردیده است.

جنس	مرد		زن	
سیگار / سن	<50	>50	<50	>50
میکشد (+)	...	...	...	...
نمیکشد (−)	...	...	...	...

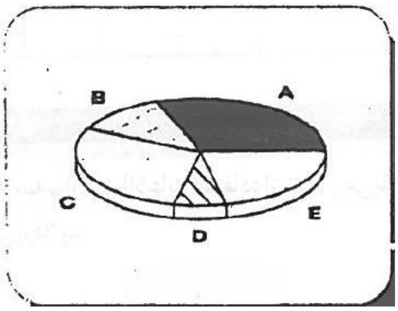
نمودار

یکی دیگر از روشهای خلاصه سازی اطلاعات استفاده از نمودار می باشد که ذیلا به ذکر انواع نمودارها و مورد استفاده هر یک از آنها می پردازیم.



❖ ***انواع نمودار***

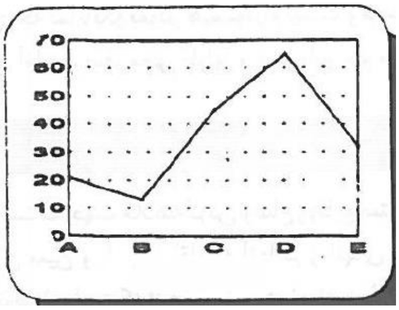
1. نمودار نرده ای، ستونی، میله ای[1]: فراوانی متغیرهای کیفی (اسمی و رتبه ای) را با این نمودار نشان می دهند. در این نمودار: الف) پهنای ستونها با هم برابر است. ب) فاصله ستونها با هم برابر است. ج) تنها ارتفاع ستونها برای ما مهم است.



2. نمودار دایره ای یا کلوچه ای[2]: فراوانی متغیرهای کیفی (اسمی و رتبه ای) را با این نمودار نشان میدهند.

زاویه مرکزی هر قطاع = $\alpha=\frac{F\times 360}{N}$

در این فرمول، F به مفهوم تعداد نمونه های موجود در گروه مورد نظر و N، تعداد کل نمونه های بررسی شده است.

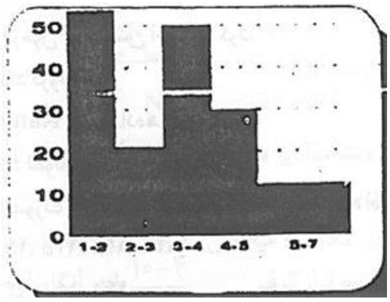


3. نمودار چندگوش[3]: . مورد استفاده آن گزارش داده های کمی گسسته است. بدین ترتیب که مقادیر صفت را روی محور طولها و فراوانی متناظر با آن مقادیر را روی محور عرضها مشخص میکنند. نموداری که از بهم وصل کردن نقاط بدست می آید، نمودار چندگوش نامیده می شود.

معمولا اولین و آخرین نقطه را با شیب خاصی به منحنی طولها وصل میکنند، شیب مورد نظر عموما بوسیله شیب یک یا دو خط آخر تعیین می شود. اگر تمام خطوط با زوایای مختلف صعودی باشند در آن صورت نمی توانیم برای خط آخر

---
1 Bar diagram
2 Pie diagram
3 Polygon diagram

شیبی که آن را به محور طولها وصل کند بیابیم، بنابراین منحنی را با توجه به شیب خطوط قبلی، بطور صعودی، با نقطه چین ادامه می دهیم.



4. نمودار هیستوگرام[1]: مورد استفاده این نمودار درباره صفات کمی پیوسته است که محور عمودی آن فراوانی هر گروه از متغیر کمی می باشد. قاعده ستونها می تواند مساوی انتخاب نشود، در این صورت سطح زیر هر ستون متناسب با فراوانی آن گروه است. چنانچه دو گروه متوالی یا بیشتر، از نظر فراوانی برابر بودند و یا اختلاف فراوانی بین گروههای سنی متوالی، بسیار کم بود، در آن صورت از کشیدن خطوط عمودی بین ستونها صرف نظر میکنیم. اما همواره متوجهیم که مساحت هر مستطیل متناسب با فراوانی آن گروه باشد.

برای راحتی کار در هنگام بررسی هر منحنی نقاط وسط ستونها را به هم وصل میکنند. باید دقت نمود که نمودار بدست آمده به هیچ وجه نمایانگر نمودار هیستوگرام نیست و ما حق نداریم ستونها را پاک کنیم.

نکته: در نمودارهای آماری، صفر محور طولها و عرضها بر هم منطبق نیستند.

[1] Histogram diagram

شاخص ها

شاخص[1] عدد یا نسبتی است که جهت خلاصه کردن اوضاع و وقایع استخراج می شود و به عنوان مشخص کننده آن اوضاع یا وقایع در زمان معین و یا برای مقایسه آنها در زمانهای مختلف بکار می روند. به عبارت دیگر شاخص آماری، شاخص توصیفی است که از داده های یک نمونه محاسبه می شود. در حالیکه، پارامتر شاخص توصیفی است که از داده های یک جمعیت محاسبه می شود. برای توصیف دقیق یک جمعیت از دو نوع شاخص مرکزی و پراکندگی استفاده میگردد که در ذیل به شرح آنها می پردازیم.

شاخص مرکزی به عددی که اکثر داده ها[2] یا نمونه های جامعه حول آن قرار دارند گفته می شود. شاخصهای مرکزی عبارت از میانگین[3] ، میانه[4] و نما[5] می باشد.

شاخص پراکندگی مشخص می کند که نمونه ها چقدر از شاخص مرکزی فاصله دارند (پراکندگی متناسب است با تنوع مقادیر داده ها). اگر مقادیر به هم نزدیک باشند پراکندگی کمتر است و بالعکس. شاخصهای پراکندگی عبارت از دامنه[6] ، میانگین انحرافات[7] ، پراش[8] و انحراف معیار[9] می باشد.

پیش از بحث درباره تک تک این شاخص ها توضیح مختصری درباره داده های گروهبندی شده میدهیم:

برای خلاصه کردن اطلاعات خام بدست آمده از یک تحقیق باید آنها را گروه بندی کرد. برای اینکار باید تعداد دسته ها (K) را محاسبه کرد و بعد دامنه هر گروه را مشخص نمود. بطور کلی هرگونه خلاصه کردن اطلاعات به معنی از دست رفتن بخشی از اطلاعات است، ولی ما به قیمت قابل فهم تر شدن اطلاعات آنها را خلاصه می کنیم. هر چه تعداد دسته ها بیشتر باشد اطلاعات کمتری از دست می رود، ولی در عوض کار مشکلتر می شود. اگرچه تعیین K بستگی به نوع آنالیز و سلیقه محقق دارد، اما بصورت یک قانون کلی جهت محاسبه تعداد دسته ها می توان از فرمول استورگس استفاده کرد:

[1] Indicator
[2] Data
[3] Mean
[4] Median
[5] Mode
[6] Range
[7] Mean deviation/MD
[8] Variance
[9] Standard deviation/SD

K: تعداد گروه ها $K= 3.323 \times \log_{10} n+1$

R = Range کلیه داده ها

n=تعداد نمونه ها $W=\frac{R}{K}$ = دامنه هرگروه

مثال: قد ۱۵ نفر بصورت خام، در جدول زیر دسته بندی شده است:

۱۳۰، ۱۴۵، ۱۴۵، ۱۶۱، ۱۷۲، ۱۵۳، ۱۴۷، ۱۳۴

۱۷۰، ۱۶۵ ۱۶۰، ۱۴۳، ۱۷۰، ۱۵۰، ۱۵۷، ۱۴۹

فراوانی	دامنه گروه (CM)
2	130-139
4	140-149
3	150-159
3	160-169
3	170-179

نکته: حد بالای هر دسته نباید مساوی حد پایین دسته بعدی باشد و باید این دو یک واحد اعشار یا صحیح، بسته به نوع اعداد، با هم اختلاف داشته باشند.

شاخصهای مرکزی

❖ میانگین حسابی[1]

مشهورترین شاخص مرکزی است و اکثریت مردم به آن معدل می گویند. برای محاسبه آن تمام مقادیر را با هم جمع زده بر کل تعداد تقسیم می کنیم.

$$\mu = \frac{\sum_{i=1}^{n} X_i}{N}$$ = میانگین = میانگین حسابی

X_i= مقدار هریک از متغیر های تصادفی

N= تعداد متغیرها

(علامت $\sum_{i=1} n$ یا بدین مفهوم است که همه مقادیر متغیر را از اولین تا آخرین آنها با هم جمع کنیم).

در مورد متغیرهای پیوسته با داده های دسته بندی شده چون مقداری از اطلاعات حذف شده اند، میانگین واقعی بطور دقیق محاسبه نمی شود. برای محاسبه میانگین تقریبی این گونه داده ها فرض میکنیم که تمام مقادیر دامنه گروه در نقطه وسط دامنه گروه قرار گرفته و خلاصه شده اند. آنگاه آن مقدار (نقطه وسط) را در فراوانی آن گروه ضرب میکنیم. پس از ضرب کردن مقادیر هر دسته در فراوانی اش، آنها را با هم جمع می کنیم و بر تعداد کل داده ها تقسیم میکنیم. فرمول:

$$\overline{X} = \frac{\sum_{i=1}^{n} m_i N_i}{\sum_{i=1}^{n} N_i}$$ = میانگین

n= تعداد گروه ها

m_i= نقطه وسطِ دامنه گروه شمارهi

N_i= فراوانی در گروه i

نکته: میانگین یک نمونه را با نماد X و میانگین جامعه را با نماد μ نشان می دهند.

[1]Arithmetic mean

- **خواص میانگین حسابی**

1. برای مجموعه ای از داده ها تنها یک میانگین حسابی وجود دارد.
2. روش محاسبه و درک مفهوم آن ساده است.
3. چون تمام مقادیر در محاسبه میانگین بکار می روند، پس هر یک از مقادیر روی میانگین اثر می گذارند و همین خاصیت میانگین باعث میشود که مقادیر افراطی (خیلی کم یا خیلی زیاد) روی آن تأثیر گذاشته و آن را چنان منحرف کنند که دیگر شاخص خوبی برای مرکزیت نباشد.

⇐ **مثال**: چهار پزشک در دانشگاه علوم پزشکی ایران، برای تعیین مخارج شیوه معینی در تشخیص پزشکی، آن را ارزیابی کرده و گزارش مخارج را بصورت زیر ارائه داده اند: 500،۵۰۰، 550، 2000 تومان

میانگین =5 .۸۸۷ واضحا این عدد به هیچ وجه نمی تواند معرف میزان مخارج باشد. لذا در هنگام استفاده از میانگین، باید به مقادیر افرطی[1]توجه لازم داشت و اثر آنها را حذف نمود.

❖ **میانه[2]**

اگر چه میانگین و میانه هر دو شاخصهای خوبی برای نشان دادن مرکز توزیع میباشند، کاربرد میانگین در تعبیر و تفسیر اطلاعات و انجام آزمونهای آماری قابل اعتمادتر است. در موارد خاص مثلا زمانی که می خواهیم سطح درآمد و با سطح مصرف را در جامعه ای که اختلاف طبقاتی زیاد است (بگونه ای که عده ای درآمد کلان و اکثریت مردم درامد بسیار محدودی دارند) تعیین کنیم، میانه شاخص مناسب تری خواهد بود. چون میانه درآمدی را که نصف مردم کمتر و نیم دیگر بیشتر از آن دارند مشخص می کند. ولی میانگین چنانچه درآمدهای کلان مقادیر بسیار افراطی را به خود اختصاص دهند نمی تواند شاخص مناسبی از وضعیت درآمد جامعه باشد. میانه یک توزیع برابر است با مقداری که نیمی از افراد، از نظر داشتن مقدار متغیر مورد نظر، از آن بزرگتر و نیم دیگر از آن کوچکتر باشند. نفر وسط هر گروه از فرمول $\frac{n+1}{2}$ بدست می آید. (n= تعداد افراد)

مفهوم میانه را با یک مثال روشن میکنیم؛ اگر بخواهیم میانه ۹۹ سرباز را از نظر قد پیدا کنیم آنها را به ترتیب صعودی و یا نزولی به صف مینماییم. آنگاه از اول یا آخر صف، تا نفر پنجاهم ($\frac{99+1}{2}$ =50) شمرده و قد نفر پنجاهم، میانه قد سربازان خواهد بود.

---

[1] outlier
[2] Median

تعداد مقادیر اگر زوج باشد، میانگین دو مقدار وسط را به عنوان میانه انتخاب میکنیم. مثلا میانه ۶ عددِ ۲، ۷، ۱۳، ۲۱، ۳۵ و ۳۹؛ $17 = \frac{21+13}{2}$ می شود.

روش محاسبه میانه را در داده های دسته بندی شده، با یک مثال بیان میکنیم (مقدار بدست آمده تقریبی است):

اولین گروهی که فراوانی تجمعی آن بیش از $\frac{N+1}{2}$ میگردد را به عنوان گروه میانه در نظر میگیریم یعنی تعیین میکنیم نفر $\frac{N+1}{2}$ ام در کدام گروه است.

$\frac{N+1}{2} = \frac{57+1}{2} = 29$

۳۴ > ۲۹> ۲۴

نفر بیست و نهم در گروه سوم قرار دارد، اما مقدار دقیق میانه چقدر است؟ چون نفر ۲۹ام در گروه سوم است مقدار آن بین 39-۳۰ می باشد. از روی جدول می فهمیم که مقدار مربوط به نفر بیست و چهارم، ۳۰ است. به ازای ۱۰ نفر (N_j) در گروه سوم مقدار، ۹ واحد (h=39-30) تغییر کرده، پس به ازاء یک نفر 0.9واحد، تغییر کرده است (بفرض مساوی بودن فاصله مقدار داده ها). اختلاف بین نفرِ ۲۹= $\frac{N+1}{2}$ از نفر ۲۴ (F_{j-1}) مساوی ۵ نفر است. پس اگر ۵ تا0.9 به مقدار ۳۰ =L_i یعنی کمترین مقدار گروه میانه بیفزاییم، مقدار مربوط به نفر ۲۹ را بدست آورده ایم.

فراوانی تجمعی	فراوانی	دامنه گروه
5	5	10-19
24(F_{j-1})	19	20-29
34	10(N_j)	30-39 ⇐ (L_i)
47	13	49-40
57	10	50-59
	جمع=57	

$$\text{میانه} = L_i + \frac{[(N+1)/2]-F_{j-1}}{N_j} h = 30 + \frac{29-24}{10} \times 9 = 34.5 \times$$

L_i = کرانه پایین گروه میانه = کمترین مقدار گروه میانه F_{j-1}= فراوانی تجمعی گروه قبل از گروه میانه

N_j= فراوانی گروه میانه h= اختلاف مقدار گروه میانه

نکته: در محاسبه میانه باید دقت شود تعداد نفرات (ستون فراوانی) با مقدار مربوط به هر نفر یا دسته (ستون دامنه) اشتباه نشود.

- **خواص میانه:**
  1. برای مجموعه ای از اعداد فقط و فقط یک میانه وجود دارد
  2. از مقادیر افراطی تأثیر پذیر نیست.

❖ نما[1]

عبارت است از داده یا داده هایی که بیشترین فراوانی را دارند؛ به عبارت دیگر صفتی که نسبت به صفات دیگر، افراد بیشتری دارای آن باشند را نما گویند. اگر همه مقادیر با هم متفاوت باشند، مجموعه شاخص نما ندارد، زیرا فراوانی هر داده برابر یک است. ممکن است یک مجموعه بیش از یک نما داشته باشد. به عنوان مثال در بررسی دمای بدن افراد یک جمعیت ممکن است دو نمای ۳۶.9 و 38 داشته باشیم که نشان دهنده این است که احتمالا عده ای از افراد جمعیت مورد مطالعه تب دارند.

مورد استفاده نما کمتر از دو شاخص قبلی است. در اپیدمیولوژی از نما مثلا برای مشخص کردن سنی که در آن یک بیماری بیشترین شیوع را دارد استفاده می شود.

برای محاسبه نما در داده های دسته بندی شده، از فرمول زیر استفاده میکنیم:

[1] Mode

$$Mode=\frac{d_1-d_2}{d_1}\times h+L_i$$

L_i= کرانه پایین گروهی که بیشترین فراوانی در آن است.

h= طول دسته ای که بیشترین فراوانی در آن است.

d_1= اختلاف فراوانی ساده گروه مد با گروه قبل از خودش ($d_1=F_i-F_{i-1}$)

d_2= اختلاف فراوانی ساده گروه مد با گروه بعد از خودش ($d_2=F_i-F_{i+1}$)

شاخص های پراکندگی

❖ دامنه[1]

تفاوت بین کوچکترین و بزرگترین مقدار داده ها می باشد ($R= X_{max} -X_{min}$). چون در محاسبه دامنه تنها از بیشترین و کمترین مقدار متغیر استفاده می گردد، سایر حالات بین این دو اندازه در آن مؤثر نیست، نمی تواند به نحو مطلوبی گویای پراکندگی صفت باشد. به عنوان مثال دو گروه عدد (۱۸، ۱۰، ۱۰، ۱۰، ۱۰، ۲) و (۱۸، ۱۶، ۱۰، ۱۰، ۴، ۲) را در نظر بگیرید. در این دو سری عدد میانگین هر سری برابر ۱۰ و دامنه برابر ۱۶ میگردد. این در حالی است که پراکندگی سری دوم اعداد بیش از سری اول می باشد.

❖ میانگین انحراف[2]

متوسط قدرمطلق انحرافات از میانگین یا متوسط انحراف خطی می باشد. یک راه خوب برای تعیین پراکندگی داده ها این است که فاصله تمام مقادیر را از میانگین حسابی بدست آورده و بدون در نظر گرفتن علامتشان (با استفاده از قدرمطلق) آنها را با هم جمع کرده و بر تعداد داده ها تقسیم کنیم. در واقع فاصله تمام مقادیر را تا میانگین بدست می آوریم.

$$\text{میانگین انحراف}=\frac{\sum|\overline{X}-X_i|}{N}$$

⇐ **مثال**: در یک بررسی نتایج مشاهدات بر روی سن کودکان بستری در یک بخش به این ترتیب بوده است:
۷، ۵،۴،۶ و ۳

[1] Range
[2] Mean deviation

برای محاسبه میانگین انحراف ابتدا میانگین اعداد را بدست می آوریم:

$$\text{میانگین} = \frac{3+4+6+5+7}{5} = 5$$

$$\text{میانگین انحراف} = \frac{|3-5|+|4-5|+|6-5|+|5-5|+|7-5|}{5} = 1.2$$

در این مثال مفهوم عدد 1.2این است که این پنج عدد بطور متوسط از میانگین خود به اندازه 1.2اختلاف دارند. میانگین انحراف عملا در محاسبات آماری کاربردی ندارد و تنها جنبه نظری دارد.

❖ پراش[1]

چون در محاسبه میانگین انحرافات از قدرمطلق اختلاف استفاده شده است و انجام عملیات جبری روی قدر مطلق خالی از اشکال نیست بمنظور رفع این نقیصه و همچنین تأثیر بیشتر اعداد دور از میانگین (که اختلافشان با میانگین زیاد است) و تأثیر کمتر اعداد حول میانگین هر یک از عبارات را مجذور می کنیم. به عبارت دیگر بجای میانگین قدر مطلق انحرافات، از میانگین مجذور انحرافات استفاده میگردد که از جنس خود صفت نیست بلکه از جنس مربع صفت است. واریانس را به حرف یونانی δ^2(سیگما دو) نشان میدهند.

$$2\delta = \frac{\sum(\overline{X} - X_i)^2}{N}$$

- **محاسبه واریانس در داده های دسته بندی شده:**

در محاسبه واریانس و انحراف معیار داده های دسته بندی شده فرض بر این است که همه مقادیر هر دسته در نقطه وسط دامنه هر دسته قرار میگیرند. (یادآوری می شود که این فرض در محاسبات میانه و میانگین نیز بکار رفته اند.)

$$2\delta = \frac{N\sum_{Xi} 2 - \sum_{fiXi} 2}{N}$$

F_i= فراوانی ساده در گروه شماره i

$\overline{X}$ = میانگین کل مقادیر

X_i= نقطه وسط گروه شماره i

N= فراوانی کل مقادیر

[1] Variance

❖ انحراف معیار[1]

چون واریانس از جنس مربع صفت است، برای رفع این اشکال از واریانس جذر گرفته و آنرا انحراف معیار می نامند.

SD= $\sqrt{2\delta}$

- **تأثیر تغییرات یکنواخت صفات در شاخص ها:**

چون در پاره ای از محاسبات آماری تغییرات یکنواخت مشاهدات بار عملیات را میکاهد، تأثیر کم کردن یا اضافه کردن یک عدد ثابت به نتیجه مشاهدات و همچنین ضرب یا تقسیم کردن بر یک عدد ثابت را در شاخص ها بررسی میکنیم:

1. هرگاه داده ها را با عدد ثابتی جمع (یا تفریق) کنیم میانگین به همان اندازه زیاد (یا کم) می شود ولی در انحراف معیار تغییری حاصل نمی شود.
2. هرگاه داده ها را در عدد ثابتی ضرب (یا تقسیم) کنیم میانگین و انحراف معیار به همان نسبت بزرگ (یا کوچک) میشود ولی واریانس به نسبت مجذور عدد ثابت بزرگ (یا کوچک) می شود.

⇐ **مثال**: می خواهیم میانگین و انحراف معیار سه داده ۱۰، ۱۱، ۱۲ را حساب کنیم برای راحتی کار از همه أنها عدد ۱۰ را کم میکنیم. داده های جدید 1،0 و ۲ میشوند که میانگین این سه عدد یک است. برای محاسبه میانگین اصلی، ۱۰ واحد به آن اضافه میکنیم، یعنی

$$SD = \sqrt{\frac{5-3}{2}} = 1 \quad و \quad \mu = \frac{2+1+0}{3} + 10 = 11$$

[1] Standard deviation (SD)

❖ ضریب تغییرات[1]

چون انحراف معیار از نوع خود صفت است، در نتیجه برای مقایسه پراکندگی دو صفت با دو واحد متفاوت قابل استفاده نمی باشد و مطالعه انحراف معیار در چنین شرایطی به تنهایی گمراه کننده خواهد بود. مثلا اگر در یک جامعه انسانی مقایسه پراکندگی توزیع افراد از نظر فشارخون (mmHg) و وزن بدن (Kg) مورد نظر باشد و انحراف معیار آنها به ترتیب ۲۰ و ۵۰ باشد نمی توان نتیجه گرفت که پراکندگی وزن در جامعه بیش از فشارخون است زیرا اصولا عدد وزن و به طبع آن میانگین وزن بیش از فشار خون می باشد. لذا چون انحراف معیار دو صفت از نظر واحد برابر نیست، قابل مقایسه نمی باشد. به منظور رفع این اشکال از نسبت انحراف معیار به میانگین که معمولا بصورت درصد بیان می شود استفاده می گردد که به آن ضریب تغییرات گویند. چون انحراف معیار و میانگین هر دو از یک جنس هستند حاصل تقسیم آنها بر هم (CV) بدون واحد خواهد بود و لذا ضریب تغییرات دو متغیر با واحدهای متفاوت قابل مقایسه می باشد.

$$CV=\frac{SD}{X} \times 100$$

منبع: کتاب اصول پایه روش تحقیق در علوم پزشکی، شورای نویسندگان، کمیته پژوهشی دانشجویی دانشگاه علوم پزشکی ایران

تنظیم کنندگان: عارفه کبیرزاده، حامد رفیعی، پریا صانعی، غزال علوی، فاطمه هاشمی

[1] Coefficient of variation