alireza hesaraki

کارگاه مفاهیم ساده آماری

دکتر فاطمه نوری

متخصص آمار زیستی

عضو هیات علمی دانشگاه علوم پزشکی اصفهان

سرفصل های درخواستی

11- مشخصات دروس دوره آموزشی				
ردیف		تعدادساعات تدریس		
		نظری	عملی	جمع
1	تفسیر آمار توصیفی و خلاصه سازی داده های پیوسته با ان	1	–	1
2	آشنایی با مفهوم فاصله اطمینان و نحوه محاسبه آن	2	–	2
3	آشنایی با مفاهیم آزمون فرضیه	1	–	1
4	آشنایی با آزمون های مقایسه ای	2	–	2
5	آشنایی با تجزیه و تحلیل رگرسیون	2	–	2
6	آشنایی با تکنیک های ناپارامتریک	1	–	1

- علم آمار مشتمل بر
- روشهای جمع آوری داده (Data collection)

(استخراج نمونه از جامعه مورد نظر طبق ضوابطی مقبول بطوریکه نمونه حاضر نماینده خوبی برای جامعه باشد)

- خلاصه سازی اطلاعات (آمار توصیفی Descriptive statistics)

(روشهایی برای سازماندهی و خلاصه کردن خصوصیات مهم مجموعه داده ها، متناسب با نوع داده)

- تجزیه و تحلیل داده ها (آمار استنباطی Inferential statistics)

(استفاده از اطلاعاتی که یک نمونه از جامعه آماری در اختیارمان می‌گذارد تا رسیدن به نتیجه ای در مورد این جامعه، مراحل یک تحلیل آماری هستند)

- استخراج و تفسیر نتایج از داده ها (Interpretation of results)

(فرآیند تفسیر و معنا بخشیدن به یافته های بدست آمده)

روشهای جمع آوری داده (Data collection)

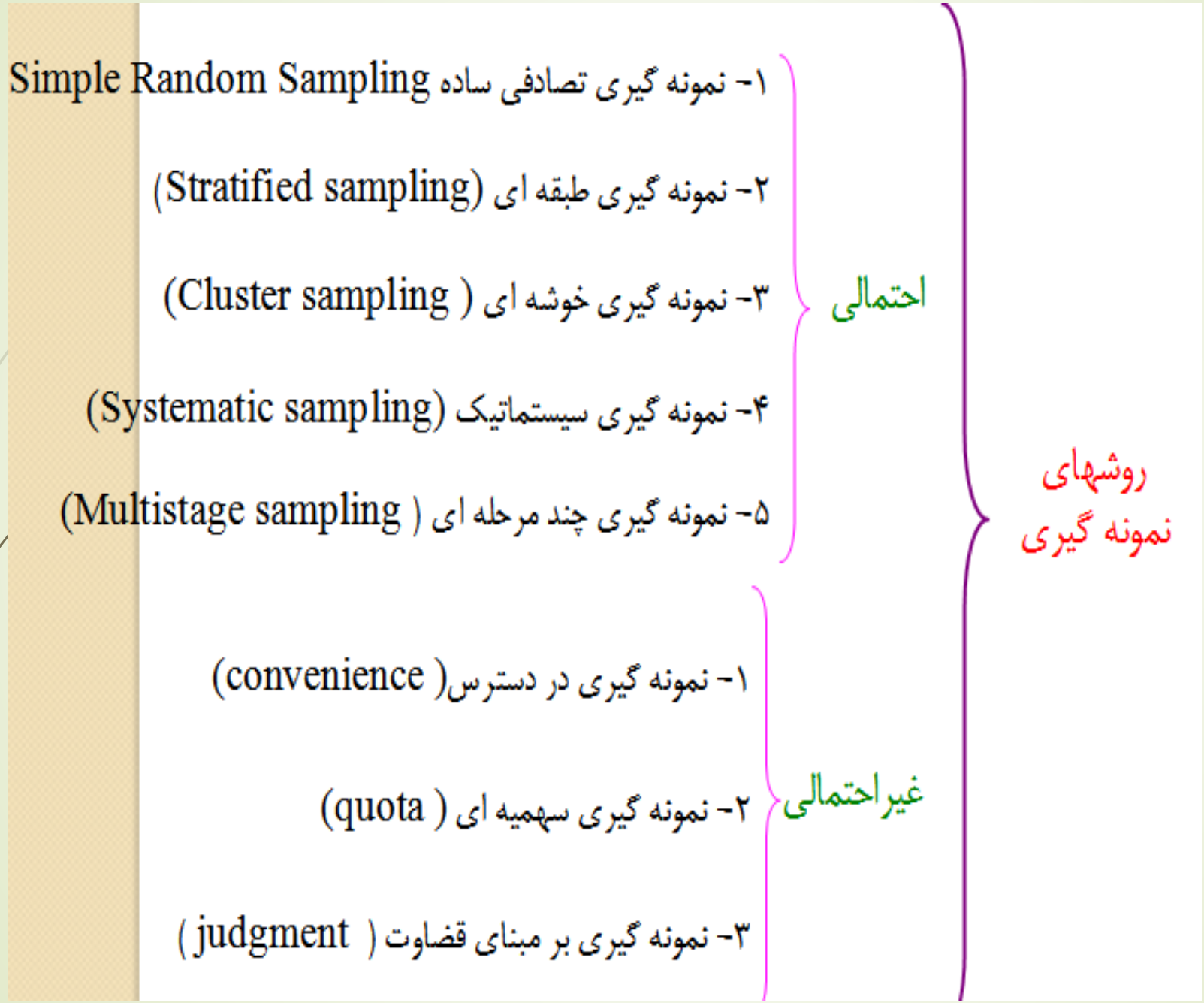
تعیین حجم نمونه (Sample Size) مورد نیاز

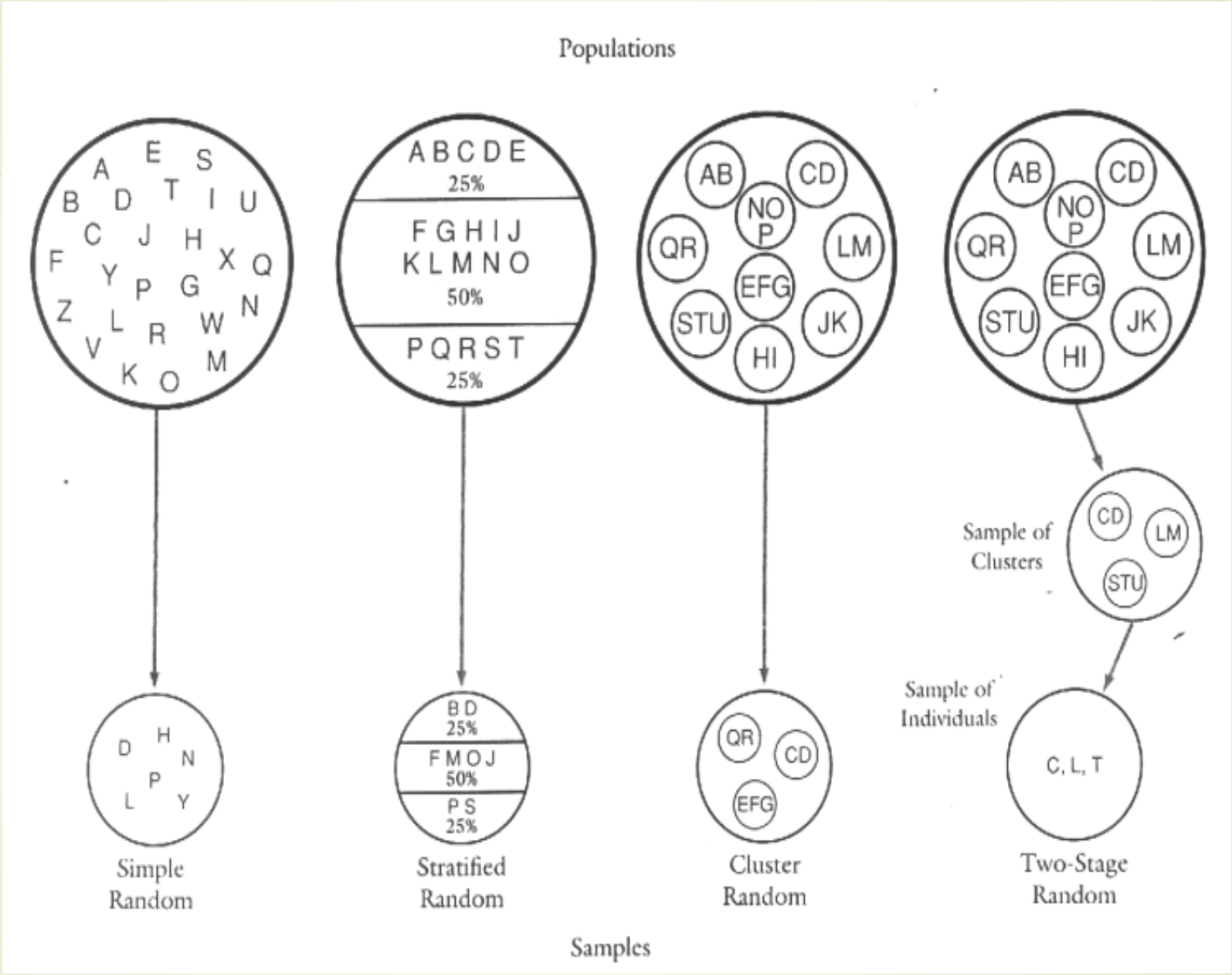
تعیین تعداد داده ها به طوریکه نتیجه های حاصل از تجزیه و تحلیل داده ها را بتوان با درجه دقت مورد نظر بیان کرد.

همچنین اگر در زمینه ای از مطالعات، انجام آزمایشات پر هزینه باشد علم آمار در تعیین تعداد داده های لازم برای بدست آوردن نتیجه هایی که از میزان اعتبار مطلوب برخوردار باشد کاربرد دارد.

تعیین روشهای نمونه گیری (Sampling methods)

تصمیم گیری در مورد نحوه انتخاب آزمودنی های مورد مطالعه و چگونگی جمع آوری اطلاعات از جامعه مورد نظر

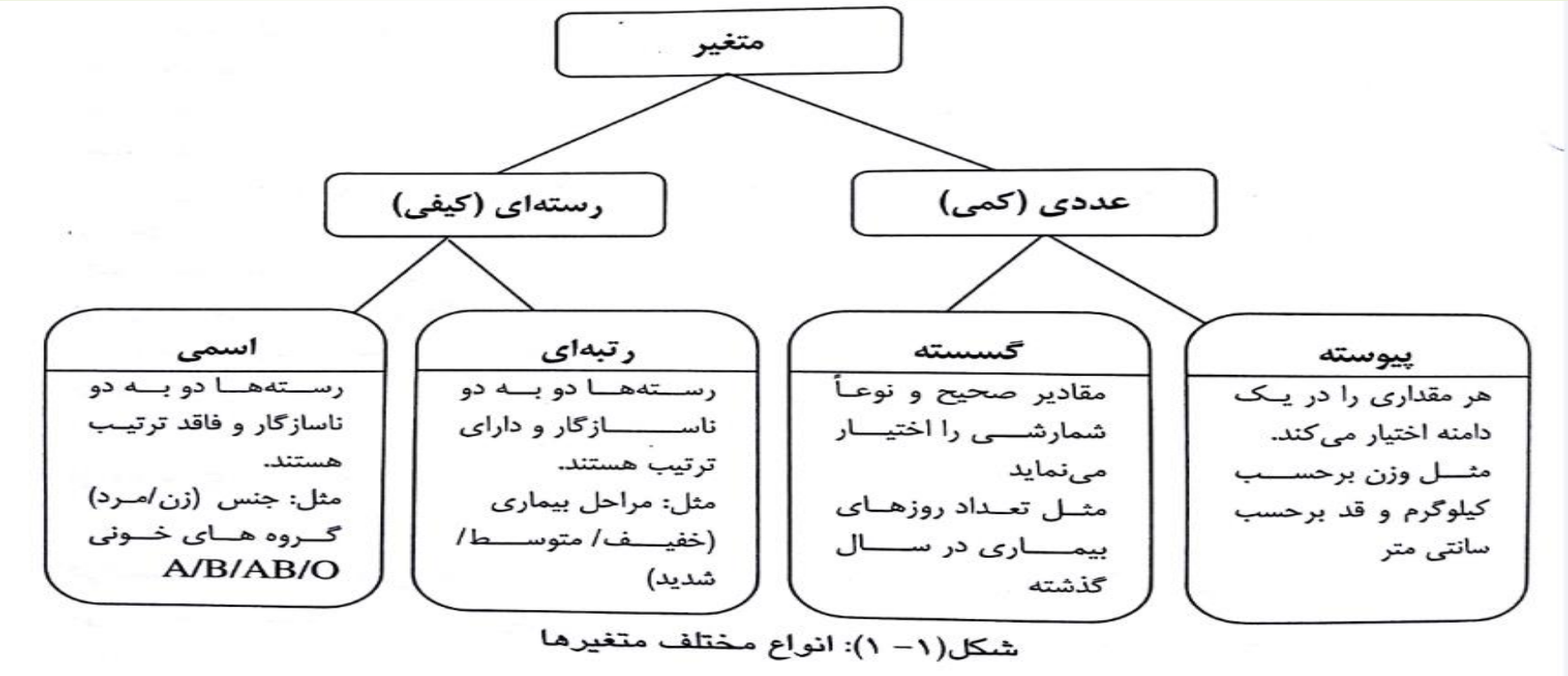

روشهای نمونه گیری
احتمالی
۱- نمونه گیری تصادفی ساده Simple Random Sampling
۲- نمونه گیری طبقه ای (Stratified sampling)
۳- نمونه گیری خوشه ای (Cluster sampling)
۴- نمونه گیری سیستماتیک (Systematic sampling)
۵- نمونه گیری چند مرحله ای (Multistage sampling)
غیراحتمالی
۱- نمونه گیری در دسترس(convenience)
۲- نمونه گیری سهمیه ای (quota)
۳- نمونه گیری بر مبنای قضاوت (judgment)


Populations
A E S B D T I U C J H X Q F Y P G N Z L R W V K O M
ABCDE 25% FGHIJ KLMNO 50% PQRST 25%
AB CD NO P QR LM EFG STU JK HI
Sample of Clusters
CD LM STU
Sample of Individuals
C, L, T
D H N P L Y
BD 25% FMOJ 50% PS 25%
QR CD EFG
Simple Random
Stratified Random
Cluster Random
Two-Stage Random
Samples

خلاصه سازی اطلاعات (Descriptive statistics)

آمار توصیفی: روشهایی برای سازماندهی و خلاصه کردن خصوصیات مهم مجموعه داده ها، متناسب با **نوع داده یا در اصطلاح نوع متغیر**

نوع هر متغیر و در نتیجه داده های مبتنی بر آن، به دو نوع اصلی و چهار نوع فرعی تقسیم می شوند:



شکل(۱– ۱): انواع مختلف متغیرها

- نوع هر متغیر و در نتیجه داده های مبتنی بر آن، به دو نوع اصلی تقسیم می شوند:

1- *متغیرها یا داده های رسته ای(کیفی) Qualitative*

نمی توان آنها را اندازه گیری کرد بلکه می توان آنها را طبقه بندی نمود که خود به دو گروه تقسیم می شود

الف) متغیر رسته ای اسمی: Nominal

نتیجه سنجش متغیرها در آزمودنی ها گاهی چنان است که تنها می توان آزمودنی را بر اساس آن ویژگی به یک گروه منتسب کرد و نام آن گروه را به آنها داد مثل گروه خونی A, B, AB, O یا وضعیت تاهل (1-مجرد، 2-متاهل، 3-همسر فوت شده، 4-مطلقه) یا جنس (1-زن و 2-مرد)

در داده های اسمی ، رسته ها دوبه دو ناسازگار و فاقد ترتیب هستند.

ب) متغیر رسته ای ترتیبی:Ordinal

در این نوع داده ها، رسته های متغیر از جهاتی دارای ترتیب هستند. مثلا فرض کنید شدت یک بیماری به سه رسته " 1-خفیف، 2-متوسط و 3-شدید" تقسیم شود. بر اساس ضوابطی هر فرد به یکی از رسته های مذکور منتسب می شود. دقت کنید اگرچه این رسته ها حاوی رتبه یا ترتیب هستند اما نمی توان گفت شدت بیماری در گروه متوسط دو برابر گروه ضعیف است. در داده های ترتیبی ، رسته ها دوبه دو ناسازگار و دارای ترتیب هستند.

2- *متغیر (داده) عددی(کمی) Quantitative*

به متغیری گفته می شود که قابل اندازه گیری باشد مثل قد و وزن یا سن افراد. این دسته خود به دو گروه فرعی تقسیم می شود

الف) متغیر (داده) عددی(کمی) پیوسته Continuous

کمیت پیوسته می تواند بین دو مقدار خود ، تمامی اعداد حقیقی را اختیار کند یا به تعبیری دیگر، متغیرهای پیوسته، می توانند هر مقداری را در دامنه اعداد حقیقی اختیار کنند مانند قد، وزن، کلسترول خون، قند خون

ب) متغیر (داده) عددی(کمی) گسسته Discrete

کمیتی است که بتواند به عنوان مقادیر خود، اعداد صحیح یا مجموعه شمارش پذیر اعداد (یا زیر مجموعه ای از آن) را اختیار کند. مانند تعداد افراد خانوار، تعداد دندانهای فاسد هر فرد، تعداد بیماران مراجعه کننده به یک پزشک در یک سال، تعداد روزهای بیماری در سال گذشته

خلاصه سازی اطلاعات (Descriptive statistics)

آمار توصیفی : روشهایی برای سازماندهی و خلاصه کردن خصوصیات مهم مجموعه داده، متناسب با نوع داده یا در اصطلاح نوع متغیر

آمار توصیفی مشتمل بر

۱- تلخیص داده ها با استفاده از **محاسبه شاخص های عددی**

شاخص های مرکزی (مانند میانگین ، میانه، مد)

- شاخص های پراکندگی (مانند انحراف معیار، واریانس، دامنه، دامنه میان چارکی)

۲- تلخیص داده ها با استفاده از **نمودارها**

خلاصه سازی اطلاعات (Descriptive statistics)

آمار توصیفی:
معرفی شاخص های مرکزی

میانگین (Mean)

میانگین حسابی با جمع کردن همه مقادیر یک جامعه یا نمونه و تقسیم آنها بر تعداد مقادیری که باهم جمع شده اند بدست میآید. میانگین جامعه را با نماد μ و میانگین نمونه را با نماد $\bar{x}$ نمایش می دهند. فرمول های میانگین جامعه و نمونه به ترتیب به صورت زیر است:

$$\mu = \frac{X_1 + X_2 + \cdots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

مثال: فرض کنید جامعه ما متشکل از 200 بزرگسال بالاتر از 20 سال مبتلا به دیابت است. نمونه ای 5 نفره از آنها استخراج شده سن آنها 22، 44، 45، 37 ، 70 سال می باشد میانگین نمونه ای سن افراد مبتلا به دیابت چقدر است؟

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i}{5} = \frac{22 + 44 + 45 + 37 + 70}{5} = 43.6$$

میانگین (Mean)

- به تک تک داده هاحساس است و از همه اعضای نمونه جهت محاسبه میانگین استفاده می گردد
- میانگین در نمونه های مختلف استخراجی از یک جامعه، بیشتر از نما و میانه به همدیگر نزدیک می باشند
- اگر تمام داده ها با عدد ثابتی جمع یا تفریق یا ضرب یا تقسیم شوند میانگین نیز با آن عدد جمع یا تفریق یا ضرب یا تقسیم می شود.
- مجموع انحراف داده ها از میانگین، برابر صفر است
- مجموع مجذور انحراف داده ها از میانگین همیشه کوچکتر یا مساوی مجموع مجذور انحراف داده ها از هر عدد یا شاخص دیگری مثل میانه یا مد است.
- میانگین تحت تاثیر داده های خیلی کوچک و یا خیلی بزرگ (**داده های پرت** Outliers) می باشد.
- از شاخص میانگین جهت توصیف **متغیرهای کمی پیوسته و گسسته** استفاده می شود.

نما (Mode):

- یکی دیگر از شاخص های مرکزی است. عددی در مجموعه داده ها که بیش از همه تکرار شده باشد را مد گویند.
- توجه شود اگر همه مقادیر با هم متفاوت باشند یعنی هر یک از مقادیر فقط یکبار رخ داده باشد یا همه داده ها دارای فروانی یکسانی باشند این مجموعه داده، نما ندارد.
- از سویی دیگر ممکن است یک مجموعه داده چند نما داشته باشد.
- **مثال:** نمای مجموعه داده زیر را مشخص کنید

20, 21, 20, 20, 34, 22, 24, 27, 27, 27

می توان گفت این مجموعه داده دارای دو نما با مقادیر ۲۷ و ۲۰ می باشد.

- توجه شود که علاوه بر داده های کمی گسسته و پیوسته، برای داده های کیفی هم می توان مد را محاسبه نمود. آن رسته ای که بیش از همه به وقوع پیوسته است را رسته نمایی می گویند.
- نما همیشه در مرکز توزیع فراوانی قرار ندارد از این رو نمی توان به عنوان یک شاخص مرکزی تمام عیار به آن اطمینان داشت. همچنین از همه داده ها جهت محاسبه نما استفاده نمی گردد. زمانی از نما استفاده می شود که نیاز به تعیین شاخص مرکزی به صورت تقریبی و سریع باشد.

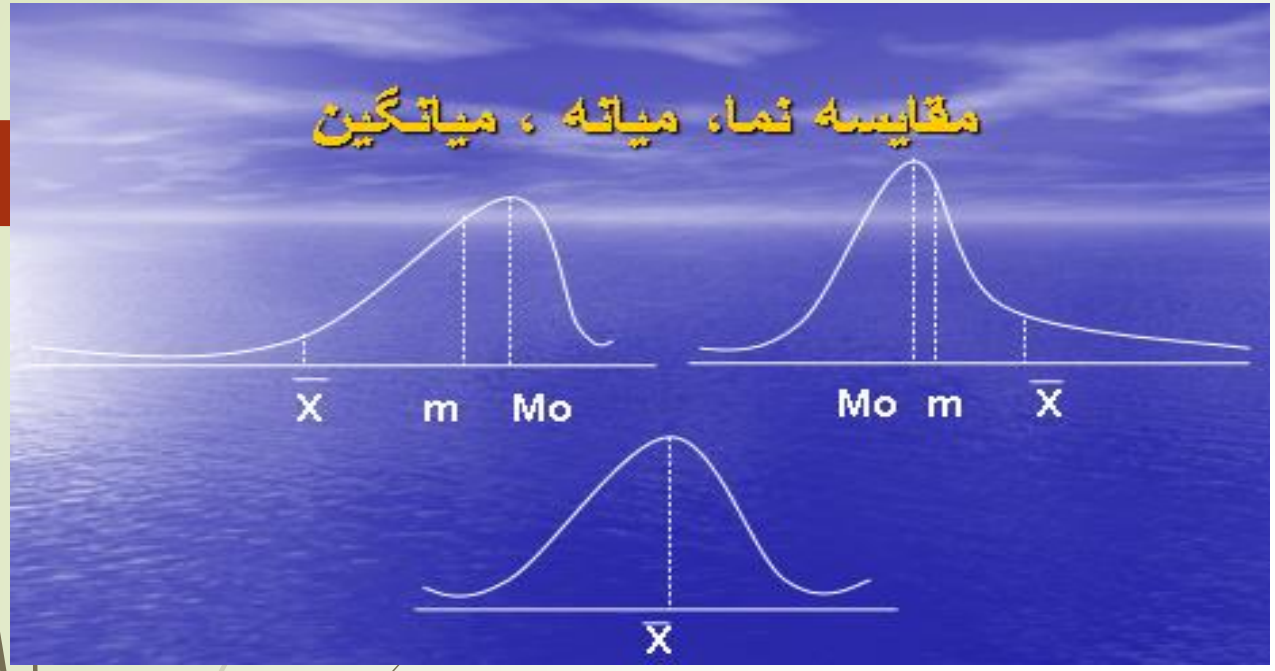
میانه (Median):

- اگر داده ها را به ترتیب از کوچکترین به بزرگترین مقدار مرتب کنیم داده ای که در وسط داده های دیگر قرار می گیرد را میانه گویند.
- به عبارت دیگر میانه داده ها را به دو دسته تقسیم می کند نصف داده ها (۵۰ درصد) کمتر از میانه و نصف دیگر بالای میانه قرار دارند.
- توجه کنید اگر تعداد داده ها عددی فرد باشد و همه مقادیر را به ترتیب صعودی مرتب کنیم میانه ، عدد میانی این داده ها خواهد شد.
- در صورتی که تعداد مشاهدات عدد زوجی باشد مشاهده میانی وجود ندارد ولی دو عدد میانی خواهیم داشت که میانه عبارت است از میانگین این دو عدد میانی.
- مثال: فرض کنید جامعه ما متشکل از ۲۰۰ بزرگسال بالاتر از ۲۰ سال مبتلا به دیابت است. نمونه ای ۵ نفره از آنها استخراج شده سن آنها 22, 37, 44, 45, 70سال می باشد میانه نمونه ای سن افراد مبتلا به دیابت ۴۴ سال است.
- برای محاسبه میانه از همه داده های جمع آوری شده استفاده نمی شود. نهایتا از دو داده استفاده می گردد.
- میانه نسبت به داده های خیلی بزرگ یا داده های خیلی کوچک (داده های پرت Outliers) حساس نیست بنابراین در این شرایط، شاخص خوبی است که تمرکز داده ها را در وسط توزیع نشان دهد
- از شاخص میانه برای توصیف متغیرهای کمی پیوسته و گسسته و متغیرهای رتبه ای می توان استفاده نمود.

میانه (Median)

- بطور خلاصه، هرگاه داده ها را از کوچک به بزرگ مرتب کنیم آنگاه میانه، مقدار میانی داده های مرتب شده می باشد چنانچه تعداد مشاهدات یعنی n ، عددی فرد باشد براحتی مقدار میانه قابل محاسبه و برابر $\frac{n+1}{2}$امین مشاهده مرتب شده خواهد بود.
- مثال: اگر n=211 باشد آنگاه میانه $\frac{211+1}{2} = 106$ امین مشاهده در مجموعه داده های مرتب شده خواهد بود.
- چنانچه n عددی زوج باشد میانه به طور صریح در داده های موجود قرار ندارد بلکه با محاسبه میانگین دو مشاهده میانی بدست می آید. (یعنی میانگین دو مشاهده $\frac{n}{2}$ و $\frac{n}{2}+1$).
- مثال: اگر n=200 باشد میانگین دو مشاهده 100 ام ($\frac{n}{2} = \frac{200}{2} = 100$) و 101 ام ($\frac{n}{2}+1 = \frac{100}{2}+1 = 101$) در مجموعه داده های مرتب شده، میانه خواهد بود.

مقایسه نما، میانه ، میانگین



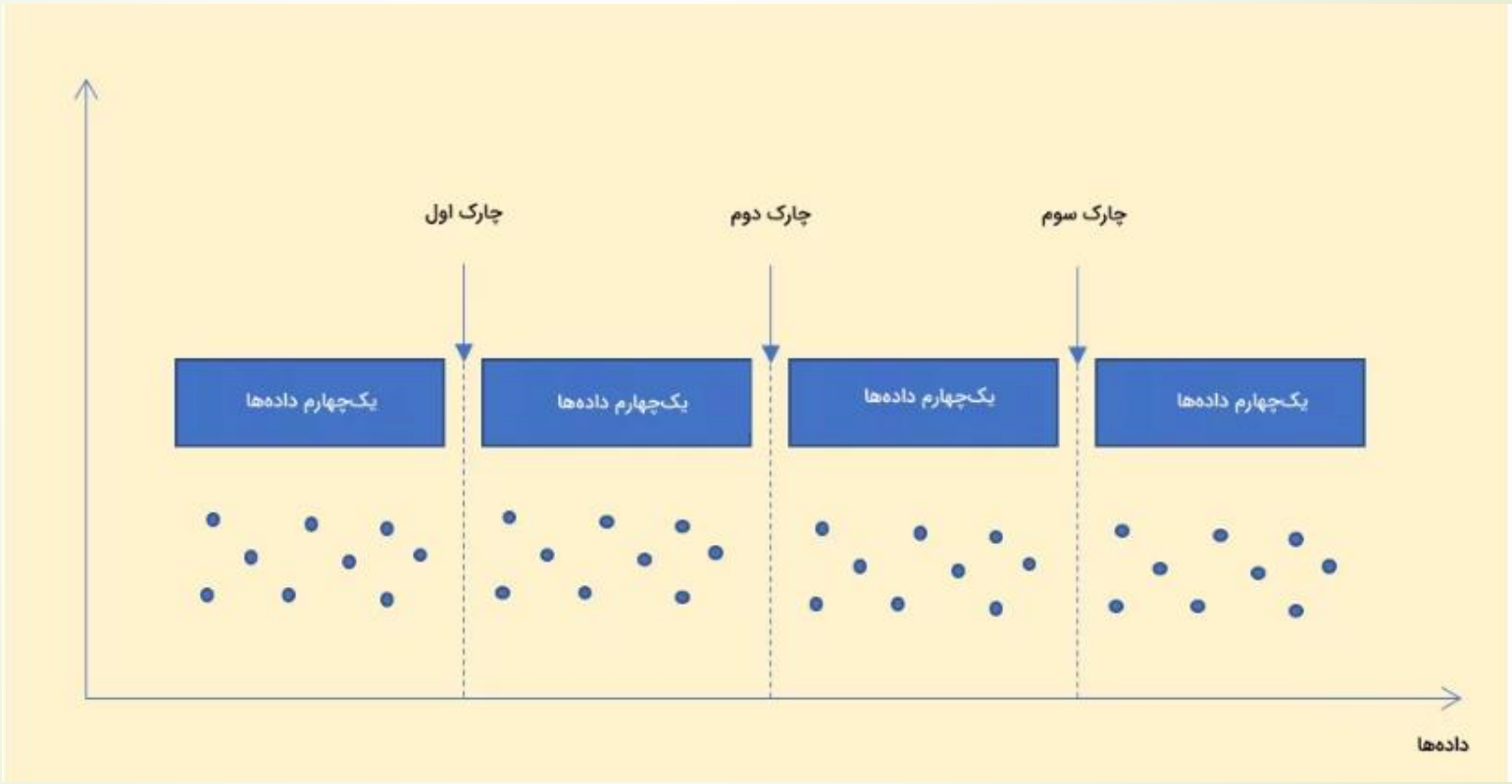
$MO = m = \overline{X}$ اگر منحنی نرمال باشد

$MO < m < \overline{X}$ اگر کجی مثبت باشد

$MO > m > \overline{X}$ اگر کجی منفی باشد

چارک ها Quartile در مجموعه داده

- گاهی مفید است که مفهوم میانه را تعمیم دهیم و مجموعه داده های مرتب شده را به چهار قسمت تقسیم کنیم همانطور که نقطه تفسیم داده ها به دو نیمه، میانه خوانده می شود نقاط تقسیم داده ها به چهار قسمت را چارک Quartile می نامند.
- در یک نمونه یا در یک مجموعه داده، چارک معادل است با سه مقدار عددی که داده‌ها را به چهار گروه تقسیم‌بندی می‌کند که عبارتند از چارک اول یا چارک پایین (عددی است که بین مینیمم داده ها و عدد وسط قرار می‌گیرد)، چارک دوم یا میانه و چارک سوم یا چارک بالا (عددی است که بین عدد وسط و ماکزیمم داده ها قرار می‌گیرد).
- چارک‌ها نوعی «چندک» Quantileیا به‌طور دقیق‌تر، نوعی «صدک» Percentileمحسوب می‌شوند.
- صدک‌ها Percentiles: متشکل است از ۹۹ صدک که داده‌ها را به صد بخش تقسیم می‌کنند.
- به‌طور کلی، kدرصد از داده‌ها زیر صدک kام قرار می‌گیرند.
- دهک‌ها Deciles: متشکل است از نه دهک که داده‌ها را به ده بخش تقسیم می‌کنند.
- چارک اول یا چارک پایین‌تر صدک ۲۵ام است، به این معنا که ۲۵ درصد از داده‌ها زیر چارک اول قرار می‌گیرند
- چارک دوم یا میانه صدک ۵۰ام است، به این معنا که ۵۰ درصد از داده‌ها زیر چارک دوم قرار می‌گیرند.
- چارک سوم یا چارک بالاتر صدک ۷۵ام است، به این معنا که ۷۵ درصد از داده‌ها زیر چارک سوم قرار می‌گیرند.


چارک اول
چارک دوم
چارک سوم
یک‌چهارم داده‌ها
یک‌چهارم داده‌ها
یک‌چهارم داده‌ها
یک‌چهارم داده‌ها
داده‌ها

طریقه محاسبه چارک ها در مجموعه داده

به طور کلی چارک اول عبارت است از

$Q1= x_r+w*(x_{r+1}-x_r)$

که در آن r رقم صحیح حاصل از (n+1)/4 است و w رقم اعشار آن است

برای تعیین چارک اول :

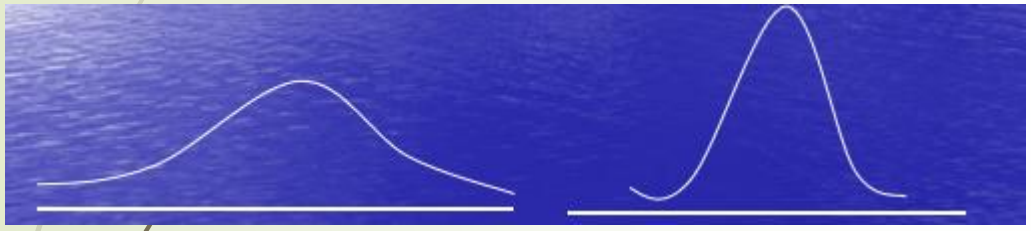
1- ابتدا مجموعه داده ها را مرتب می کنیم
2- سپس(n+1)/4 را حساب میکنیم و اسم آنرا b می گذاریم
3- اگر b یک عدد صحیح بود چارک اول،b امین مشاهده از مجموعه داده های مرتب شده خواهد بود
4- اگر b عدد صحیح نباشد پس رقم صحیح آنرا r و رقم اعشار آنرا w می نامیم
5- تفاضل مشاهده r+1 ام و مشاهده r ام را محاسبه کرده و حاصل را با w ضرب می کنیم و آنرا c می نامیم
6- چارک اول عبارت است از مقدار مشاهده r ام به اضافه c

توجه شود برای یافتن چارک سوم کافی است در همه مراحل بالا به جای (n+1)/4 از 3*(n+1)/4

آمار توصیفی:
معرفی شاخص های پراکندگی

شاخص های پراکندگی:

- شاخص های پراکندگی میزان تغییرات یا پراکندگی که در بین داده ها وجود دارد را نشان می دهد به عبارت دیگر، پراکندگی یک مجموعه داده به تنوعی که مقادیر مشاهدات از خود نمایش می دهند بر می گردد
- اگر همه مقادیر یکسان باشند پراکندگی وجود ندارد. هرچقدر داده ها از هم دورتر باشند پراکندگی بزرگتر و هرچقدر داده ها به هم نزدیکتر باشند پراکندگی کوچکتر خواهند بود.
- ممکن است میانگین دو مجموعه داده یکشان باشد اما پراکندگی این دو مجموعه داده در اطراف میانگین متفاوت باشد



- جهت توصیف هر داده یا متغیری لازم است یک شاخص مرکزی به همراه یک شاخص پراکندگی گزارش گردد.
- انواع شاخص های پراکندگی عبارتند از دامنه تغییرات ، دامنه میان چارکی، واریانس و انحراف معیار

- دامنه تغییرات (Range)
- تفاوت بین کوچکترین و بزرگترین داده را دامنه تغییرات گویند.
- چون تنها دو مقدار در محاسبه این معیار دخالت دارند این شاخص به عنوان شاخص ضعیفی از پراکندگی به حساب می آید و قادر به توصیف حقیقی پراکندگی داده ها نیست
- حسن عمده آن سهولت محاسبه آن است.
- اگر در داده ها، مشاهدات دور افتاده (داده های خیلی بزرگ یا خیلی کوک) وجود داشته باشد دامنه یک شاخص گمراه کننده جهت توصیف پراکندگی خواهد بود.

- **دامنه میان چارکی (Inter Quartile Range) :**
- یکی دیگر از شاخص های پراکندگی است که با تفاضل بین چارک اول و سوم بدست می آید.
- **دامنه میان چارکی = چارک سوم – چارک اول**
- دامنه میان چارکی، طول فاصله ای را که نیمه مرکزی مشاهدات در بر دارد را نمایش می دهد
- این شاخص تحت تاثیر داده های پرت قرار نمی گیرد پس در این شرایط، شاخص خوبی جهت توصیف پراکندگی دادهاست.
- هرجا گزارشی از میانه داشته باشیم از این شاخص به عنوان معیاری برای پراکندگی استفاده می کنیم
- برای محاسبه این شاخص، از همه داده های جمع آوری شده استفاده نمی شود.
- از چارک ها به همراه دامنه میان چارکی می توان در تشخیص داده های پرت به طور تقریبی استفاده نمود ، هر دادهای که از $Q3+1.5IQR$ بیشتر و از $Q1-1.5IQR$کمتر باشد، داده پرت محسوب می‌شود.

واریانس (Variance)

- به دنبال یک شاخص پراکندگی باثبات و معتبر هستیم به طوری که از همه اعداد در محاسبه استفاده نماید
- یکی از راههای اندازه گیری پراکندگی در داده ها، ارزیابی میزان انحراف مشاهدات از میانگین داده ها است.

- اگر نمونه ای مشتمل بر n مشاهده داشته باشیم

1. میانگین این مجموعه داده را پیدا می کنیم

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1. تفاضل تک تک مشاهدات از این میانگین محاسبه می شود که به هریک، انحراف از میانگین گفته می شود

- $(x_1 - \bar{x}), (x_2 - \bar{x}) , (x_3 - \bar{x}), \ldots\ldots, (x_n - \bar{x})$

1. تک تک این اختلافات را به توان دو می رسانیم

- $(x_1 - \bar{x})^2, (x_2 - \bar{x})^2 , (x_3 - \bar{x})^2, \ldots\ldots, (x_n - \bar{x})^2$
-

1. همه توان دوم انحرافات را باهم جمع می کنیم

- $(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots\ldots + (x_n - \bar{x})^2$

1. حاصل را به n-1 تقسیم می کنیم

- فرمول کلی واریانس نمونه به صورت زیر است

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

انحراف معیار (Standard Deviation)

- توجه شود واحد واریانس توان دوم واحد داده های اصلی است چون در محاسبه آن از توان دوم انحرافات استفاده شده است.
- یعنی اگر متغیر مورد نظر، وزن افراد بر حسب کیلوگرم باشد واحد واریانس کیلوگرم به توان دو است.
- مطلوب است معیار پراکندگی بسازیم که به واحد داده های اصلی باشد.
- کافی است از واریانس مجموعه داده، جذر بگیریم. این معیار را انحراف معیار (انحراف استاندارد) داده ها می نامند که یک معیار پراکندگی مفید و در واقع معتبرترین معیار است.
- $SD = \sqrt{var(x)}$
- واحد انحراف معیار به واحد داده های اصلی است.
- از شاخص انحراف معیار در همراهی با شاخص مرکزی میانگین، برای توصیف داده ها استفاده می گردد
- در صورتی که همه داده ها با یک عدد ثابت جمع شود یا از آن کم شود واریانس و انحراف استاندارد تغییری نمی کنند
- در صورتی که عدد ثابتی در داده ها ضرب یا بر آن تقسیم شود انحراف استاندارد نیز در همان عدد ضرب یا بر آن تقسیم خواهد شد
- در صورتی که عدد ثابتی در داده ها ضرب یا بر آن تقسیم شود واریانس نیز در توان دوم آن عدد ضرب یا بر آن تقسیم خواهد شد

ضریب تغییرات (Coefficient of Variation)

- اگر یک مجموعه داده یا در اصطلاح یک متغیر داشته باشیم انحراف معیار شاخص خوبی برای ارزیابی پراکندگی است اما اگر دو متغیر مورد بررسی قرار گیرند دیگر انحراف معیار شاخص مناسبی نیست
- تنحراف معیار ممکن است منجر به نتیجه گیری نادرستی گردد زیرا دو نوع متغیری که ارزیابی می گردد ممکن است با واحدهای مختلف اندازه گیری شده باشند
- مثلا بخواهیم سطح کلسترول خون که بر حسب میلی گرم در دسی لیتر اندازه گیری شده است را نسبت به وزن بر حسب کیلوگرم مقایسه کنیم و ببینیم کدامیک تغییرات بیشتری دارد؟ از آنجا که این دو با واحدهای متفاوتی اندازه گیری شده اند و انحراف معیار هم واحدی شبیه واحد داده های اندازه گیری شده دارد پس مقایسه انحراف معیار آن دو براحتی صورت نمی گیرد.
- علاوه بر آن، اگر واحدهای اندازه گیری یکسانی داشته باشند ممکن است میانگین های کاملا متفاوتی از هم داشته باشند و مقایسه انحراف معیارها مشکل باشد.

ضریب تغییرات (Coefficient of Variation)

- پس به معیاری برای مقایسه پراکندگی دو مجموعه داده نیاز داریم که **هم، واحد اندازه گیری متفاوت و هم، اثر اختلاف میانگین در آن** لحاظ شود.
- به این شاخص، ضریب تغییرات گفته می شود که در آن انحراف معیار به میانگین تقسیم می شود و معمولا با درصد بیان می شود
- $CV = \frac{SD}{\bar{X}} \times 100$
- چون انحراف معیار و میانگین هم واحدند و هر دو به واحد داده های اصلی هستند، ضریب تغییرات به واحد اندازه گیری بستگی ندارد.

مثال: دو گروه از افراد یکی ۲۵ ساله و دیگری ۱۱ ساله در نمونه ما حضور دارند میانگین وزن و انحراف معیار آنها بر حسب کیلوگرم در جدول زیر آمده است پراکندگی وزن ۲۵ ساله ها بیشتر است یا ۱۱ ساله ها؟ مقایسه کنید

سن	25 ساله ها	11 ساله ها
میانگین وزن	145 کیلوگرم	80 کیلوگرم
انحراف استاندارد	10 کیلوگرم	10 کیلوگرم
ضریب تغییرات	6.9	12.5
پراکندگی وزن 11 ساله بیشتر است.		

خلاصه سازی اطلاعات (Descriptive statistics)

آمار توصیفی : روشهایی برای سازماندهی و خلاصه کردن خصوصیات مهم مجموعه داده، متناسب با نوع داده یا در اصطلاح نوع متغیر

آمار توصیفی مشتمل بر

۱- تلخیص داده ها با استفاده از محاسبه شاخص های عددی

شاخص های مرکزی (مانند میانگین ، میانه، مد)

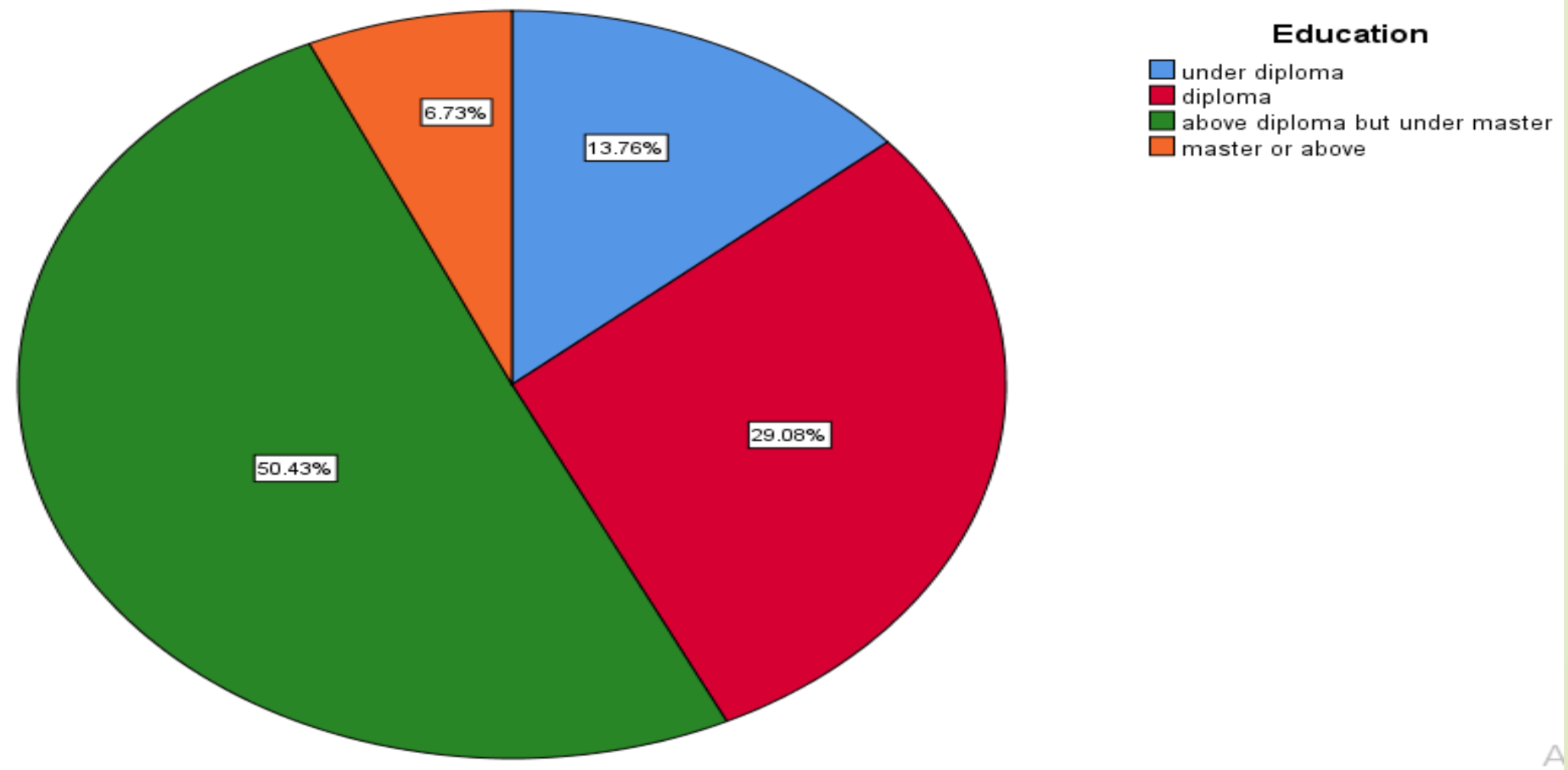
- شاخص های پراکندگی (مانند انحراف معیار، واریانس، دامنه، دامنه میان چارکی)

۲- تلخیص داده ها با استفاده از نمودارها

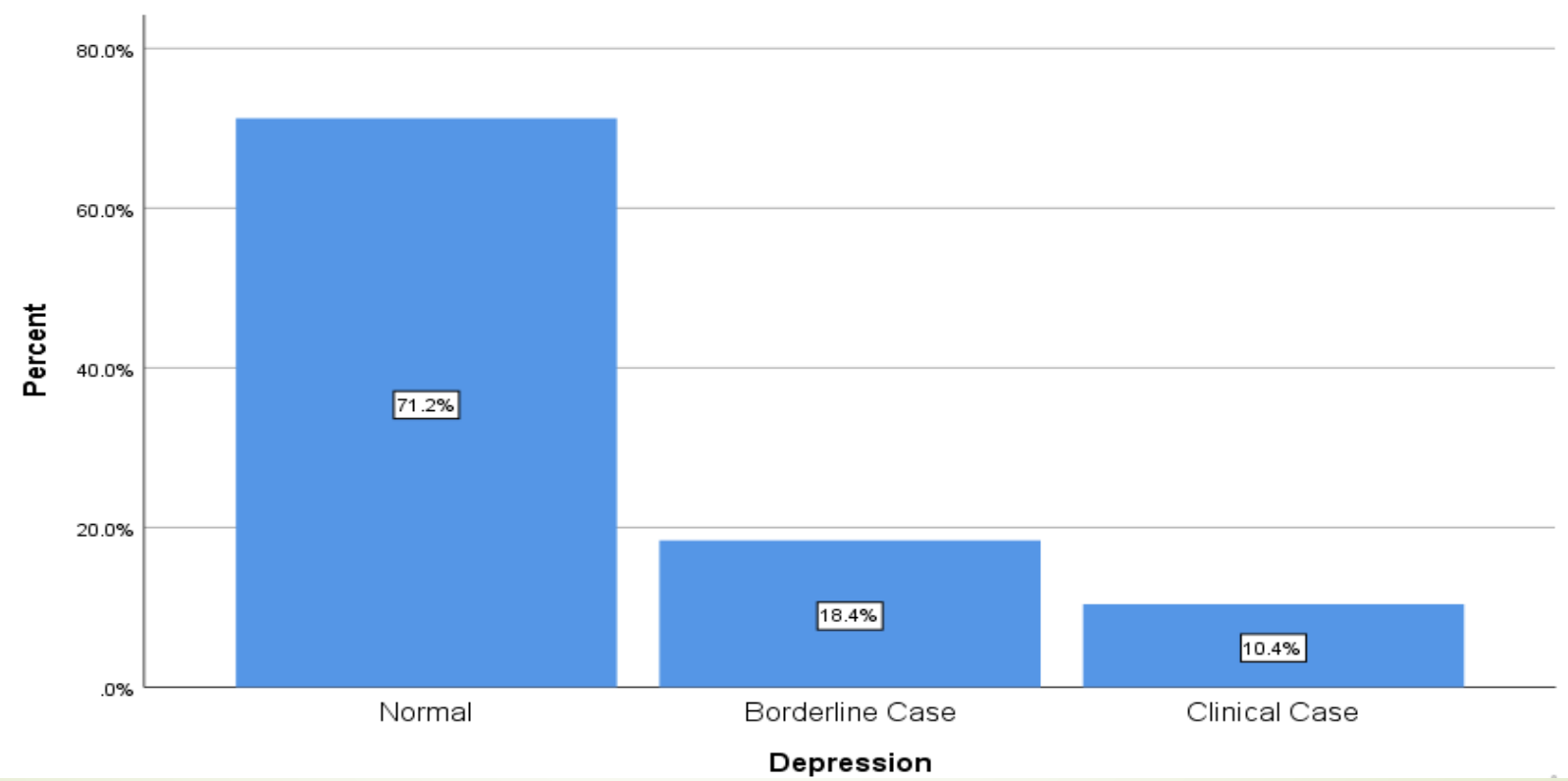
استفاده از نمودارها برای توصیف داده ها

- در نمودارها سرعت انتقال اطلاعات به مخاطب بیشتر است
- بسته به نوع داده مورد نظر (داده های کمی یا کیفی) می توان از نمودارهای مختلف جهت توصیف داده استفاده نمود
- متداولترین نمودارها برای گزارش داده های کیفی : Bar plot , Pie plot
- نمودارهای مناسب برای داده های کمی : , Histogram, Error bar, Box plot

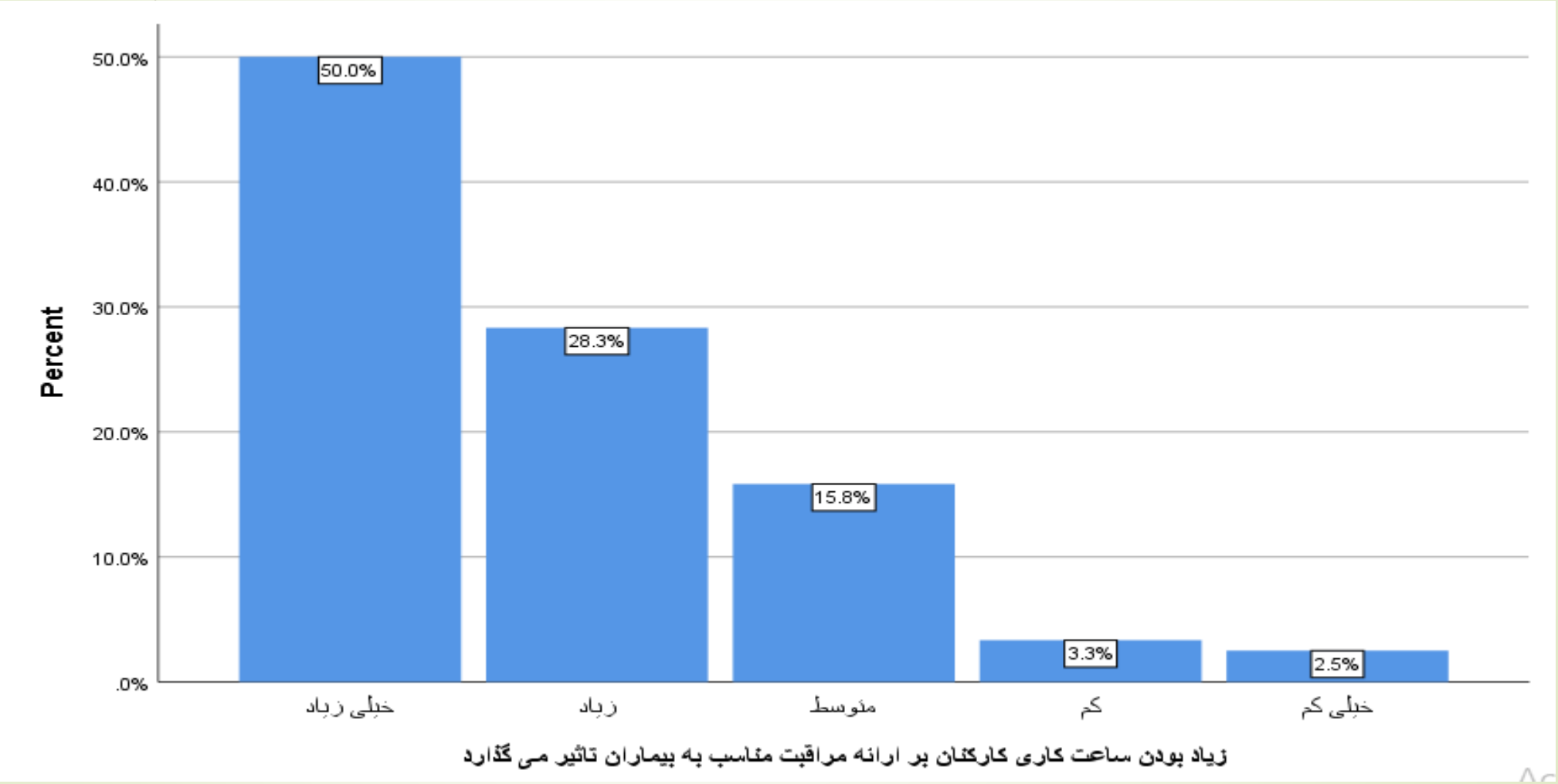
نمودار دایره ای (Pie plot) از میزان تحصیلات نمونه ای از کارمندان دانشگاه علوم پزشکی اصفهان در سال ۱۳۹۰



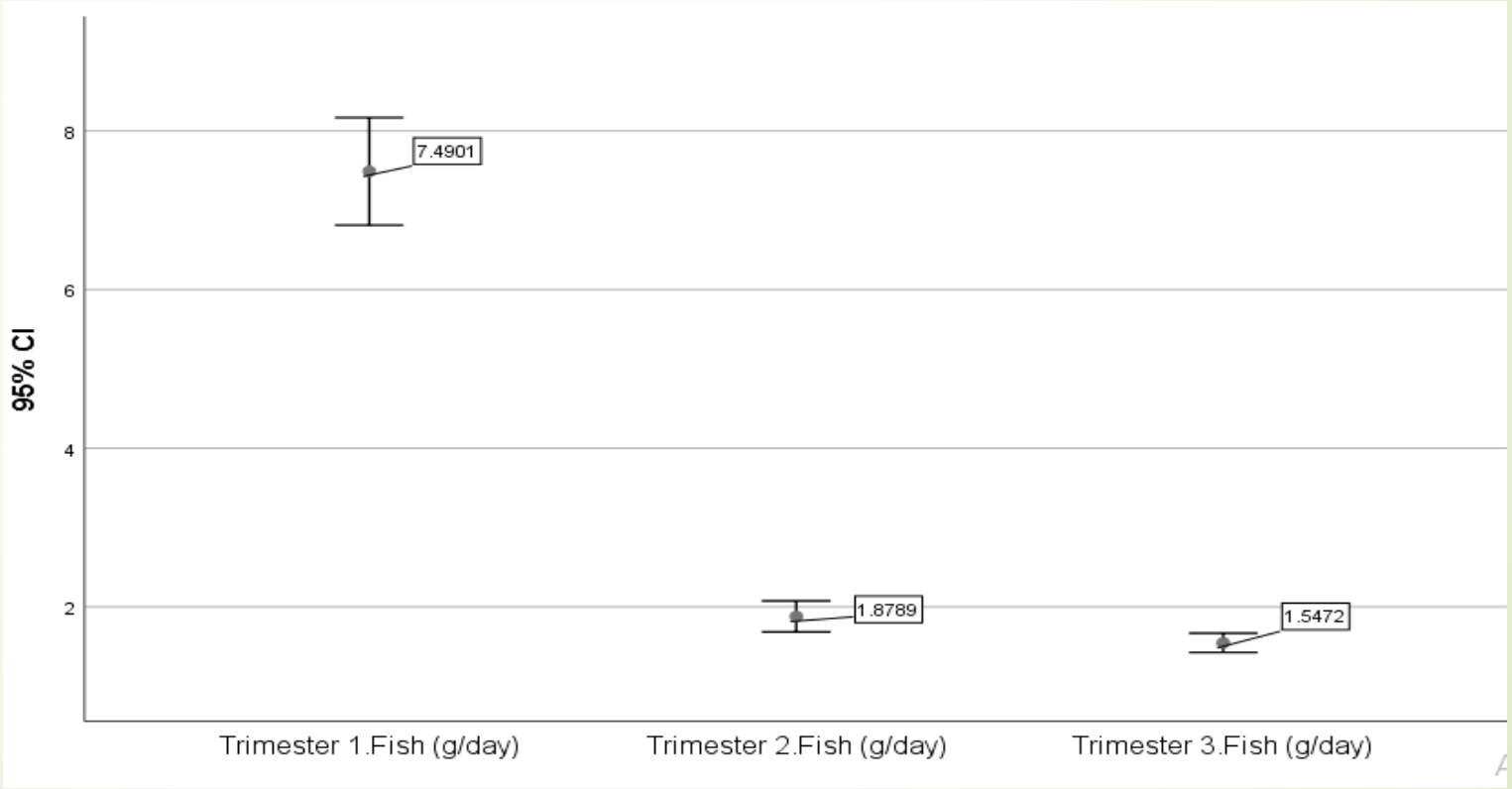
نمودار ستونی(Bar plot) از شدت افسردگی در نمونه ۵ هزار نفری از کارمندان دانشگاه علوم پزشکی اصفهان در سال ۱۳۹۰



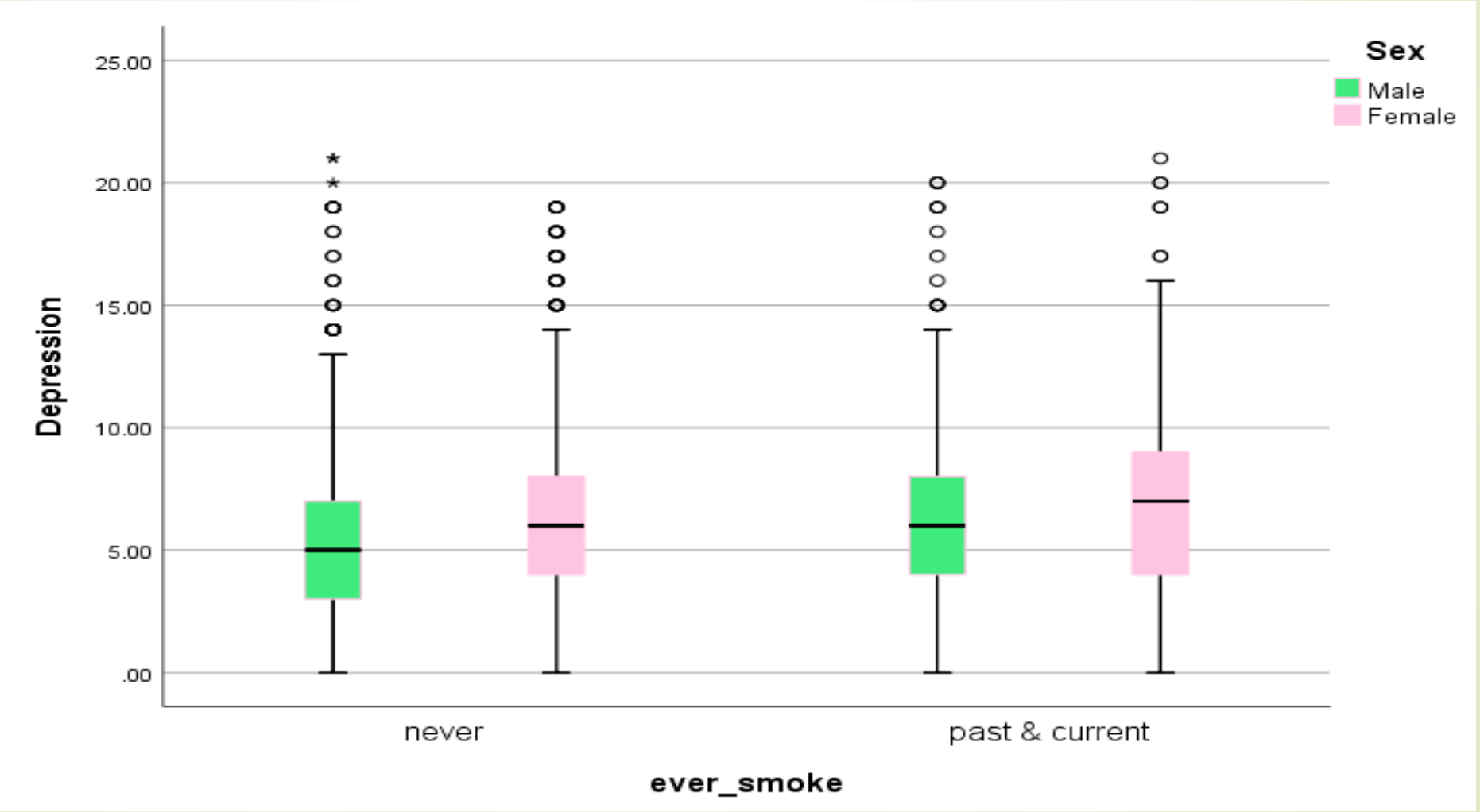
نمودار ستونی نظر کارکنان سلامت نسبت به کیفیت ارائه خدمات



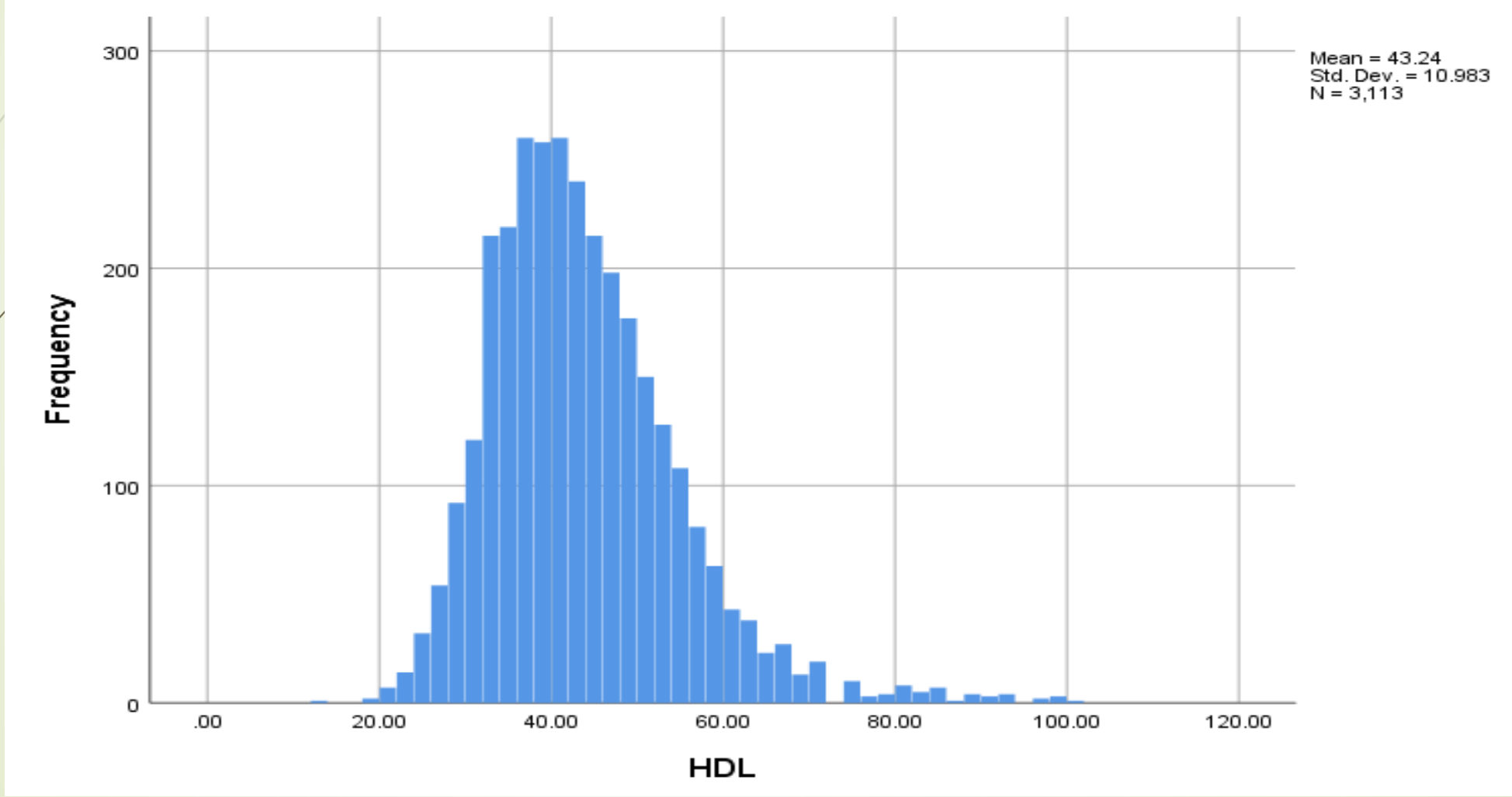
نمودار میانگین و فاصله اطمینان ۹۵ درصد (Error Bar) مصرف ماهی (گرم در روز) در سه ماهه اول، دوم و سوم بارداری مادر



نمودار جعبه ای (Boxplot) میانگین نمره افسردگی در زنان و مردان با و بدون تجربه استعمال سیگار



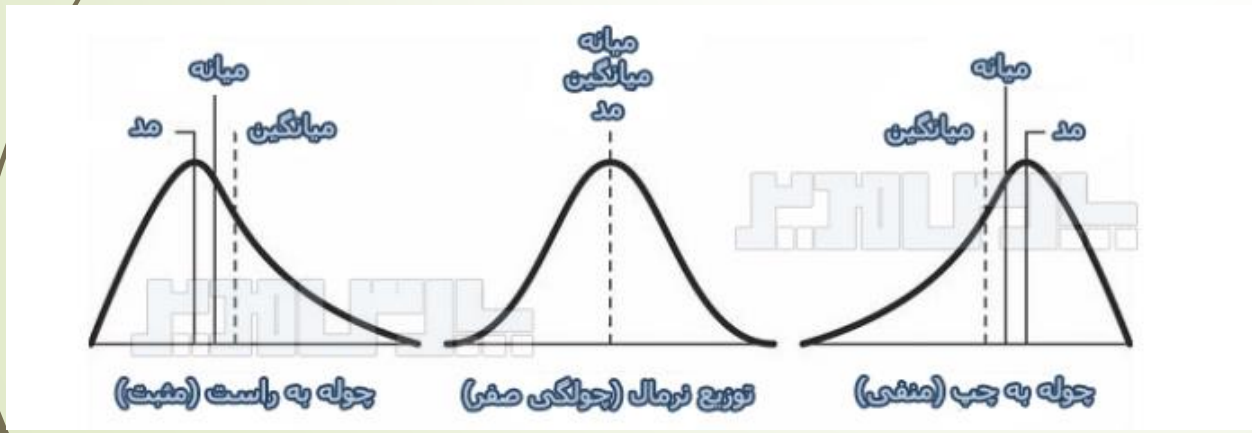
نمودار هیستوگرام چربی خوب HDL در بین افراد زیر ۶۰ سال



چولگی Skewness داده ها

- چولگی در آمار نشان دهنده میزان تقارن یا عدم تقارن توزیع داده ها است. اگر داده‌ها نسبت به میانگین متقارن باشند، چولگی برابر صفر خواهد بود.
- به طور کلی سه نوع چولگی داریم:

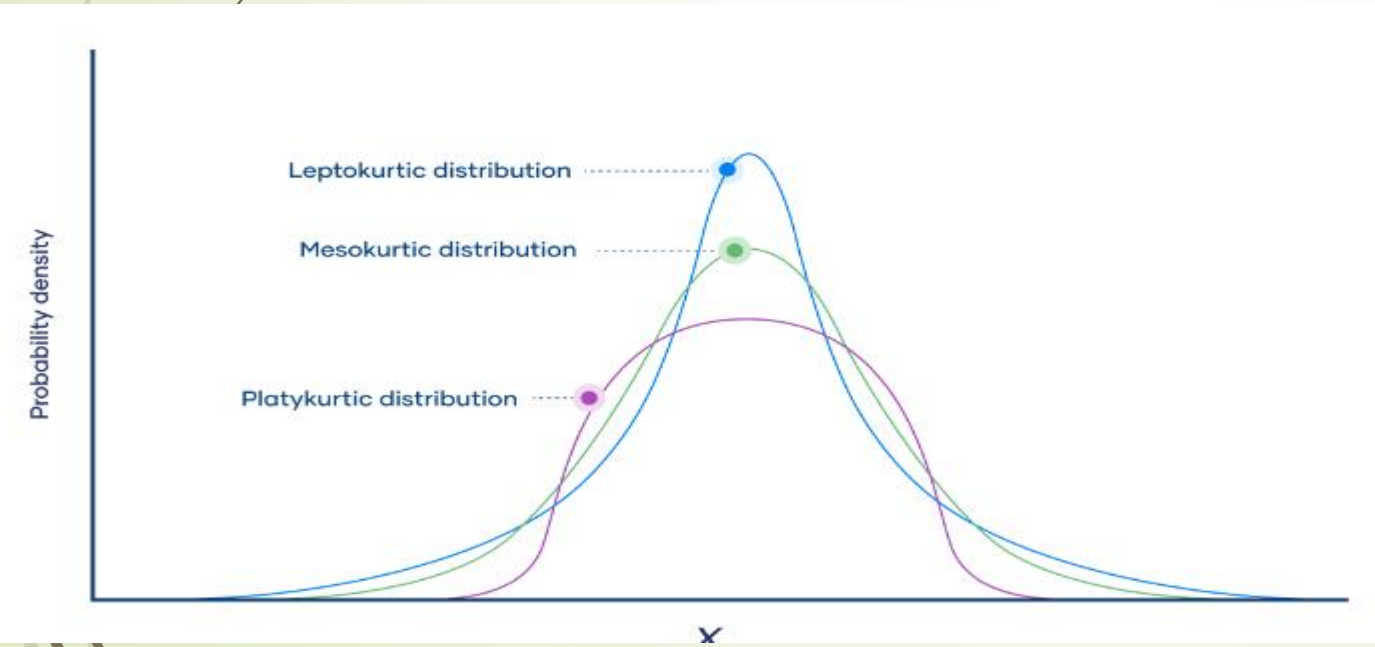
- متقارن (symmetrical): زمانی که چولگی نزدیک به صفر است و میانگین و میانه تقریبا برابر هستند.
- چولگی منفی (negative skew): زمانی که دنباله (tail) چپ نمودار هیستوگرام ، طولانی تر است و اکثر مشاهدات در دنباله راست متمرکز شده است. در این حالت میانه بزرگتر از میانگین است.
- چولگی مثبت (positive skew): زمانی که دنباله (tail)راست نمودار هیستوگرام ، طولانی تر است و اکثر مشاهدات در دنباله چپ متمرکز شده است. در این حالت میانه کوچکتر از میانگین است.



- تفسیر مقادیر چولگی:
- متقارن (symmetric): مقادیر بین ۰.۵ - تا ۰.۵
- داده با چولگی متوسط: مقادیر بین ۰.۵ - و۱ - یا بین ۰.۵ و ۱
- داده با چولگی بالا: مقادیر کمتر از -۱ و بزرگتر از ۱

کشیدگی Kurtosis داده ها

- کشیدگی توصیف کننده میزان قله‌ای بودن و مسطح بودن توزیع داده ها است.
- سه نوع کشیدگی وجود دارد:
  - Mesokurtic: به حالتی اشاره دارد که در آن منحنی داده‌ها نه پهن‌تر و نه باریک‌تر از توزیع نرمال است. میزان کشیدگی توزیع نرمال برابر ۳ می باشد
  - لپتوکورتیک (Leptokurtic): این توزیع دارای دنباله های عریض تر و طولانی تر و قله تیزتر است. در این حالت مقدار کشیدگی مثبت و مقدار آن از ۳ بزرگتر است.
- پلاتیکورتیک (Platykurtic): توزیع دارای قله پهن تر و پایین تر و دنباله های باریک تر و کوتاهتر است. در این حالت کشیدگی منفی و مقدار کمتر از ۳ است



برای بررسی نرمال بودن داده‌ها باید میزان چولگی و کشیدگی را بر خطای استاندارد مربوطه تقسیم کرد. اگر عدد حاصل، در بازه [۲ و ۲-] قرار گیرد، می‌توان توزیع داده‌ها را نرمال در نظر گرفت.

تشخیص تقریبی چولگی با استفاده از چارک ها

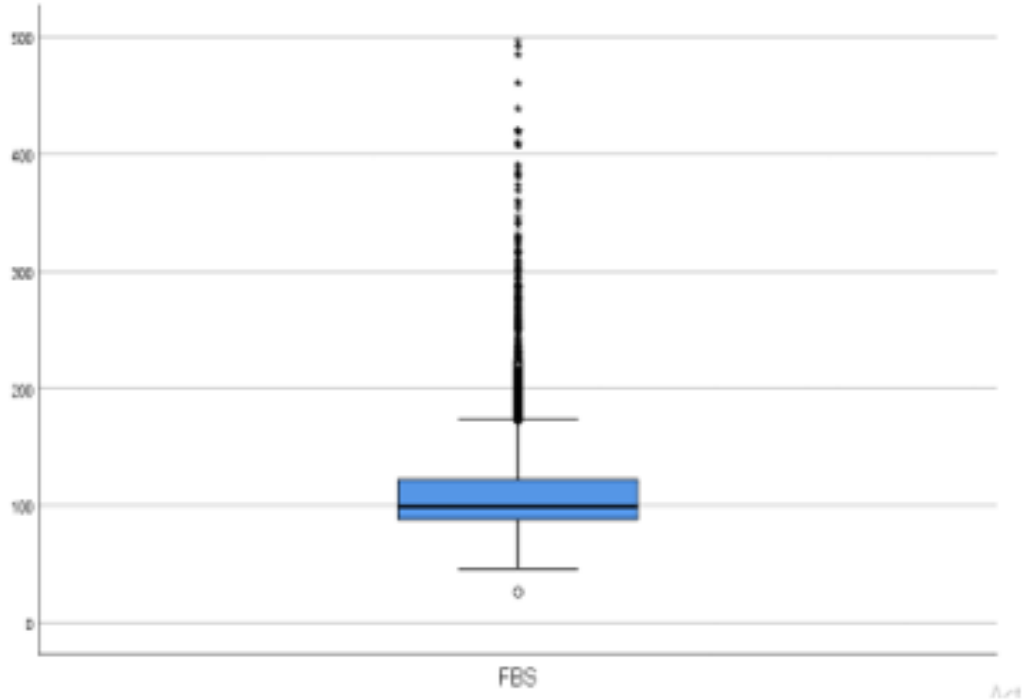
نشان دادن کجی منحنی با استفاده از چارکها

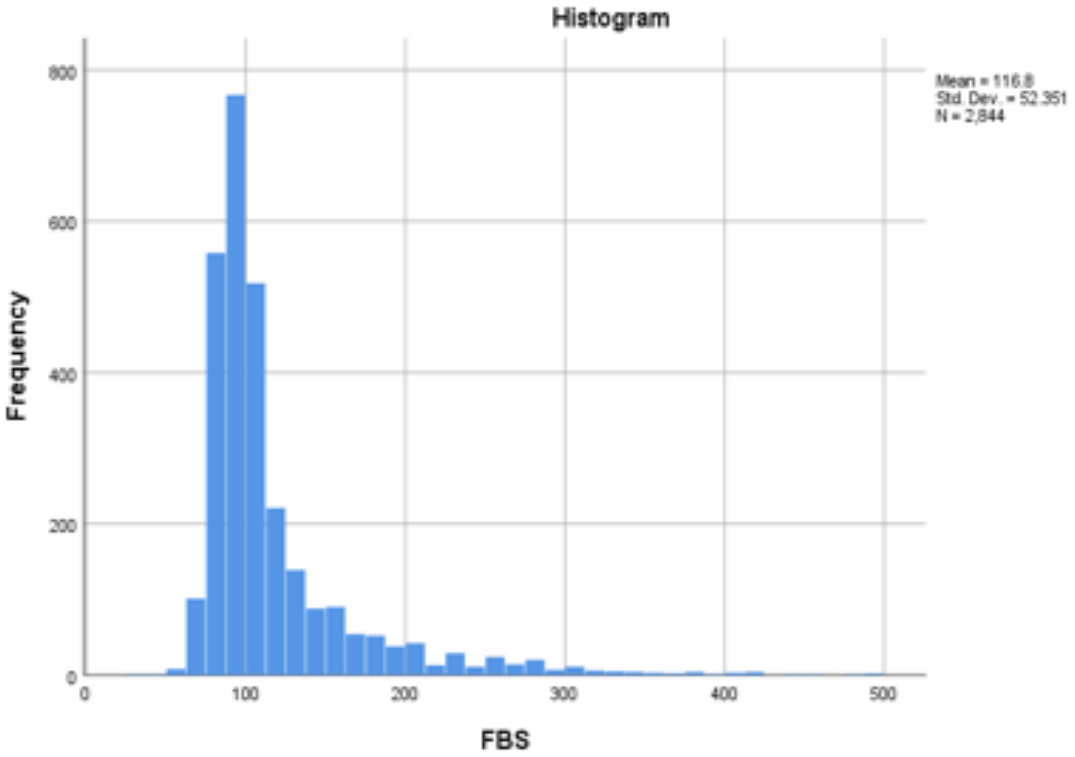
$Q_3 - Q_2 > Q_2 - Q_1$ = کجی مثبت

$Q_3 - Q_2 = Q_2 - Q_1$ = متقارن

$Q_3 - Q_2 < Q_2 - Q_1$ = کجی منفی

انواع خمیدگی با توجه به فواصل بین چارک‌ها





Statistics

FBS		
N	Valid	2844
	Missing	998
Percentiles	25	88.00
	50	99.00
	75	122.00

Descriptives

		Statistic	Std. Error
FBS	Mean	116.80	.982
	Median	99.00	
	Skewness	2.824	.046
	Kurtosis	10.215	.092

- علم آمار مشتمل بر
- روشهای جمع آوری داده (Data collection)
- خلاصه سازی اطلاعات (آمار توصیفی Descriptive statistics)
- **تجزیه و تحلیل داده ها (آمار استنباطی Inferential statistics)**
- استخراج و تفسیر نتایج از داده ها (Interpretation of results)

فرآیند تفسیر و معنا بخشیدن به یافته های بدست آمده

آمار استنباطی (گذار از آمار توصیفی)

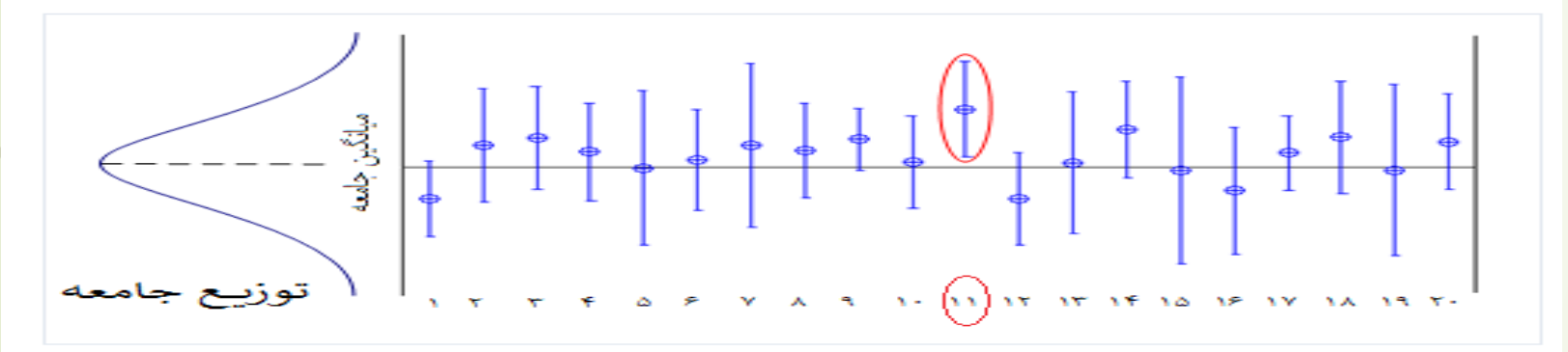
- چگونگی امکان دستیابی به یک تصمیم (رد یا پذیرش یک ادعا) در مورد مجموعه بزرگی از داده ها با بررسی بخش کوچکی از آن آمار استنباطی گویند.
- آمار استنباطی متکی بر علم احتمالات است.
- **براوردهای آماری و آزمونهای آماری (آزمون فرضیه)** دو مبحث مهم در آمار استنباطی محسوب می شود .
- اگر هدف از تحلیل آماری، برآورد پارامتر جامعه باشد بطوری که دقت برآورد نیز قابل محاسبه باشد از براوردهای آماری استفاده میکنیم
- براوردهای آماری می تواند نقطه ای یا فاصله ای (فاصله اطمینان) باشد.
- براورد نقطه ای یعنی شاخص آماری مورد نظر را از داده های نمونه محاسبه کرده و آنرا به عنوان تقریبی از پارامتر مجهول جامعه در نظر بگیریم. منظور از پارامتر جامعه، مقداری نامشخص ولی ثابت است که براساس نمونه تصادفی قصد برآورد آن را داریم.

فاصله اطمینان (Confidence Interval)

- هدف در محاسبه فاصله اطمینان، بدست آوردن حدودی برای پارامتر از روی اطلاعات نمونه است که در سطح اطمینان تعیین شده، شامل پارامتر مجهول جامعه باشد.
- براورد فاصله ای (یا فاصله اطمینان) با مفهوم مربوط به سطح اطمینان (Confidence Level) ارتباط نزدیکی دارد. منظور از سطح اطمینان، تعیین میزان شک یا یقینی است که نسبت به در برگیری پارامتر توسط فاصله اطمینان مربوطه، داریم.
- با استخراج نمونه های مختلف از جامعه، محاسبات برای این محدوده تغییر کرده و یک فاصله اطمینان متفاوت بدست خواهد آمد. به همین علت با سطح اطمینان مشخص، خانواده‌ای از فاصله‌های اطمینان با توجه نمونه‌های مختلف تولید می‌شود.
- به این ترتیب می‌توان گفت سطح اطمینان، فراوانی نسبی، فاصله‌های اطمینانی است که شامل پارامتر مجهول جامعه هستند.
- به بیان دیگر، اگر n فاصله اطمینان با سطح اطمینان ثابت ایجاد کنیم، نسبت آن‌هایی که شامل پارامتر مجهول هستند به کل فاصله‌ها، برابر با همان سطح اطمینان خواهد بود.

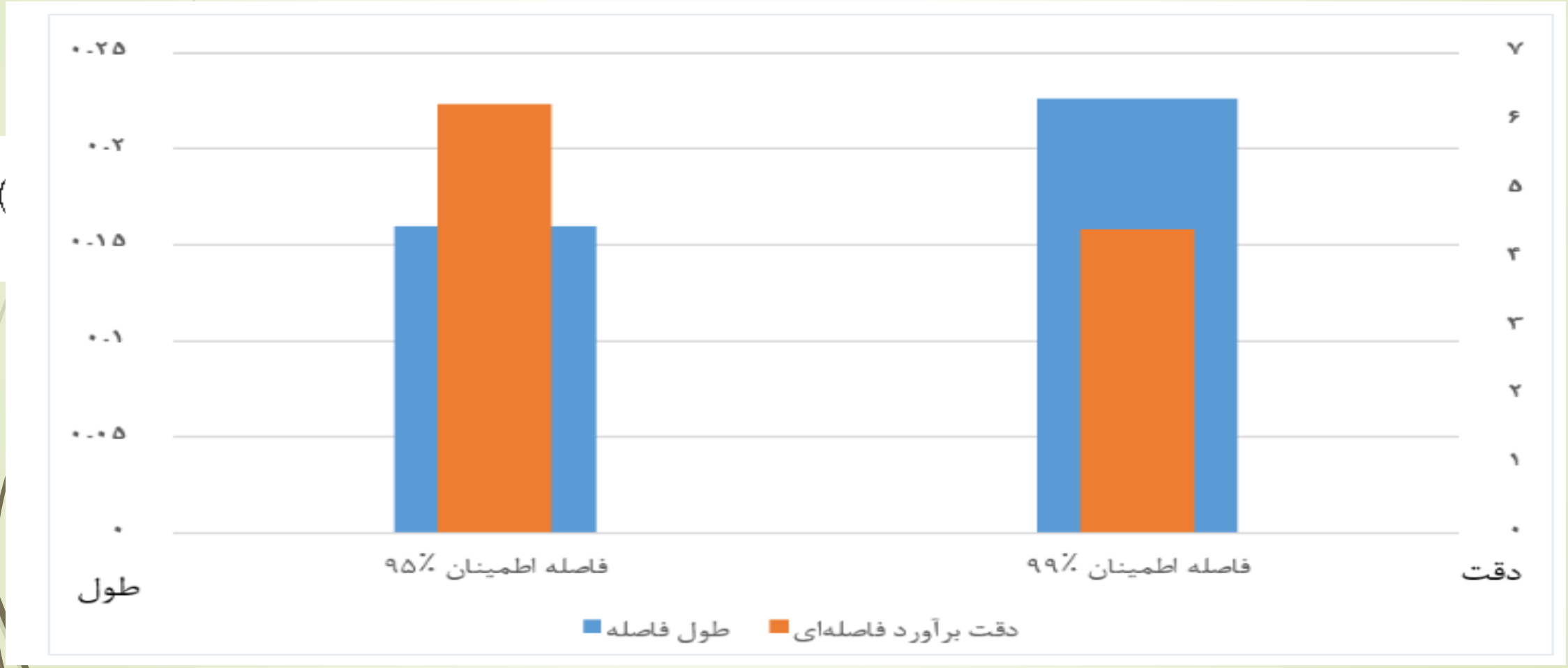
فاصله اطمینان (Confidence Interval)

- فاصله اطمینان بیان می‌کند که اگر مثلا ۱۰۰ بار نمونه‌گیری تکرار شود و ۱۰۰ فاصله اطمینان ٪۹۵ تولید شود، ۹۵ فاصله، در برگیرنده پارامتر جامعه خواهند بود و فقط ۵ تا از این فاصله‌ها شامل میانگین جامعه نمی‌شوند.
- یعنی در ٪۵ مواقع به فاصله اطمینانی برخورد کنیم که دربرگیرنده پارامتر جامعه نباشد و میزان خطا برابر با ٪۵ است



- از آنجا که پارامتری که قصد داریم برای آن یک فاصله اطمینان ایجاد کنیم، (مثلا میانگین جامعه)، نامعلوم است در نتیجه هرگز نمی‌توانیم پی‌ببریم که فاصله اطمینان بدست آمده حتما شامل پارامتر است یا خیر.
- سطح اطمینان را با $1-\alpha$ نشان می‌دهند که در آن α همان احتمال خطای نوع اول در آزمون فرض آماری است.

فاصله اطمینان (Confidence Interval) برای میانگین جامعه



آزمون فرض آماری (Statistical Hypothesis Testing)

- آزمونهای آماری یعنی با آزمایش و بررسی نمونه ای از جامعه مورد نظر، در مورد آن جمعیت به رد یا پذیرش یک ادعا می پردازیم.
- ولی مشکلی که در این میان وجود دارد، کمبود اطلاعاتی است که از جامعه آماری داریم، زیرا نمونه گرفته شده از جامعه آماری همه اطلاعات آن را منعکس نمی‌کند و نمونه فقط می‌تواند بخشی از اطلاعات جامعه آماری را به ما نشان دهد. بنابراین آزمون‌ها همواره دارای درجه‌ای از خطا هستند.تصمیمات و نتایج گرفته شده از آزمون فرض آماری، همراه با یک خطای قابل تحمل به نام احتمال خطای نوع اول Error Type I که از قبل مشخص شده، صورت می‌گیرد که اغلب آن را با α نشان می‌دهیم .α یعنی احتمال رد فرض صفر در شرایطی که فرض صفر درست باشد.
- یک آزمون فرض آماری از دو فرضیه به نام‌های فرض صفر Null Hypothesis (معمولا نظری است که از قبل در مورد پارامتر وجود داشته) و فرض مقابل Alternative Hypothesis (ادعای جدید مطرح شده توسط محقق) تشکیل شده که به صورت یک گزاره در مورد پارامتر جامعه بیان می‌شوند.
- هدف از انجام آزمون فرض آماری، مشخص کردن یک تصمیم است که منجر به رد یا پذیرش فرض صفر شود.
- برای اینکه مشخص شود که آیا براساس نمونه می‌توان فرض صفر را رد کرد باید ملاکی در نظر گرفته شود که به آن آماره آزمون می‌گویند. به علاوه، با توجه به توزیع آماره آزمون، یک ناحیه بحرانی ساخته می‌شود. اگر مقدار آماره آزمون با توجه به نمونه تصادفی در ناحیه بحرانی قرار بگیرد، فرض صفر رد می‌شود. در غیر اینصورت دلیلی برای رد کردن فرض صفر وجود ندارد.

آزمون فرض آماری (Statistical Hypothesis Testing)

- چند مثال از تنظیم فرضیه ها، آماره آزمون و ناحیه بحرانی

فرض صفر	فرض مقابل	آماره آزمون	ناحیه بحرانی
$\mu = \mu_0$	$\mu = \mu_1, \quad \mu_1 > \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$T > t_{(1-\alpha)}(n-1)$
$\mu = \mu_0$	$\mu < \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$T < -t_{(1-\alpha)}(n-1)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$\lvert T \rvert > t_{(1-\frac{\alpha}{2})}(n-1)$

آزمون فرض آماری (Statistical Hypothesis Testing)

- مثال:

جامعه آماری، از پسرانی که در محدوده سنی 10 تا 12 سال هستند تشکیل شده است. اطلاعات قبلی نشان می‌دهد که متوسط قد این افراد برابر است با 75 سانتی‌متر با انحراف استاندارد 11.6 است. با توجه به تغییر شیوه تغذیه اعتقاد داریم که میانگین قد پسرها در جامعه افزایش داشته و به 80 سانتی‌متر رسیده است. براساس یک نمونه ۲۵ تایی میانگین قدها برابر با 80.94 سانتی‌متر بدست آمده است. آیا می‌توان از اطلاعات قبلی در مورد قد اطمینان داشت یا می‌توان به کمک آزمون آماری نشان داد که تغییر محسوسی در میزان قد پسران رخ داده است؟ فرضیات این آزمون به صورت زیر نوشته می‌شود:

$$\begin{cases} H_0 : \mu = 75 \\ H_1 : \mu = 80 \end{cases}$$

فرضیه صفر، نتایج و نظراتی که از قبل وجود داشته را بیان کرده ولی فرض مقابل نظر محقق را نشان می‌دهد. سعی داریم به کمک نمونه تصادفی تهیه شده، در مورد فرض صفر قضاوت کنیم. با توجه به آماره آزمون معرفی شده و ناحیه بحرانی آزمون فرض را انجام می‌دهیم.

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{80.94 - 75}{\frac{11.6}{\sqrt{25}}} = 2.56$$

اگر میزان احتمال خطای نوع اول را $\alpha = 0.05$ در نظر بگیریم، خواهیم داشت: $Z > z_{(1-\alpha)} = 2.56 > 1.64$

در نتیجه با خطای 0.05، رای به رد فرض H_0 می‌دهیم.

آزمون فرض آماری (Statistical Hypothesis Testing)

- در بیشتر نرم‌افزارهای آماری برای سهولت در تصمیم‌گیری نسبت به نتیجه آزمون فرض آماری، شاخصی به نام مقدار احتمال P-value ارائه می‌شود. این مقدار به محقق کمک می‌کند که بدون آگاهی از فرمول ها و محاسبه آمار آزمون و ناحیه بحرانی، بتواند در مورد رد یا پذیرش فرض صفر تصمیم بگیرید.
- P-value بر اساس اطلاعات نمونه حاضر محاسبه می شود و نشان می‌دهد که آزمون مورد نظر، درصورت صحیح بودن فرض صفر، چقدر وجود چنین نمونه‌ای را محتمل می‌داند. قاعدتا هر چه P-value کوچکتر باشد یعنی انتظار نمی رود اگر فرض صفر صحیح بود چنین نمونه ای مشاهده نماییم. پس به نظر می رسد فرض صفر درست نباشد.
- به عبارت دیگر، P-value کوچکترین مقدار خطای مجاز محقق (خطای نوع اول α) است برای اینکه با اطلاعات نمونه حاضر بتواند به رد فرض صفر مبادرت ورزد.
- بنا بر یک قاعده مشخص، در آزمون های آماری، اگر مقدار P-value کمتر از خطای نوع اول $\alpha=0.05$ باشد فرض صفر را رد می کنیم و اگر بزرگتر باشد نمونه حاضر شواهدی برای رد فرض صفر ارائه نمی کند

- علم آمار مشتمل بر
- روشهای جمع آوری داده (Data collection)

(استخراج نمونه از جامعه مورد نظر طبق ضوابطی مقبول بطوریکه نمونه حاضر نماینده خوبی برای جامعه باشد)

- خلاصه سازی اطلاعات (آمار توصیفی Descriptive statistics)

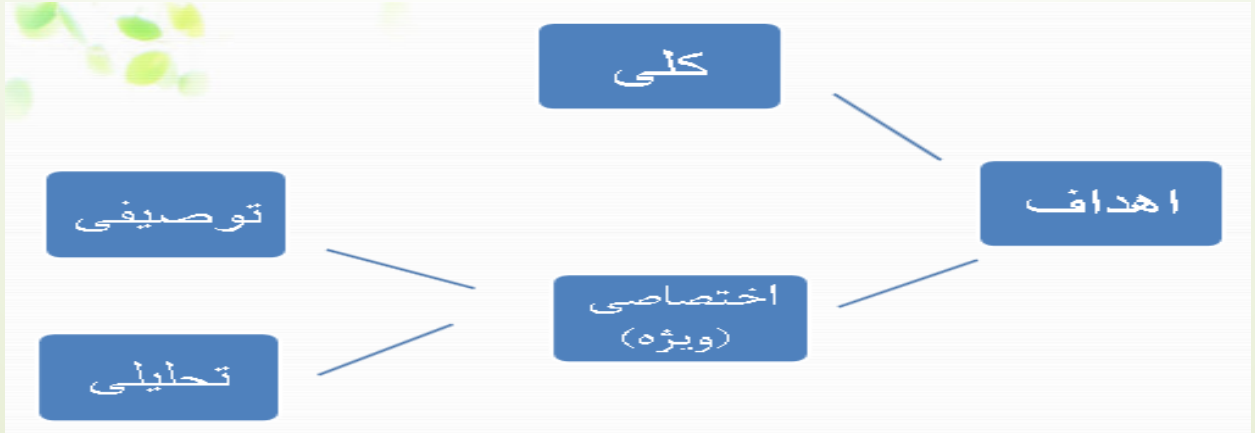
(روشهایی برای سازماندهی و خلاصه کردن خصوصیات مهم مجموعه داده ها، متناسب با نوع داده)

- **تجزیه و تحلیل داده ها (آمار استنباطی Inferential statistics)**

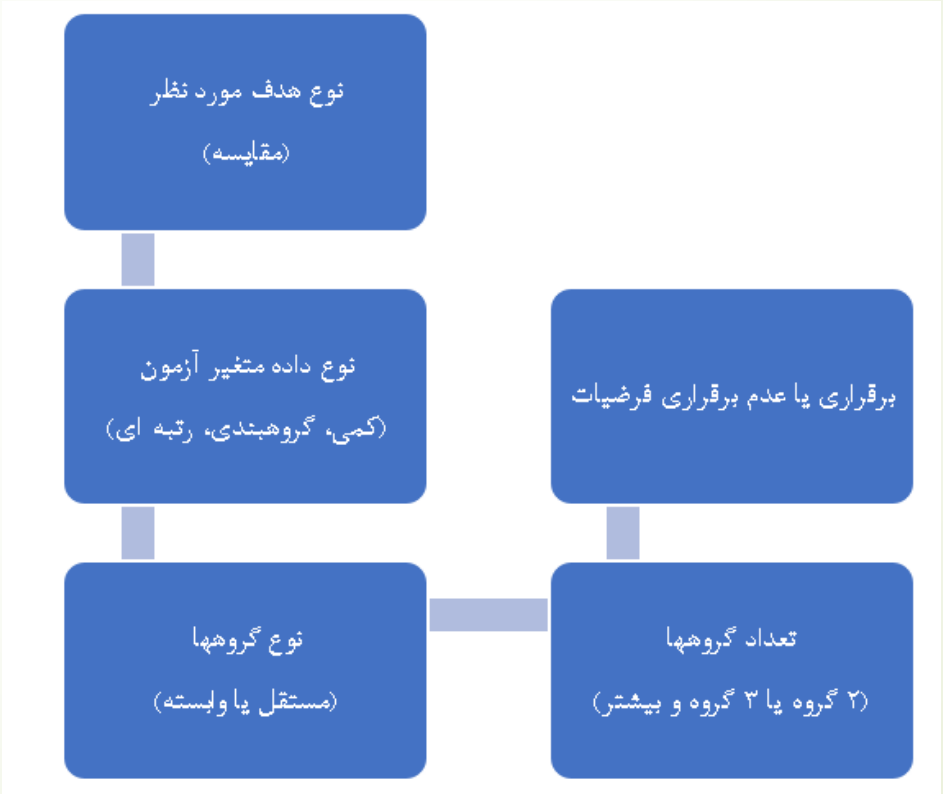
(استفاده از اطلاعاتی که یک نمونه از جامعه آماری در اختیارمان می‌گذارد تا رسیدن به نتیجه ای در مورد این جامعه، مراحل یک تحلیل آماری هستند)

- استخراج و تفسیر نتایج از داده ها (Interpretation of results)

(فرآیند تفسیر و معنا بخشیدن به یافته های بدست آمده)

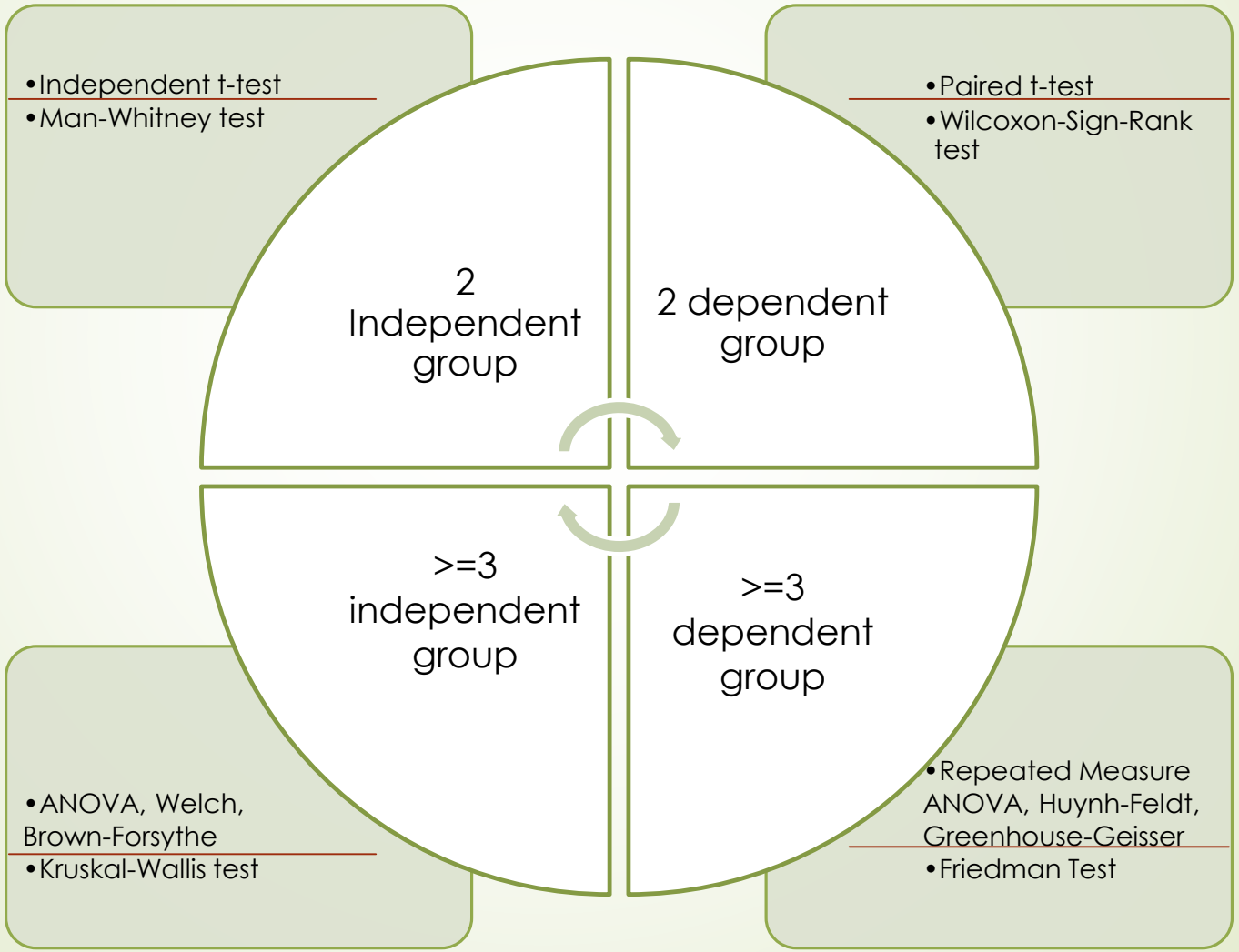


انتخاب آزمون آماری مناسب برای اهداف مربوط به اختلاف معنی داری (اهداف مقایسه ای)



انتخاب آزمون آماری مناسب برای اهداف مربوط به اختلاف معنی داری (اهداف مقایسه ای): برای متغیرهای کمی

Continuous variable



- **Independent t-test (Mann-Whitney, if required):**

Group Statistics

	SEX	N	Mean	Std. Deviation	Std. Error Mean
Body Mass Index	FEMALE	4636	26.53459	4.90733743	.0720733
	MALE	4665	24.80901	4.11837260	.0602976

$H_0: \mu_1 = \mu_2$, $H_1: \mu_1 \neq \mu_2$

$H_0: \mu_1 - \mu_2 = 0$, $H_1: \mu_1 - \mu_2 \neq 0$

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Body Mass Index	Equal variances assumed	120.321	.000	18.373	9299	.000	1.725578	.09391913	1.541476	1.909680
	Equal variances not assumed			18.363	9008.290	**.000**	1.725578	.09396996	1.541376	1.909781

	Female N=4636	Male N=4665	p-value
BMI	26.53±4.91	24.81±4.12	<0.001

فرضیات آزمون t-test:
برقراری توزیع نرمال در گروهها
برقراری فرض همگنی واریانس

- **Paired t-test (Wilcoxon sign rank test, if required):**

فرضیات آزمون paired t-test:
برقراری توزیع نرمال

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	bmi1	27.1622	736	3.90773	.14404
	bmi2	26.6077	736	3.81147	.14049

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	bmi1 - bmi2	.55449	1.03278	.03807	.47976	.62923	14.566	735	.000

	Before	After	Difference	p-value
BMI	27.16±3.91	26.61±3.81	0.55±1.03	<0.001

Analysis Of Variance (ANOVA) (Kruskal-Wallis test, if required) :

فرضیات آزمونANOVA:
برقراری توزیع نرمال در گروهها
برقراری همگنی واریانس بین گروهها

Descriptives

bmi

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
<8year	2319	28.5431	5.02926	.10444	28.3383	28.7479	14.46	54.42
9-12y	632	27.7424	4.55019	.18100	27.3870	28.0978	16.19	43.56
>12y	348	27.7227	4.06614	.21797	27.2940	28.1514	14.67	42.24
Total	3299	28.3032	4.85950	.08461	28.1373	28.4690	14.46	54.42

ANOVA

bmi

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	449.486	2	224.743	9.567	.000
Within Groups	77431.827	3296	23.493		
Total	77881.313	3298			

	0 – 5 year	6 – 12 year	>12 year	P-value
BMI	26.60±4.71^{bc}	25.05±4.46^{ac}	24.35±4.07^{ab}	<0.001

Robust Tests of Equality of Means

bmi

	Statistic[a]	df1	df2	Sig.
Welch	10.788	2	850.747	.000
Brown-Forsythe	11.513	2	1387.298	.000

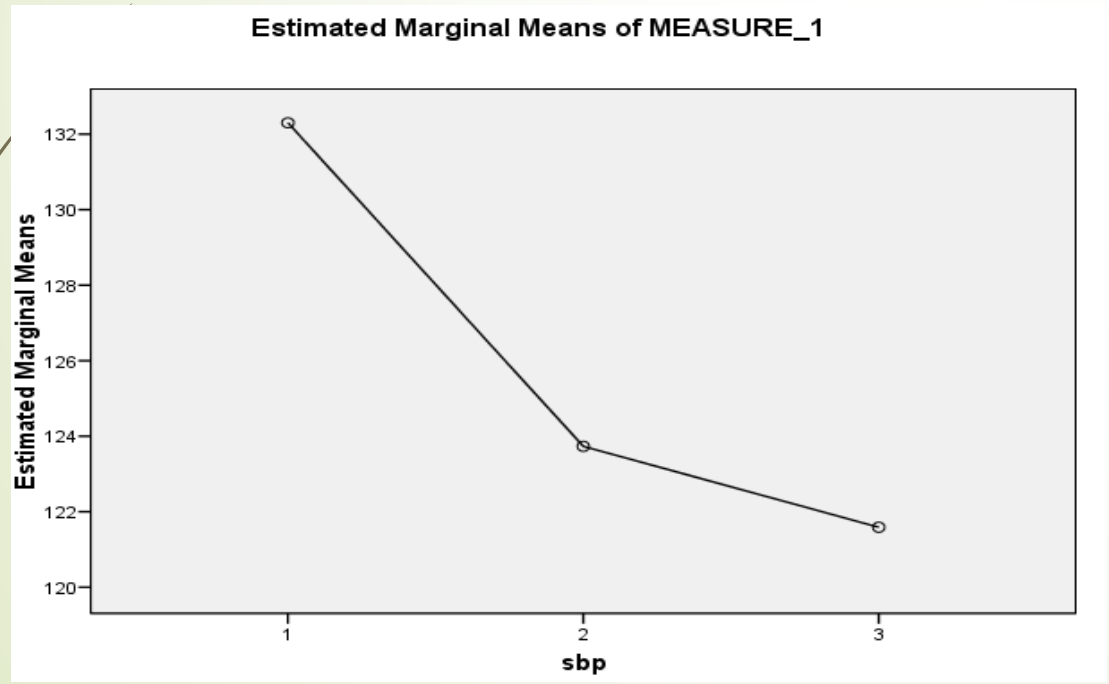
a. Asymptotically F distributed.

- One sample Repeated Measured ANOVA (Friedman Test) :

Descriptive Statistics

	Mean	Std. Deviation	N
SBP1	132.3016	18.17978	63
SBP3	123.7302	16.58235	63
SBP5	121.5873	13.22343	63

	Time 1	Time 2	Time 3	P-value
SBP	132.30±18.18	123.73±16.58	121.59±13.22	???



One sample Repeated Measured ANOVA (Friedman Test) :

- فرضیات آزمون
برقراری توزیع نرمال
فرض کرویت Sphericity test

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.
sbp	Pillai's Trace	.410	21.232[a]	2.000	61.000	.000
	Wilks' Lambda	.590	21.232[a]	2.000	61.000	.000
	Hotelling's Trace	.696	21.232[a]	2.000	61.000	.000
	Roy's Largest Root	.696	21.232[a]	2.000	61.000	.000

a. Exact statistic

b.
Design: Intercept
Within Subjects Design: sbp

Mauchly's Test of Sphericity

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
sbp	1.000	.000	2	1.000	1.000	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

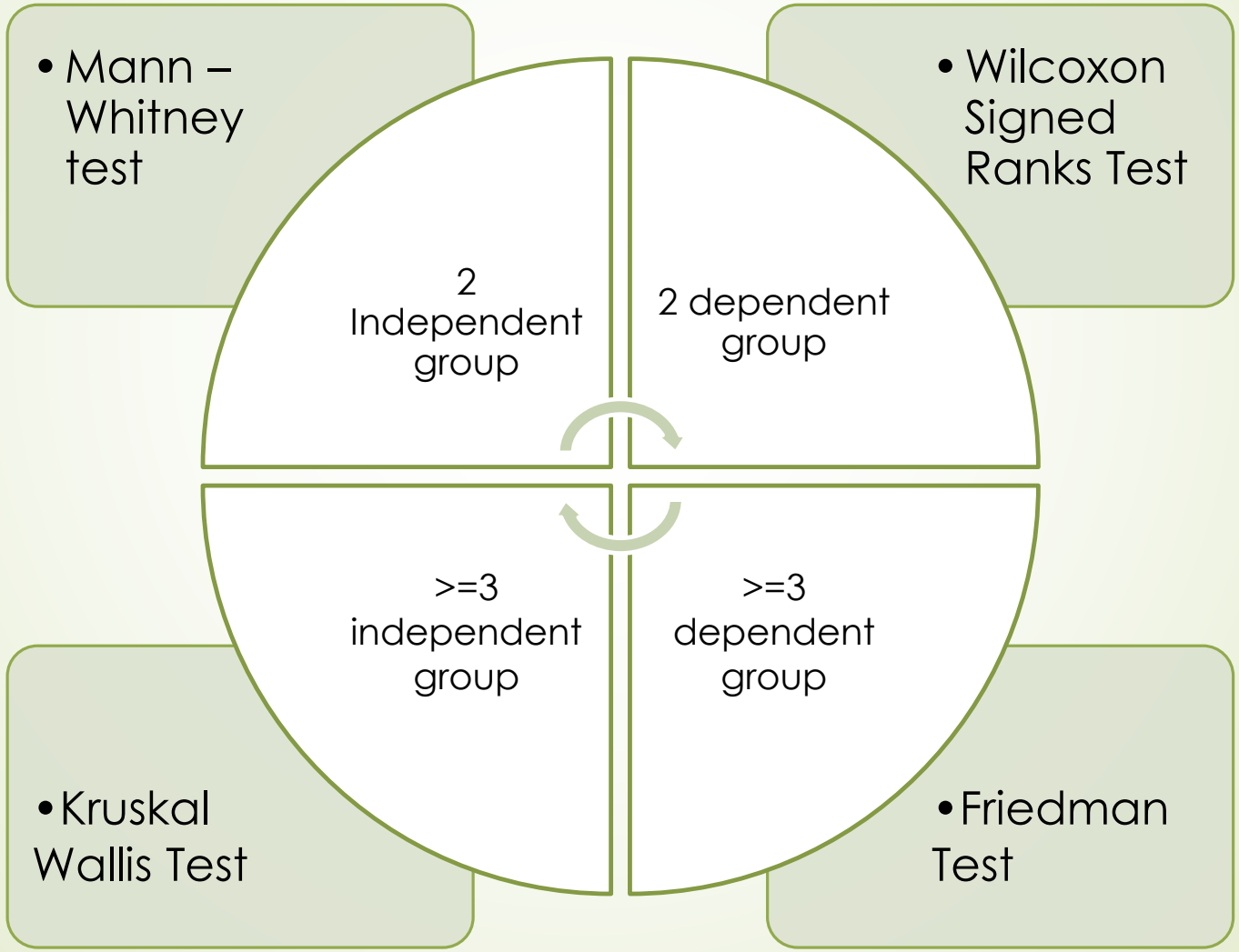
Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
sbp	Sphericity Assumed	4050.000	2	2025.000	21.585	.000
	Greenhouse-Geisser	4050.000	2.000	2025.002	21.585	.000
	Huynh-Feldt	4050.000	2.000	2025.000	21.585	.000
	Lower-bound	4050.000	1.000	4050.000	21.585	.000

انتخاب آزمون آماری مناسب برای اهداف مربوط به اختلاف معنی داری (اهداف مقایسه ای): برای متغیرهای رتبه ای

Ordinal variable



- Mann-Whitney test :

Descriptive Statistics

SEX		N	Mean	Std. Deviation	Minimum	Maximum	Percentiles 25th	50th (Median)	75th
FEMALE	Body Mass Index	4636	26.5346	4.907337	15.0356	47.6299	22.99184	26.2486836	29.60651
MALE	Body Mass Index	4665	24.8090	4.118373	15.2249	54.0123	21.73651	24.5351240	27.44257

Test Statistics[a]

	Body Mass Index
Mann-Whitney U	8572406.0
Wilcoxon W	19455851
Z	-17.308
Asymp. Sig. (2-tailed)	**.000**

a. Grouping Variable: SEX

	Female	Male	p-value
BMI	26.25 (22.99 , 29.61)	24.53 (21.74 , 27.44)	<0.001

Median (inter quartile range IQR)

- Wilcoxon signed rank test :

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
bmi1	803	27.2595	3.94531	15.77	43.15	24.5268	26.7229	29.6158
bmi2	736	26.6077	3.81147	15.77	43.57	24.0101	26.0731	28.8970

Test Statistics[b]

	bmi2 - bmi1
Z	-13.803[a]
Asymp. Sig. (2-tailed)	.000

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

	Before	After	p-value
BMI	26.72 (24.53 , 29.62)	26.07 (24.01 , 28.89)	<0.001

Median (inter quartile range IQR)

- Kruskal Wallis Test:

Descriptive Statistics

education		N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
							25th	50th (Median)	75th
0 - 5 year	Body Mass Index	4212	26.60261	4.70977161	15.04748	54.01235	23.14748	26.3039595	29.54192
6 - 12 year	Body Mass Index	3875	25.05305	4.45657974	15.03558	47.62993	21.79931	24.7474543	27.73230
>12 year	Body Mass Index	1197	24.35179	4.06507419	15.65454	38.15967	21.33821	24.1516725	26.97473

Test Statistics[a,b]

	Body Mass Index
Chi-Square	18.827
df	1
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: education

	0 – 5 year	6 – 12 year	>12 year	P-value
BMI	26.30 (23.15 , 29.54)	24.75 (21.79 , 27.73)	24.15(21.34 , 26.97)	<0.001

- Friedman test :

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
BMI1	70	30.2076	3.79081	23.58	40.57	27.8091	30.0083	32.3995
BMI2	70	306652.8	38554.24650	235846.7	409752.9	280019.5	305529.0383	326946.5
BMI3	70	30.8714	4.13511	24.00	43.00	27.0000	31.0000	34.0000

Test Statistics[a]

N	70
Chi-Square	110.600
df	2
Asymp. Sig.	.000

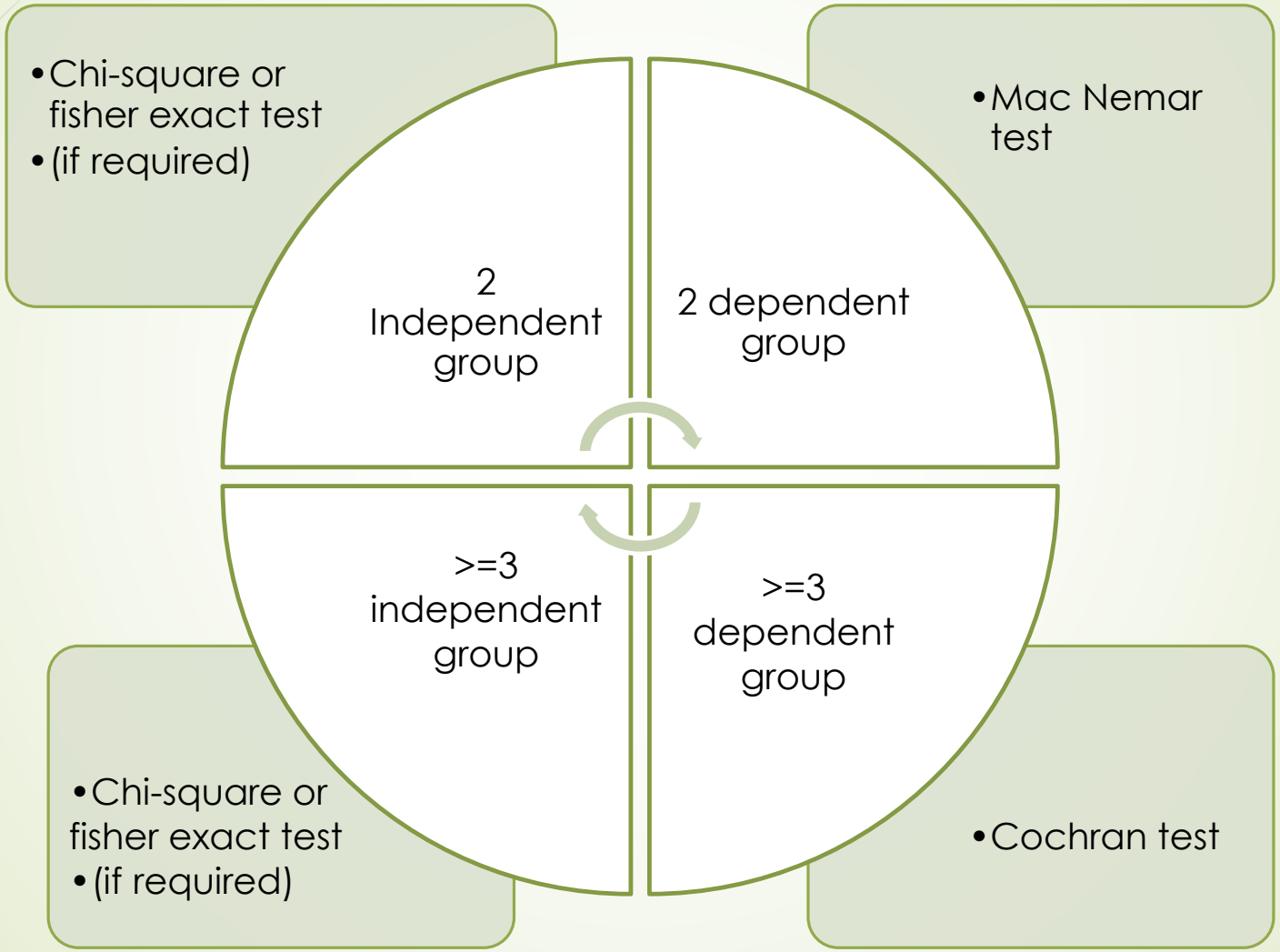
a. Friedman Test

	Time 1	Time 2	Time 3	P-value
BMI	30.01 (27.81 , 32.39)	30.55 (28.00 , 32.69)	31.00 (27.00 , 34.00)	<0.001

Median (inter quartile range)

انتخاب آزمون آماری مناسب برای اهداف مربوط به اختلاف معنی داری (اهداف مقایسه ای): برای متغیرهای گروهبندی شده

Nominal variable



Chi- square test :

metabolic syndrome * SEX Crosstabulation

			SEX		
			FEMALE	MALE	Total
metabolic syndrome	.00	Count	3453	4058	7511
		% within metabolic syndrome	46.0%	54.0%	100.0%
		% within SEX	72.9%	85.8%	79.3%
	1.00	Count	1286	669	1955
		% within metabolic syndrome	65.8%	34.2%	100.0%
		% within SEX	27.1%	14.2%	20.7%
Total		Count	4739	4727	9466
		% within metabolic syndrome	50.1%	49.9%	100.0%
		% within SEX	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	243.443[b]	1	**.000**		
Continuity Correction[a]	242.651	1	.000		
Likelihood Ratio	246.864	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	243.417	1	.000		
N of Valid Cases	9466				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 976.26.

	Female	Male	p-value
Metabolic syndrome	1286 (27.1)	669 (14.2)	<0.001

- Chi- square test :

SEX * education Crosstabulation

			education			Total
			0 - 5 year	6 - 12 year	>12 year	
SEX	FEMALE	Count	2588	1721	469	4778
		% within SEX	54.2%	36.0%	9.8%	100.0%
		% within education	60.2%	43.1%	37.1%	50.0%
		% of Total	27.1%	18.0%	4.9%	50.0%
	MALE	Count	1711	2271	795	4777
		% within SEX	35.8%	47.5%	16.6%	100.0%
		% within education	39.8%	56.9%	62.9%	50.0%
		% of Total	17.9%	23.8%	8.3%	50.0%
Total		Count	4299	3992	1264	9555
		% within SEX	45.0%	41.8%	13.2%	100.0%
		% within education	100.0%	100.0%	100.0%	100.0%
		% of Total	45.0%	41.8%	13.2%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	338.764[a]	2	.000
Likelihood Ratio	341.226	2	.000
Linear-by-Linear Association	314.481	1	.000
N of Valid Cases	9555		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 631.93.

Mc Nemar test :

- برای آزمون مک نمار نیاز است متغیر مورد بررسی یک متغیر دو رده ای باشد

bmi1_cat * bmi2_cat Crosstabulation

			bmi2_cat		
			.00	Obese	Total
bmi1_cat	.00	Count	622	**7**	629
		% within bmi1_cat	98.9%	1.1%	100.0%
		% within bmi2_cat	91.6%	5.6%	78.3%
		% of Total	77.5%	**.9%**	78.3%
	Obese	Count	**57**	117	174
		% within bmi1_cat	32.8%	67.2%	100.0%
		% within bmi2_cat	8.4%	94.4%	21.7%
		% of Total	**7.1%**	14.6%	21.7%
Total		Count	679	124	803
		% within bmi1_cat	84.6%	15.4%	100.0%
		% within bmi2_cat	100.0%	100.0%	100.0%
		% of Total	84.6%	15.4%	100.0%

	بدتر شده	بهبود یافته	p-value
BMI	7 (0.9 %)	57 (7.1%)	<0.001

Test Statistics[b]

	bmi1_cat & bmi2_cat
N	803
Chi-Square[a]	37.516
Asymp. Sig.	.000

a. Continuity Corrected

b. McNemar Test

	Time 1	Time 2	p-value
BMI category (obese)	174(21.7%)	124(15.4)	<0.001

Frequency (percent)

Cochran Test

- برای آزمون کوکران نیاز است متغیر مورد بررسی یک متغیر دو رده ای باشد

anxiety1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	98	39.2	39.2	39.2
	Clinical case	152	60.8	60.8	100.0
	Total	250	100.0	100.0	

anxiety2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	156	62.4	62.4	62.4
	Clinical case	94	37.6	37.6	100.0
	Total	250	100.0	100.0	

anxiety3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	158	63.2	63.2	63.2
	Clinical case	92	36.8	36.8	100.0
	Total	250	100.0	100.0	

Count

			anxiety3		
anxiety1			Normal	Clinical case	Total
Normal	anxiety2	Normal	38	8	46
		Clinical case	40	12	52
	Total		78	20	98
Clinical case	anxiety2	Normal	60	50	110
		Clinical case	20	22	42
	Total		80	72	152
Total	anxiety2	Normal	98	58	156
		Clinical case	60	34	94
	Total		158	92	250

Test Statistics

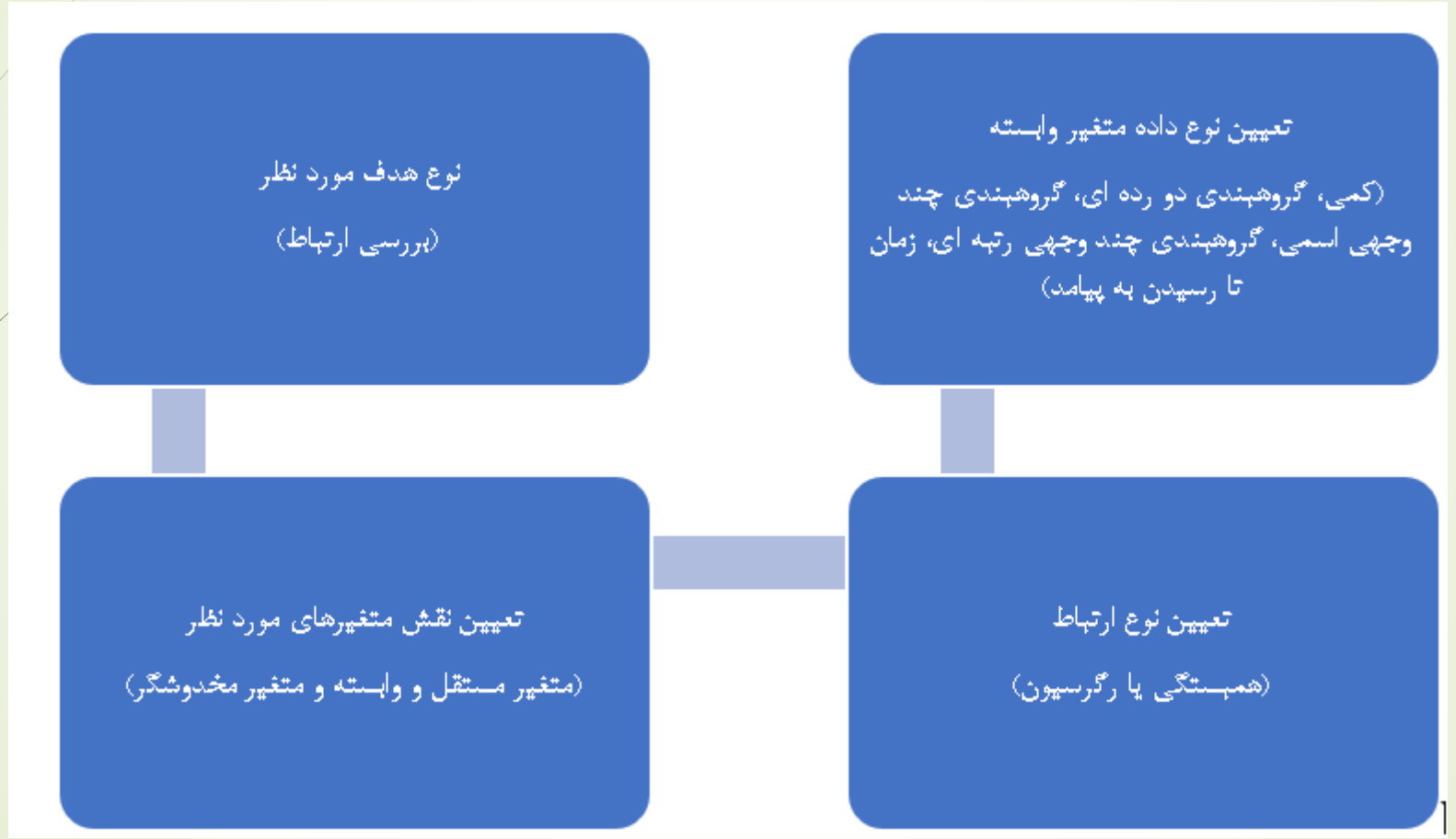
N	250
Cochran's Q	36.674[a]
df	2
Asymp. Sig.	.000

a. 0 is treated as a success.

	Time 1	Time 2	Time 3	p-value
Anxiety	152(60%)	94(37.65)	92(36.8%)	<0.001

Frequency (percent)

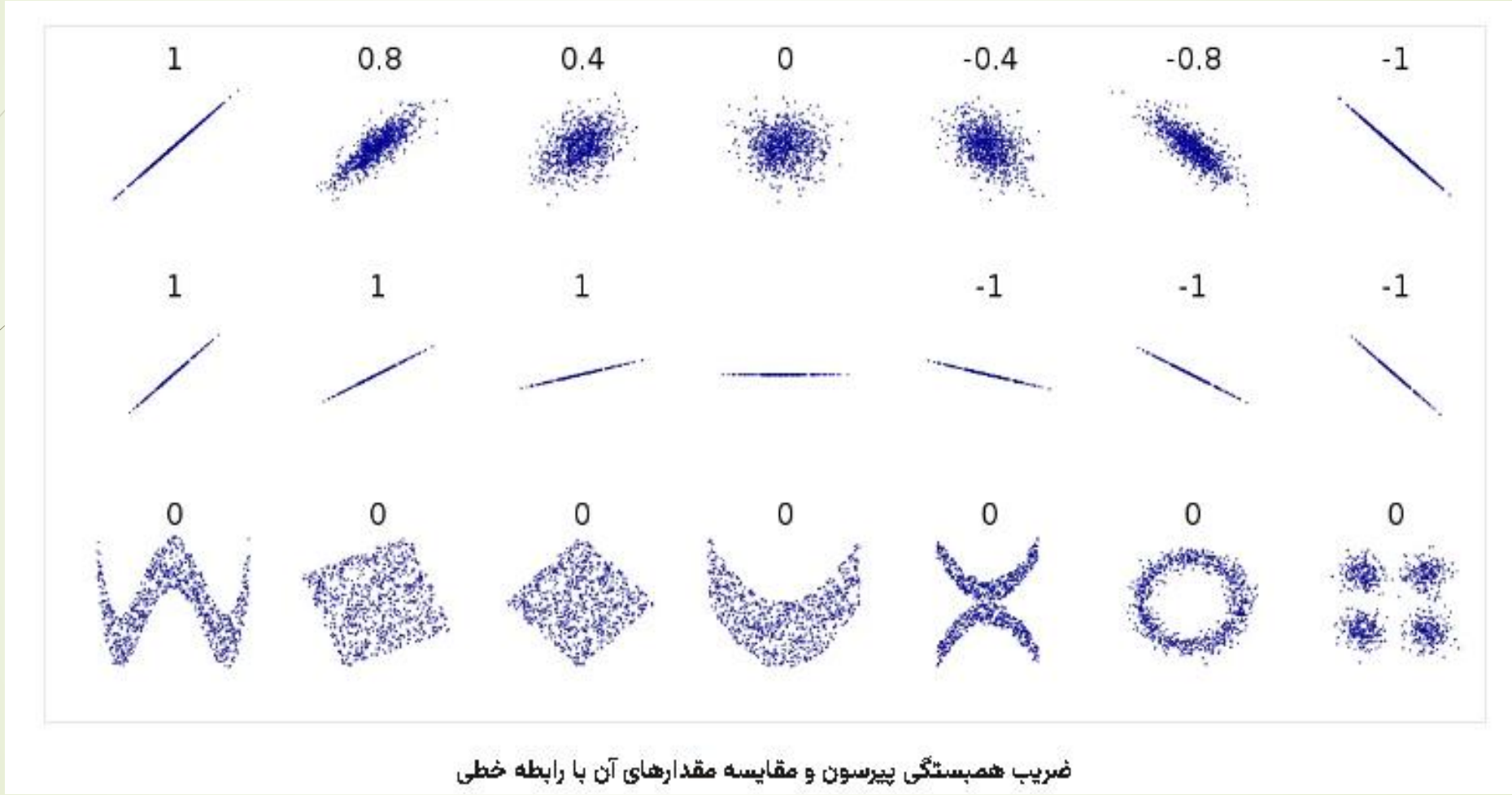
انتخاب آزمون آماری مناسب برای اهداف مربوط به بررسی ارتباط



انتخاب آزمون آماری مناسب برای اهداف مربوط به بررسی ارتباط: ضرایب همبستگی

- ضریب همبستگی پیرسن برای بررسی ارتباط دو متغیر کمی استفاده می گردد
- هر چه مقدار شاخص همبستگی بزرگتر باشد، نشان‌دهنده ارتباط بیشتر بین دو متغیر است.
- این ضریب ارتباط خطی بین دو متغیر را شناسایی می کند
- مقدار این ضریب بین -۱ تا ۱ می باشد صفر بودن به معنی عدم وجود ارتباط خطی دو متغیر است
- جهت ارتباط دو متغیر نیز مشخص می گردد یعنی اگر ارتباط مثبت باشد با افزایش یکی دیگری نیز تمایل به افزایش دارد. و بالعکس
- این شاخص به واحد اندازه گیری متغیرها بستگی ندارد
- این شاخص در مقابل داده‌های دورافتاده، منحرف شده و میزان همبستگی را به درستی نشان نمی دهد.
- اگر نوع متغیرهای مورد نظر رتبه ای باشند یا نوع ارتباط غیر خطی باشد یا داده پرت وجود داشته باشد یا نحوه توزیع متغیرهای کمی، نرمال نباشد می توان از ضریب همبستگی اسپیرمن استفاده نمود
- ضریب همبستگی کندال نیز جهت بررسی ارتباط دو متغیر رتبه ای کاربرد دارد.

انتخاب آزمون آماری مناسب برای اهداف مربوط به بررسی ارتباط: ضرایب همبستگی



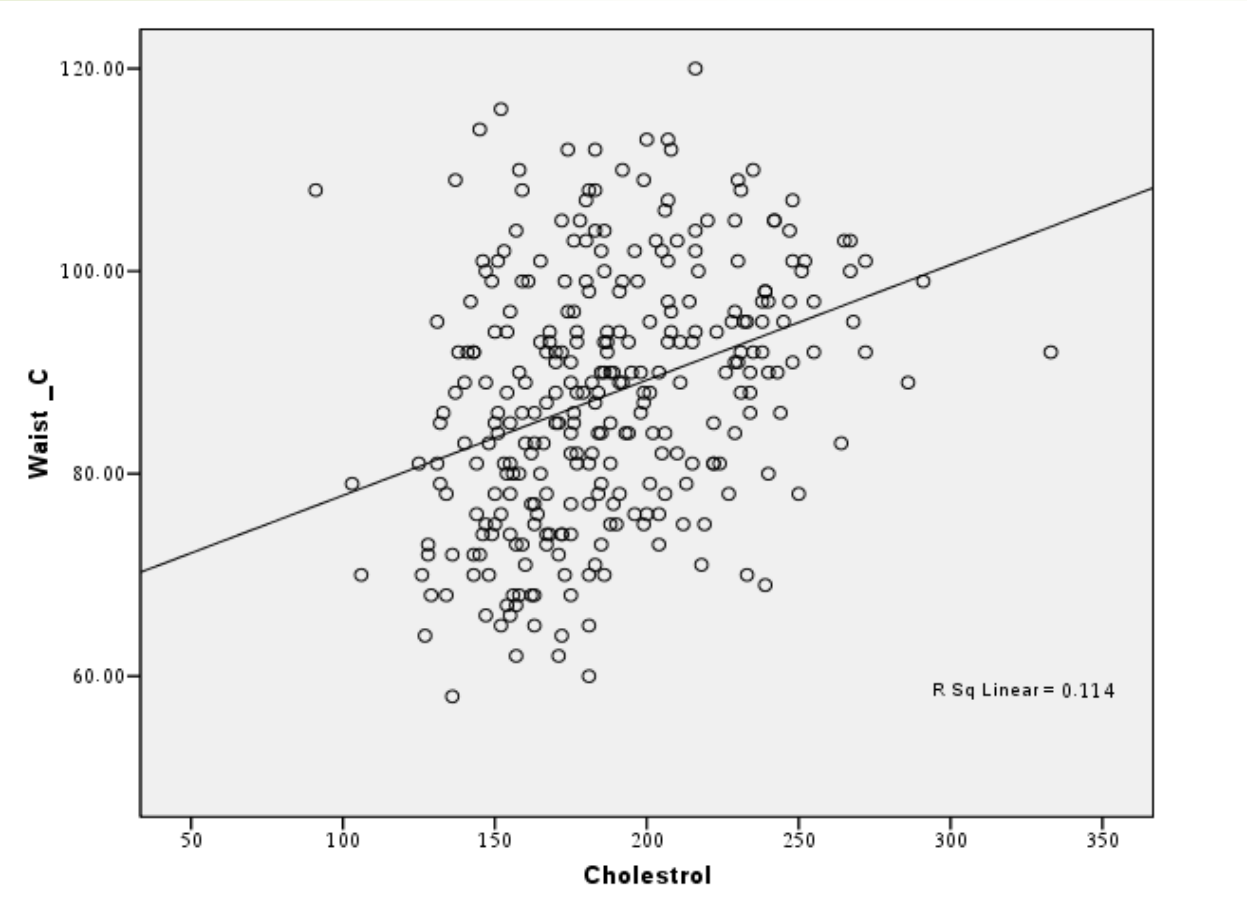
ضریب همبستگی پیرسون و مقایسه مقدارهای آن با رابطه خطی

Correlation :

Correlations

		Cholestrol	Triglycerides	HDL	LDL by Friedewald Formula	Systolic Blood Pressure	Diastolic Blood Pressure
Waist _C	Pearson Correlation	.338**	.363**	-.152**	.283**	.293**	.300**
	Sig. (2-tailed)	.000	.000	.009	.000	.000	.000
	N	299	299	298	299	303	303

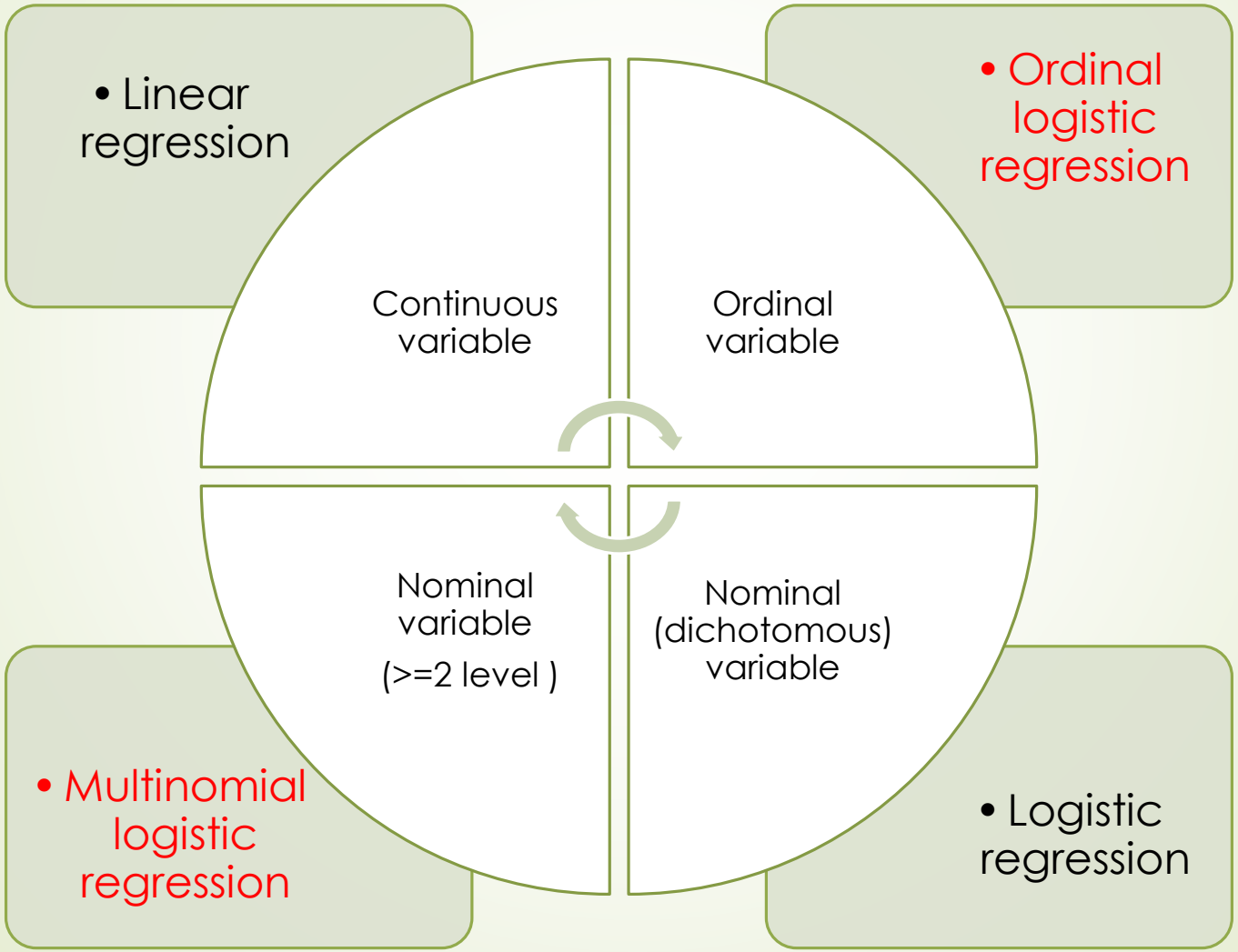
**. Correlation is significant at the 0.01 level (2-tailed).



- ضریب تعیین، به مجذور ضریب همبستگی دو متغیر گفته می شود
- در این مثال، برای دور کمر و کلسترول برابر 0.11 می باشد.
- یعنی میزان تبیین واریانس مشترک بین دو متغیر ۱۱ درصد است.

انتخاب مدل آماری مناسب برای اهداف مربوط به بررسی ارتباط: برای انواع متغیرهای وابسته

Dependent variable



- Univariate simple Linear regression (crude) :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.338[a]	.114	.111	35.042	.114	38.378	1	297	.000

a. Predictors: (Constant), Waist _C

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	98.637	14.378		6.860	.000	70.341	126.934
	Waist _C	1.005	.162	.338	6.195	.000	.686	1.324

a. Dependent Variable: Cholestrol

	Cholesterol							
	Format 1		Format 2			Format 3		
Waist. C	Standardized Beta	p-value	Unstandardized		p-value	Unstandardized		p-value
			Beta	S E		Beta	Confidence interval (95 %)	
	0.34	<0.001	1.01	0.16	<0.001	1.01	(0.69 , 1.32)	<0.001

- Univariate multiple Linear regression (adjusted) :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.400[a]	.160	.151	34.250	.160	18.685	3	295	.000

a. Predictors: (Constant), AGE(years), SEX, Waist _C

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	103.518	15.345		6.746	.000	73.319	133.717
	Waist _C	.718	.174	.242	4.119	.000	.375	1.061
	SEX	-.957	3.964	-.013	-.241	.809	-8.757	6.844
	AGE(years)	.560	.141	.233	3.970	.000	.283	.838

a. Dependent Variable: Cholestrol

Logistic Regression :

تیپ شخصیتی	Odds Ratio (95 % confidence interval)	P-VALUE
روان رنجور خویی	1.19 (1.18,1.82)	<0.001
برون گرایی	0.92 (0.9,0.93)	<0.001

Categorical Variables Codings

		Frequency	Parameter coding (1)
ever_smoke	never	3253	.000
	past & current	722	1.000
Sex	Male	1633	1.000
	Female	2342	.000
marital2	Married	3250	1.000
	Single & Divorced or Widowed	725	.000
IBS	No	3098	.000
	Yes	877	1.000

Original Value	Internal Value
No Case	0
Case	1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	IBS(1)	.625	.098	40.469	1	.000	1.868	1.541	2.264
	Sex(1)	-.415	.096	18.556	1	.000	.660	.547	.798
	Age	.025	.006	18.534	1	.000	1.026	1.014	1.038
	marital2(1)	-.312	.118	6.967	1	.008	.732	.581	.923
	ever_smoke(1)	.383	.112	11.725	1	.001	1.467	1.178	1.828
	N	.174	.008	490.585	1	.000	1.189	1.171	1.208
	E	-.087	.009	88.837	1	.000	.917	.900	.933
	O	.002	.010	.032	1	.859	1.002	.982	1.022
	A	-.015	.010	2.453	1	.117	.985	.967	1.004
	C	.016	.009	3.054	1	.081	1.016	.998	1.034
	Constant	-2.974	.465	40.828	1	.000	.051		

a. Variable(s) entered on step 1: IBS, Sex, Age, marital2, ever_smoke, N, E, O, A, C.

- Univariate Multiple Logistic Regression :

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	3281.697[a]	.313	.448

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			dep_cat2		Percentage Correct
	Observed		No Case	Case	
Step 1	dep_cat2	No Case	2570	260	90.8
		Case	503	642	56.1
	Overall Percentage				80.8

a. The cut value is .500

Specificity

Sensitivity

Accuracy

- Univariate Multiple Multinomial Logistic Regression :

Parameter Estimates

bmi.group[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
25 – 30	Intercept	−1.427	.073	382.130	1	.000			
	age.years	.025	.002	232.306	1	.000	1.025	1.022	1.028
	[gender=1.00]	.416	.047	77.383	1	.000	1.516	1.382	1.663
	[gender=2.00]	0[b]	.	.	0	.	.	.	.
	[tertile_nuts=1]	.129	.058	4.909	1	.027	1.138	1.015	1.276
	[tertile_nuts=2]	−.045	.058	.593	1	.441	.956	.853	1.072
	[tertile_nuts=3]	0[b]	.	.	0	.	.	.	.
>=30	Intercept	−2.883	.100	830.665	1	.000			
	age.years	.032	.002	251.452	1	.000	1.032	1.028	1.036
	[gender=1.00]	1.052	.063	278.409	1	.000	2.864	2.531	3.241
	[gender=2.00]	0[b]	.	.	0	.	.	.	.
	[tertile_nuts=1]	.161	.075	4.578	1	.032	1.175	1.014	1.362
	[tertile_nuts=2]	.025	.077	.107	1	.743	1.025	.883	1.191
	[tertile_nuts=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: bmi<25.

b. This parameter is set to zero because it is redundant.

	25<BMI <30			BMI >=30		
	Odds ratio	Confidence interval (95%)	P-value	Odds ratio	Confidence interval (95%)	P-value
Tertile of nuts						
Tertile 1	1.14	(1.01 , 1.27)	0.03	1.17	(1.01 , 1.36)	0.032
Tertile 2	0.96	(0.85 , 1.07)	0.44	1.03	(0.88 , 1.19)	0.74
Tertile 3	R	R	R	R	R	R

Univariate Multiple Ordinal Logistic Regression :

pcad_severity	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
male	2.082182	.1629475	9.37	0.000	1.786098	2.427348
fastfood1	1.005889	.0009192	6.43	0.000	1.004089	1.007693
smoking						
ever	1.231716	.0960963	2.67	0.008	1.057064	1.435224
/cut1	.0237873	.0535573			-.0811832	.1287577
/cut2	1.218839	.0582158			1.104738	1.33294
/cut3	2.097151	.0652304			1.969302	2.225

	Odds Ratio (95 % confidence interval)	P-VALUE
Smoking (ever vs. never)	1.23 (1.06,1.43)	0.008