مقدمات آمار کاربردی

در مطالعات علوم پزشکی

کارگاه روش تحقیق
دوشنبه ۱۹ آبان ۹۵
دانشگاه علوم پزشکی استان سمنان

دکتر مجید میر محمدخانی
اپیدمیولوژیست

داده ، اطلاعات،آگاهی

- بین کلمات "داده" (data) ، "اطلاعات" (information) و "آگاهی" (intelligence) تفاوت وجود دارد.
  - **داده** شامل مشاهدات جداگانه‌ای از صفات یا رویدادها است که وقتی به تنهایی مورد توجه قرار می گیرند دارای معنای کمی هستند؛ به طوری که داده‌های جمع آوری شده از نظام‌ها یا مراکز اداری سلامت برای برنامه‌ریزی کفایت لازم را ندارند.
  - داده‌ها باید تبدیل به **اطلاعات** شوند و این کار با خلاصه‌سازی، جمع‌بندی و تطبیق آنها برای متغیرهایی نظیر ترکیب سنی و جنسی جامعه صورت می گیرد تا انجام مقایسه‌های زمانی و مکانی میسر گردد.
  - اطلاعات با یکپارچه‌سازی و پردازش براساس درک و تجربیات برآمده از ارزش‌های اجتماعی و سیاسی تبدیل به **آگاهی** می گردد.

داده‌هایی که به اطلاعات و اطلاعاتی که به آگاهی تبدیل نمی‌شوند، ارزش زیادی برای راهنمایی تصمیم گیرندگان، سیاست‌گذاران، برنامه‌ریزان، مجریان و خود کارکنان مراقبت سلامت نخواهند داشت.

ارائه‌ی داده‌های آماری

- در هر زمینه‌ای از تحقیق و جستجو ، ابتدا داده‌ها جمع‌آوری می‌شوند و پس از طبقه‌بندی و تحلیل، درستی نتایج برآمده از آنها با استفاده از آزمون‌های آماری ارزیابی می گردد.

- بعد از این که داده‌های آماری جمع آوری شد، بایستی هدفمندانه مرتب شوند تا بتوان نکات مهم را به‌طور واضح و کامل از آن‌ها استخراج نمود. روش های متفاوتی برای این کار وجود دارد:

  - ترسیم جداول (tables) ، نمودارها(charts) ، طرح‌ها (diagrams) ، نگاره‌ها(graphs) ، تصاویر(pictures) ، منحنی ها (curves) ، ....
  - استفاده از شاخص های آماری شامل آماره های مرکزی (معدل های آماری) و مقادیر پراکندگی

ترسیم جدول

- به‌وسیله ترسیم جداول می‌توان به‌سادگی انبوهی از داده‌های آماری را به نمایش گذارد.
- اولین اقدام قبل از تجزیه و تحلیل داده‌ها ساختن جداول می باشد.
- جدول ممکن است ساده یا پیچیده باشد که به تعداد یا اندازه‌ی مجموعه‌ی منفرد یا مجموعه‌های متعدد حاوی موضوعات مورد بررسی بستگی دارد.

Table 1. Means and Standard Deviations for the Duration of Retention in the Program per Week

Variable	Retention, wk [a]	P Value [b]
Prison		< 0.001
Isfahan	23.5 ± 0.9	
Hamadan	11.3 ± 1.0	
Tehran	26.4 ± 0.3	
Drug injection		0.008
Yes	25.9 ± 0.7	
No	21.9 ± 0.8	
Main used drug		0.001
Opium/opium extract	18.6 ± 3.1	
Heroin and crack	24.6 ± 0.6	
Ecstasy	25.9 ± 1.0	
Other drugs	9.8 ± 0.9	
HBV test		0.02
Negative	25.0 ± 0.6	
Positive/unknown	23.1 ± 0.8	
Occupation/income		0.05
Stable job / regular income	22.7 ± 1.1	
Instable job / irregular income	24.1 ± 0.6	
Main cause of imprisonment		< 0.001
Drug-related crime	25.8 ± 0.6	
Crimes not related to drugs	18.4 ± 1.5	
History of involvement in harm reduction programs		< 0.001
Yes	19.9 ± 1.2	
No	25.7 ± 0.6	

[a] Data are presented as Mean ± SD.
[b] Log-Rank Test.

اصول طراحی جدول

الف- جداول باید شماره‌گذاری شوند مثل جدول۱، جدول۲ و نظیر آن.

ب- برای هر جدول باید عنوانی درنظر گرفته شود. عنوان باید مختصر و به‌تنهایی گویا باشد.

ج- عناوین ستون‌ها و سطرها باید روشن و مختصر باشند.

د- داده‌ها بایستی به ترتیب براساس مقدار یا اهمیت ، تقدم و تاخرزمانی، حروف الفبا و یا طبق ملاحظات جغرافیایی ارائه شوند.

ه- اگر قرار است مقادیر به‌شکل درصد یا میانگین با هم مقایسه شوند تاآن‌جا که می شود بهتر است مقادیر آن‌ها نزدیک به‌هم ارائه شوند.

و- جدول نباید خیلی طولانی باشد.

ز- بیشتر مردم، جداول تنظیم شده به صورت عمودی را بهتر از افقی می‌دانند ، چرا که مرور داده‌ها از بالا به پایین آسانتر است.

ح- می توان برای جدول زیرنویس (Foot notes) تهیه کرد، تا در صورت لزوم ارائه‌دهنده‌ی اطلاعات اضافی یا نکات توضیحی باشد.

مثال هایی از جداول ساده

آ) جدول ۱، جمعیت برخی از ایالات هند*

States	Population 2001
Andhra Pradesh	75,727,541
Bihar	82, 878, 796
Madhya Pradesh	60,385,118
Uttar Pradesh	116,052,859

* Source: Census of India, 2001

جدول ۲، جمعیت هند

Year	Population
1901	238,396,000
1921	251,321,000
1981	685,185,000
1991	843,930,000
2001	1027,015,247

* Source: Census of India, 2001

جمعیت برخی از استان‌های ایران در سال ۱۳۹۰

استان	بیت سال
تهران	۱۲.۱۸۳.۳۹۱
خراسان رضوی	۵.۹۹۴.۴۰۲
اصفهان	۴.۸۷۹.۳۱۲
فارس	۴.۵۹۶.۶۵۸
آذربایجان شرقی	۳.۷۲۴.۶۲۰

جمعیت ایران سال‌های ۱۳۵۵ تا ۱۳۹۰

سال	کشور
۱۳۵۵	۳۳.۷۰۸.۷۴۴
۱۳۶۵	۴۹.۸۵۷.۳۸۴
۱۳۷۵	۶۰.۰۵۵.۴۸۸
۱۳۸۵	۷۰.۴۹۵.۷۸۲
۱۳۹۰	۷۵.۱۴۹.۶۶۹

جدول توزیع فراوانی

- در جدول توزیع فراوانی، ابتدا داده‌ها به گروه‌های مناسب (تحت عنوان فواصل طبقه‌ای یا Class intervals) تقسیم شده و تعداد افراد (فراوانی) هر گروه ، در ستون کنارآن نمایش داده می شود.

نمونه: اعداد زیر سن بیمارانی را نشان می‌دهد که با تشخیص فلج کودکان در یک بیمارستان بستری شده‌اند.
جدول توزیع فراوانی آن را ترسیم نمایید.

۸، ۲۴، ۱۸، ۵، ۶، ۱۲، ۴، ۳، ۳، ۲، ۳، ۲۳، ۹، ۱۸، ۱۶، ۱، ۲، ۳، ۵، ۱۱، ۱۳، ۱۵، ۹، ۱۱، ۱۱، ۷، ۱۰، ۶، ۹، ۵، ۱۶، ۲۰، ۴، ۳، ۳، ۳، ۱۰، ۳، ۲، ۱، ۶، ۹، ۳، ۷، ۱۴، ۸، ۱، ۴، ۶، ۴، ۱۵، ۲۲، ۲، ۱، ۴، ۷، ۱، ۱۲، ۳، ۲۳، ۴، ۱۹، ۶، ۲، ۲، ۴، ۱۴، ۲، ۲، ۲۱، ۳، ۲، ۹، ۳، ۲، ۱، ۷، ۱۹

داده‌های مذکور را به راحتی می‌توان به‌صورت زیر مرتب نمود:

گروه سنی	فراوانی
-	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
-	𝍸 𝍸 𝍸 ///
-	𝍸 𝍸 /
-	𝍸 ///
-	𝍸 /

داده‌هایی که به شکل مذکور مرتب شده‌اند، به‌صورت جدول فراوانی نمایش داده می‌شوند:

جدول۳، توزیع سنی بیماران مبتلا به فلج کودکان

سن	شمار بیماران
۴-۰	۳۵
۹-۵	۱۸
۱۴-۱۰	۱۱
۱۹-۱۵	۸
۲۴-۲۰	۶

جدول۳، توزیع سنی بیماران مبتلا به فلج کودکان

سن	شمار بیماران
۴-۰	۳۵
۹-۵	۱۸
۱۴-۱۰	۱۱
۱۹-۱۵	۸
۲۴-۲۰	۶

- در مثال بالا، بیماران از نظر سن به ۵ گروه تقسیم شده‌اندو تعداد مشاهدات در هر گروه *فراوانی* نام دارد.

  - ✓ سوالی که در رسم جداول توزیع فراوانی مطرح می شود این است که داده‌ها را در چند گروه طبقه‌بندی کنیم؟ و چه فواصلی باید برای طبقه‌بندی درنظر گرفته‌شود؟
  - ✓ پاسخ: بر اساس یک قانون عملی وقتی داده‌های زیادی وجوددارد حداکثر ۲۰ گروه و وقتی داده‌ها چندان زیاد نیست، حداقل ۵ گروه قابل تعیین است.

- تا جایی که مقدور است فاصله‌ی طبقات باید برابر باشد تا بتوان مشاهدات را با هم مقایسه کرد.
- مزیت جدول توزیع فراوانی آن است که در یک نگاه نشان می دهد چه تعداد از مشاهدات در هر گروه قرار گرفته‌اند و تراکم اصلی در کدام گروه واقع شده‌است.

نمودارها و طرح ها

- نمودارها و طرح‌ها روش‌های مفیدی برای نمایش داده‌های ساده‌ی آماری محسوب می شوند.
  - آن‌ها تاثیر قوی بر ذهن مردم می‌گذارند. بنابراین برای تبیین داده‌های آماری ، به خصوص در روزنامه‌ها و مجله‌ها ، رسانه‌های متداولی محسوب می‌شوند.
- در مورد نمودار و طرح سه نکته مهم قابل ذکر است:

I. طرح‌ها بهتر از جداول آماری در ذهن می‌مانند.

II. داده‌هایی که قرار است به شکل طرح نمایش داده شوند ، باید ساده باشند. به این‌ترتیب نگران این نخواهیم بود که خواننده در فهم مطلب دچار مشکل شود.

III. سادگی طرح به قیمت کاهش صحت و دقت داده‌ها تمام خواهد شد. یعنی بسیاری از جزئیات داده‌های اولیه، در نمودارها و طرح ها از بین خواهد رفت. اگر ما در پی نکات واقعی یک مطالعه هستیم، بایستی به داده های اولیه رجوع کنیم.

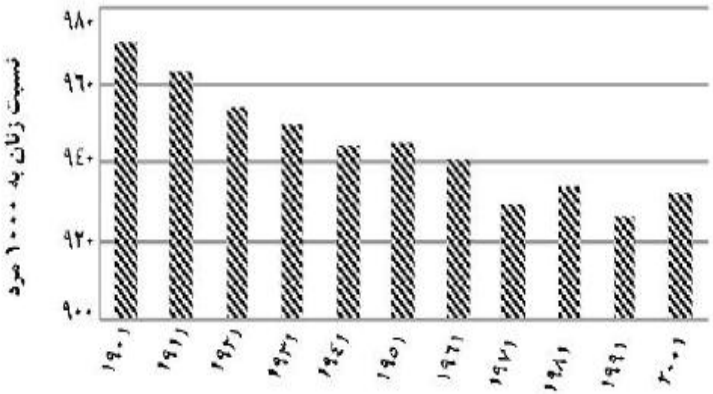
نمودارهای ستونی

- نمودارهای ستونی (Bar charts) برای نشان‌دادن مجموعه ای از اعداد به‌کار می‌روند که درآن طول هر ستون، متناسب با بزرگی اندازه‌ای است که قرار است نمایش داده شود.
- از آنجایی‌که نمودارهای ستونی به راحتی قابل ترسیم هستند و با آن‌ها می توان ارقام را به طور چشمی مقایسه نمود ، وسیله‌ی متداولی به شمار می‌آیند.

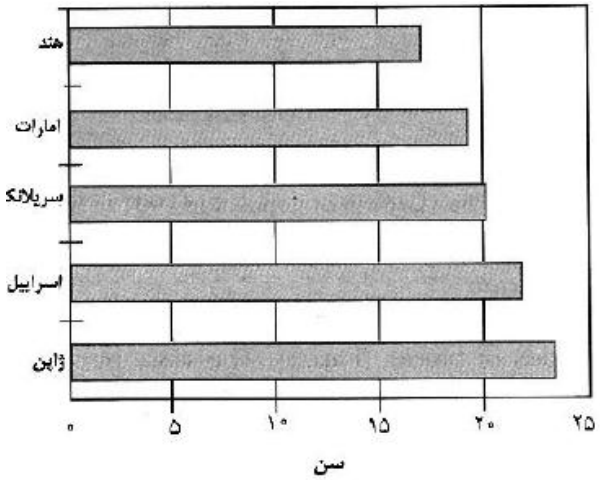
- انواع نمودارهای ستونی:
  - الف- نمودار ستونی ساده (simple)
  - ب- نمودار ستونی چندتایی (multiple)
  - ج- نمودار ستونی مرکب (component)

نمودار ستونی ساده

- ستون‌ها یا عمودی هستند یا افقی.
- ستون‌ها معمولا با فاصله‌هایی از هم جدا می‌شوند تا منظم‌تر و واضح‌تر به‌نظربرسند.
- ازنظر اندازه لازم است مقیاس مناسبی برای نمایش ستون ها انتخاب شود.



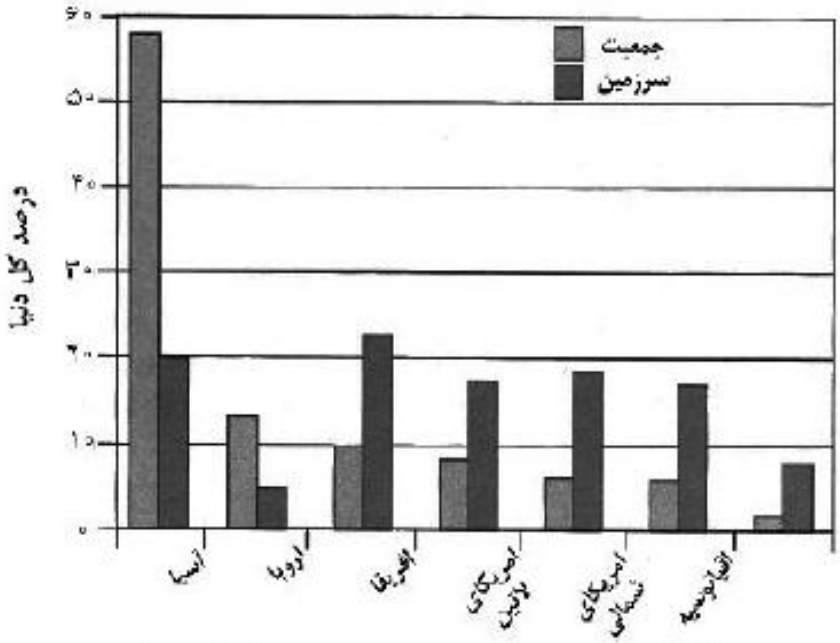
شکل ۱. کسر جنسی[1] در هند،۱۹۰۱ تا ۲۰۰۱



شکل ۲، میانگین سن ازدواج(زنان) در برخی کشورها

نمودار ستونی چندتایی

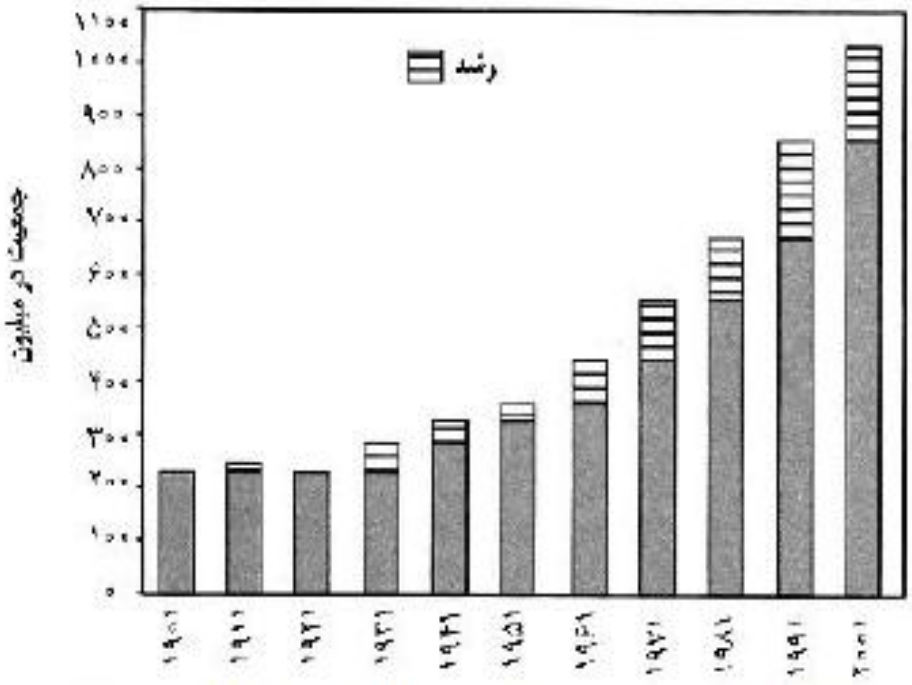
- شکل زیر ، نمونه‌ای از نمودارستونی چندتایی یا نمودار ستونی ترکیبی را نشان می‌دهد. در این نوع از نمودار، دو ستون یا بیشتر ممکن است با هم یک گروه را تشکیل دهند. در این شکل، جمعیت و وسعت سرزمین در مناطق مختلف دنیا با هم مقایسه شده‌اند.



شکل۳، جمعیت و وسعت سرزمین به تفکیک منطقه

نمودار ستونی مرکب

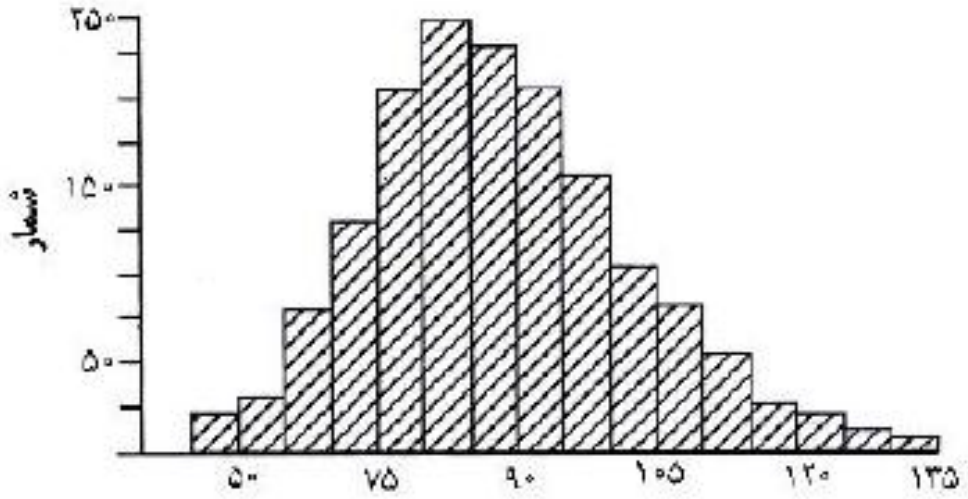
- ستون‌ها ممکن است به دو یا چند قسمت تقسیم شوند. هر قسمت مربوط به موضوع خاصی است و اندازه‌ی آن متناسب با بزرگی یا مقدار همان موضوع می‌باشد.



شکل۴: هند، رشدجمعیت در سال‌های ۱۹۰۱ تا ۲۰۰۱

هیستوگرام

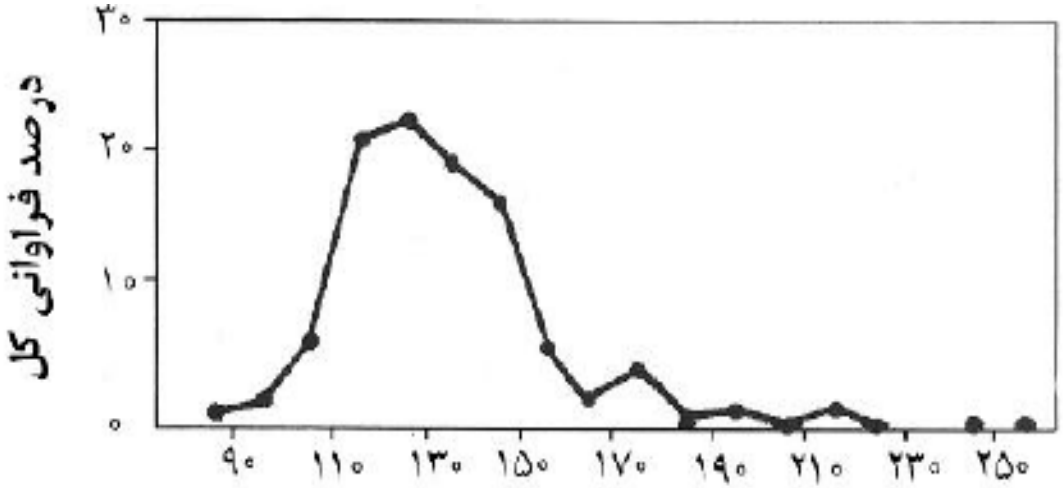
- هیستوگرام (histogram) یک طرح تصویری از توزیع فراوانی است و دارای مجموعه‌ای از قطعات مستطیل شکل (blocks) می باشد.
  - ❖ فواصل طبقه‌ای در محور افقی و فراوانی‌ها در محور عمودی قرار می گیرند.
  - ❖ مساحت هر قطعه یا مستطیل متناسب با فراوانی موضوع متناظر است.



شکل۵. توزیع فراوانی فشارخون دیاستولیک در زنان ۴۵ تا ۶۴ سال

چندگوشه‌ی فراوانی

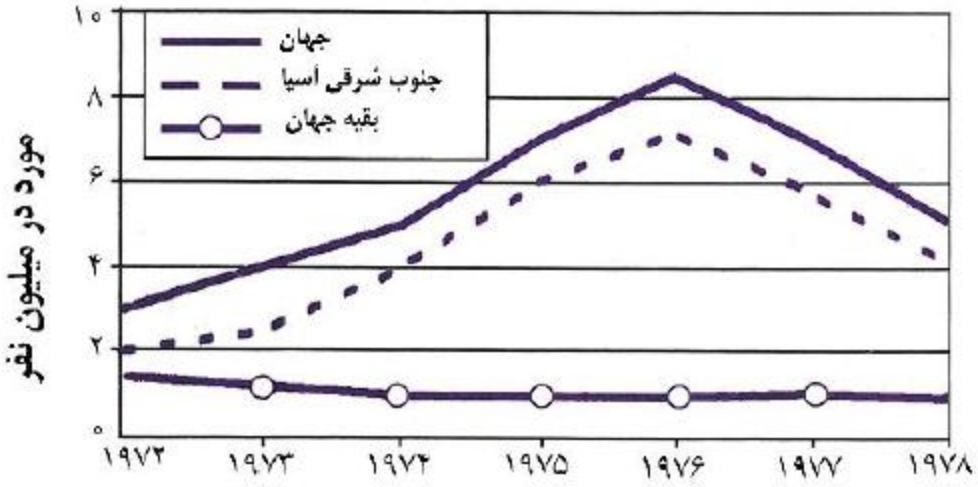
- توزیع فراوانی را می‌توان در قالب طرحی موسوم به چندگوشه‌ی فراوانی (Frequency Polygon) نشان داد.
  - این طرح از اتصال نقاط میانی قطعات هیستوگرام به دست می‌آید. شکل زیر ، چندگوشه‌ی فراوانی فشار خون سیستولیک اندازه‌گیری شده را در یک جامعه نشان می دهد.



شکل۶، توزیع فراوانی فشارخون سیستولیک اندازه‌گیری شده

طرح‌خطی

- طرح‌های خطی (Line Diagram) برای نشان دادن روند رویدادها در طول گذر زمان به کار می‌روند.
  - شکل زیر نمونه‌ای از یک طرح‌خطی می باشد که روند گزارش موارد مالاریا در جهان(غیر از منطقه‌ی آفریقا) را طی سال های ۱۹۷۲ تا ۱۹۷۸ نشان می‌دهد.



شکل۷، موارد مالاریای گزارش شده، ۱۹۷۱-۱۹۷۸ (به استثنای منطقه‌ی آفریقایی)

نمودار دایره ای

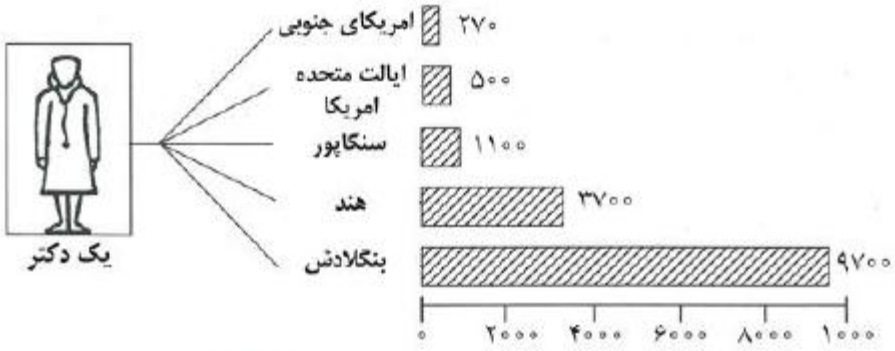
- در نمودار دایره ای یا کلوچه ای (Pie chart) برخلاف نمودار ستونی که در آن طول ستون‌ها با هم مقایسه می‌شد ، مساحت بخش‌های یک دایره با هم مقایسه می‌شود.
- مساحت هر بخش به زاویه‌ی آن بستگی دارد.
  - برخلاف مقبولیت زیاد این نوع نمودار میان عموم مردم ، آمارشناسان نمودارهای ستونی را بهتر از نمودارهای دایره‌ای می‌دانند.
  - اغلب لازم است برای هر بخش از دایره درصد فراوانی متناظر ذکر شود، چراکه گاهی مقایسه‌ی چشمی مساحت میان بخش‌ها به آسانی مقدور نمی‌باشد.

شکل ۸، جمعیت جهان

طرح‌تصویری

- طرح‌تصویری (**Pictogram**) میان مردم کوچه و بازار و برای کسانی که از نمودارهای دقیق و پیچیده کمتر سردرمی‌آورند، یک روش مقبول محسوب می‌شود.
- در این روش از نمادها یا عکس‌های کوچک استفاده می‌شود.

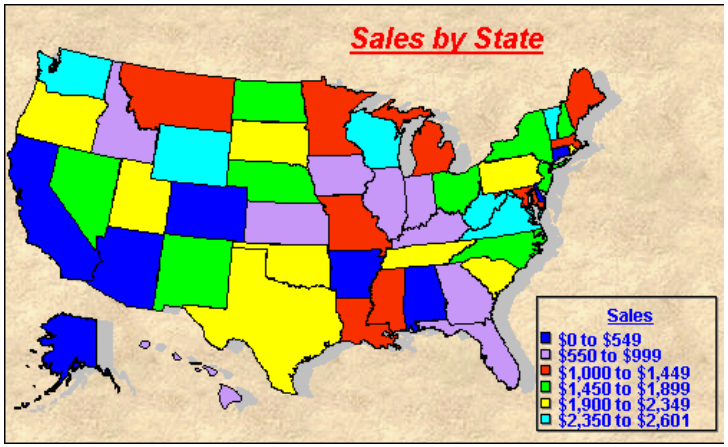
❖ مثلا در طرح تصویری زیر ، نسبت پزشک به جمعیت در کشورهای مختلف با استفاده از تصویر ساده‌ی یک پزشک نشان داده‌شده‌است.

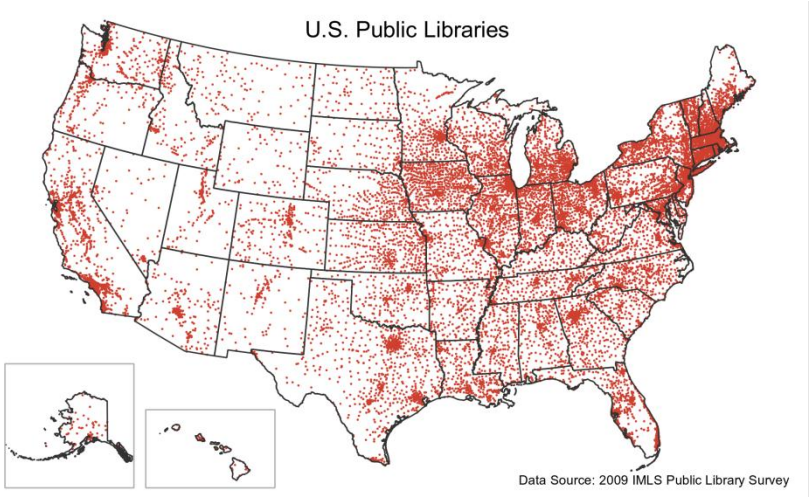


شکل ۹، نسبت جمعیت به پزشک، ۱۹۸۱

نقشه های آماری

- وقتی داده‌های آماری مربوط به مناطق جغرافیایی یا اجرایی باشند، می‌توان آن‌ها را با استفاده از "نقشه‌های هاشورزده" (Shaded maps) یا "نقشه‌های نقطه‌گذاری‌شده‌ی" (Dot maps) مناسب نیز نشان داد.
  - نقشه‌های هاشورزده برای نشان‌دادن داده‌هایی استفاده می‌شوند که از نظر اندازه متفاوت باشند.
  - قسمت‌های مختلف نقشه با استفاده از رنگ‌های مختلف یا با شدت‌های متفاوت از یک رنگ ، هاشورزده می‌شوند و درکلید نقشه برای راهنمایی بیننده تفسیر هرکدام آورده می‌شود.

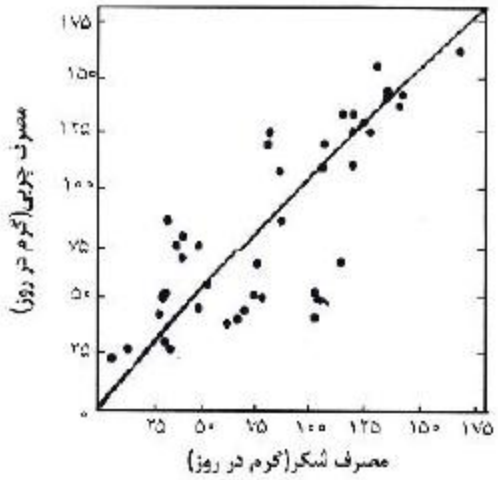




طرح نقاط پراکنده

- *طرح نقاط پراکنده* (Scatter) برای نشان‌دادن ارتباط میان دو متغیر به‌کار می رود.
- به عنوان نمونه درشکل زیر، ارتباط مثبت بین متوسط مصرف غذایی چربی و مصرف قند در ۴۱ کشور نشان داده‌شده‌است.

- نکته: اگر تجمع نقاط بیشتر حول‌وحوش یک خط صاف باشد، شواهد به نفع وجود ارتباط با ماهیت خطی بین متغیرها خواهد بود . اما اگر پراکندگی نقاط به این صورت نباشد، احتمالا ارتباطی بین متغیرها وجود ندارد.



شکل ۱۰، ارتباط بین متوسط مصرف چربی و قند در ۴۱ کشور

معدل‌های آماری

- واژه "معدل آماری" (Statistical Averages) گویای مقداری از توزیع است که دیگر مقادیر حول‌وحوش آن پراکنده شده‌اند ، به‌طوری‌که از نظر مقدار تصویری ذهنی از وسط (central) توزیع را به ما می‌دهد.

- انواع مختلفی از معدل‌های آماری وجود دارد که رایج‌ترین آنها عبارتند از:

- (۱)میانگین حسابی
- (۲)میانه
- (۳)نما

میانگین

میانگین حسابی کاربرد زیادی در محاسبات آماری دارد. گاهی آن را برای سادگی فقط میانگین (Mean) می‌نامند. برای محاسبه‌ی میانگین ابتدا اندازه‌ی همه مشاهده‌ها با هم جمع و سپس مجموع به‌دست آمده بر شمار کل مشاهده‌ها بخش می‌شود. عمل جمع کردن مشاهده‌ها را «جمع‌زنی» (summation) می‌نامند و با علامت Σ یا S نشان می‌دهند. شمار مشاهده‌ها با علامت n و میانگین را با علامت $\overline{X}$ (که «X بار» خوانده می‌شود) نمایش می‌دهند.

میانگین ($\overline{X}$) به این ترتیب محاسبه می‌شود: فرض کنیم فشار خون دیاستولیک ۱۰ نفر ۸۳، ۷۵، ۸۱، ۷۹، ۷۱، ۹۵، ۷۵، ۷۷، ۸۴، ۹۰ بوده‌است. جمع کل فشار خون‌ها برابر است با ۸۱۰ و میانگین برابر است با ۸۱۰ بخش بر ۱۰ که می‌شود ۸۱.

برتری میانگین این است که محاسبه و فهم آن آسان است. از عیب‌های آن یکی این است که گاهی ممکن است زیر تأثیر اندازه‌های غیر عادی قرار گیرد. دیگر این‌که گاهی ممکن است از نظر مفهوم مضحک به نظر برسد؛ مانند میانگین بچه‌های متولد شده از هر زن در یک منطقه خاص ۴/۷۶ محاسبه شده است، در حالی که این رقم هرگز در واقعیت معنی نمی‌دهد. با وجود این عیب‌ها، میانگین حسابی از مفیدترین معدل‌های آماری به‌شمار می‌آید.

میانه

میانه (median)، معدلی است که به نوعی از میانگین متفاوت است و به جمع مقدارها و شمار مشاهده‌ها وابسته نیست. برای به دست آوردن میانه، ابتدا داده‌ها بر اساس مقدارشان به‌صورت صعودی یا نزولی مرتب می‌شوند و سپس مشاهده‌ای که از نظر مقدار در وسط قرار گرفته‌است به‌عنوان میانه نام می‌گیرد. برای مثال فشار خون دیاستولیک ۹ نفر در شکل۱۱ درج شده‌است. میانه عدد ۷۹ است که در وسط داده‌ها واقع شده است(شکل ۱۲).

فشارخون دیاستولیک
۸۳
۷۵
۸۱
۷۹
۷۱
۹۵
۷۵
۷۷
۸۴

شکل۱۱، داده‌های مرتب نشده

فشارخون دیاستولیک
۷۱
۷۵
۷۵
۷۷
۷۹ میانه
۸۱
۸۳
۸۴
۹۵

شکل۱۲، داده‌های مرتب شده بر اساس مقدار

اگر به جای ۹ مشاهده، ۱۰ مشاهده وجود می‌داشت، میانه با محاسبه‌ی میانگین دو رقم مربوط به مشاهده‌های میانی به‌دست می‌آمد. بنابراین هرگاه شمار مشاهده‌ها زوج باشد، برای به‌دست آوردن میانه باید میانگین دو رقم میانی را حساب کرد. مانند فشار خون دیاستولیک ۱۰ نفر به شرح مندرج در شکل ۱۳ است:

فشارخون دیاستولیک
۸۳
۷۵
۸۱
۷۹
۷۱
۹۵
۷۵
۷۷
۸۴
۹۰

فشارخون دیاستولیک	
۷۱	
۷۵	
۷۵	
۷۷	
۷۹	$\frac{۷۹+۸۱}{۲}$
۸۱	
۸۳	=۸۰
۸۴	
۹۰	
۹۵	

شکل۱۳، داده‌های مرتب شده بر اساس مقدار در این نمونه میانه برابر است با ۷۹+۸۱ بخش بر ۲ که می‌شود ۸۰

نما

نما (mode) به مقداری از مشاهده‌ها می‌گویند که بیشترین شمار را در یک توزیع داده‌ای دارد. در واقع نما مقداری است که بیشتر از همه دیده می‌شود و در میان مشاهده‌ها از همه «رایج‌تر» است. برای نمونه فشار خون دیاستولیک ۲۰ نفر به شرح زیر است: ۸۵، ۷۵، ۸۱، ۷۹، ۷۱، ۹۵، ۷۵، ۷۷، ۷۵، ۹۰، ۷۱، ۷۵، ۷۹، ۷۵، ۷۵، ۷۷، ۸۴، ۷۵، ۸۱، ۷۵

نما یا پرتکرارترین مشاهده در میان این داده‌ها ۷۵ است. برتری نما این است که فهم آن آسان است و تحت تأثیر مقدارهای خیلی کم یا خیلی زیاد قرار نمی‌گیرد. از نقاط ضعف آن یکی این است که جایگاه واقعی آن مشخص نیست و دیگر این که اغلب دارای تعریف روشنی نیست. بنابراین نما معمولا در آمارهای پزشکی و علوم زیستی استفاده نمی‌شود.

مقدارهای پراکندگی

مقدار کالری مورد نیاز برای فرد سالمی که در کارش تحرک چندانی وجود ندارد، حدود ۲۴۰۰ کالری در روز است. بدیهی است که این مقدار برای همه صادق نیست. بین انسان‌ها تفاوت‌های بسیاری وجود دارد. اگر مقدارهای فشارخون، قد یا وزن افراد را در یک گروه بزرگ اندازه‌گیری کنیم، خواهیم دید که این مقادیر از فردی به فرد دیگر متفاوت است. حتی در یک فرد هم، بسته به زمان اندازه‌گیری، مقدار آن متفاوت خواهد بود. پرسش‌هایی که مطرح می‌شود عبارت‌اند از این که: مقدار طبیعی این تغییرات چقدر است؟ و چگونه می‌توان این تغییرات را اندازه گرفت؟

چندین مقدار برای اندازه گیری تغییرات (یا در اصطلاح فنی «پراکندگی») وجود دارد که شماری از شناخته شده‌ترین آن‌ها عبارت‌اند از:

آ) دامنه (Range)　　ب) میانگین یا متوسط انحرافات　　پ) انحراف معیار (Standard Deviation)

الف)دامنه

- دامنه(range) را شاید بتوان ساده ترین شاخص برای اندازه‌گیری پراکندگی دانست و عبارت است از اختلاف بین بیشترین و کمترین مقادیر در یک نمونه.

➢ برای مثال مقادیر فشار خون دیاستولیک ۱۰ نفر را به‌شرح ذیل در نظر بگیرید:

۸۳، ۷۵، ۸۱، ۷۹، ۷۱، ۹۰، ۷۵، ۹۵، ۷۷، ۹۴

- همان‌طور که مشاهده می شود بیشترین مقدار ۹۵ و کمترین مقدار ۷۱ بوده‌است. بنابراین دامنه برابر است با فاصله‌ی ۷۱ تا ۹۵ و یا تفاضل مطلق این دو مقدار یعنی ۲۴.
- در داده‌های گروه بندی شده دامنه می‌تواند از تفاضل بین نقطه‌های میانی کوچک‌ترین و بزرگ‌ترین گروه‌ها نیز به دست آید.

دامنه در عمل کاربرد زیادی ندارد ، چرا که فقط نشان‌دهنده دو مقدار انتهایی است و پراکندگی مشاهداتی که بین این دو مقدار قرار دارند در نظر گرفته نمی شود.

ب) میانگین انحرافات

میانگین انحرافات، معدل مقدارهای انحرافات هر مشاهده از میانگین حسابی داده‌هاست و از فرمول زیر محاسبه می‌شود:

$$M.D. = \frac{\Sigma - \overline{x}}{\eta}$$

نمونه: فشار خون دیاستولیک ۱۰ نفر بـه ایـن شـرح اسـت: ۸۳، ۷۵، ۸۱، ۷۹، ۷۱، ۹۵، ۷۵، ۷۷، ۸۴، ۹۰. میـانگین انحرافات را محاسبه کنید.

پاسخ (میانگین انحرافات)

فشارخون دیاستولیک	میانگین حسابی	انحراف از میانگین
x	$\overline{x}$	$(x - \overline{x})$
۸۳	۸۱	۲
۷۵	۸۱	-۶
۸۱	۸۱	۰
۷۹	۸۱	-۲
۷۱	۸۱	-۱۰
۹۵	۸۱	۱۴
۷۵	۸۱	-۶
۷۷	۸۱	-۴
۸۴	۸۱	۳
۹۰	۸۱	۹
جمع = ۸۱۰	(بدون توجه به علامت ±)	جمع = ۵۶

$$\text{میانگین انحرافات} = \frac{۵۶}{۱۰} = ۵/۶ \qquad \text{میانگین} = \frac{۸۱۰}{۱۰} = ۸۱$$

پ) انحراف‌معیار

انحراف معیار رایج‌ترین مقدار پراکندگی به شمار می‌آید. انحراف معیار به زبان ساده «ریشه‌ی دوم میانگین مجذور انحرافات» است و با حرف یونانی ـ [سیگما] و یا به شکل مخفف S.D. [1] نوشته می‌شود. فرمول اصلی برای محاسبه‌ی انحراف معیار عبارت است از:

$$S.D. = \sqrt{\frac{\Sigma(x-\overline{x})^{۲}}{\eta}}$$

هنگامی حجم نمونه بزرگتر از ۳۰ باشد، فرمول یاد شده بدون هیچ تغییری به‌کار می‌رود؛ اما برای نمونه‌های کوچک‌تر فرمول مذکور انحراف معیار را کمتر از مقدار واقعی برآورد می‌کند و نیاز به اصلاح دارد که این کار با جانشینی $(\eta-۱)$ به جای η در مخرج کسر انجام می‌شود. فرمول تصحیح شده عبارت است از:

$$S.D. = \sqrt{\frac{\Sigma(x-\overline{x})^{۲}}{\eta-۱}}$$

۱- به‌گونه‌ی کلی از حرف سیگمای یونانی (Σ) برای بیان انحراف معیار برای کل جامعه و از حروف (S یا SD) برای بیان انحراف معیار نمونه‌ها استفاده می‌گردد.(ویر)

مراحلی که باید در محاسبه انحراف معیار انجام شود بدین شرح است:

آ) پیش از هرکاری، انحراف مقدار هر مشاهده را از میانگین حسابی به دست آورید. $(x-\bar{x})$

ب) سپس مجذور هر انحراف را حساب کنید. $(x-\bar{x})^{۲}$

پ) مجذور انحرافات را با هم جمع کنید. $\Sigma(x-\bar{x})^{۲}$

ت) مقدار به‌دست آمده را بر شمار کل مشاهدات (ħ) تقسیم کنید [اگر حجم نمونه کمتر از ۳۰ باشد بر $(\eta-۱)$ تقسیم می‌کنیم]

ث) آنگاه از نتیجه جذر بگیرید تا انحراف معیار به دست آید.

نمونه: فشار خون دیاستولی ۱۰ نفر به این شرح است: ۹۰، ۸۴، ۷۷، ۷۵، ۹۵، ۷۱، ۷۹، ۸۱، ۷۵، ۸۳ انحراف معیار را محاسبه کنید.

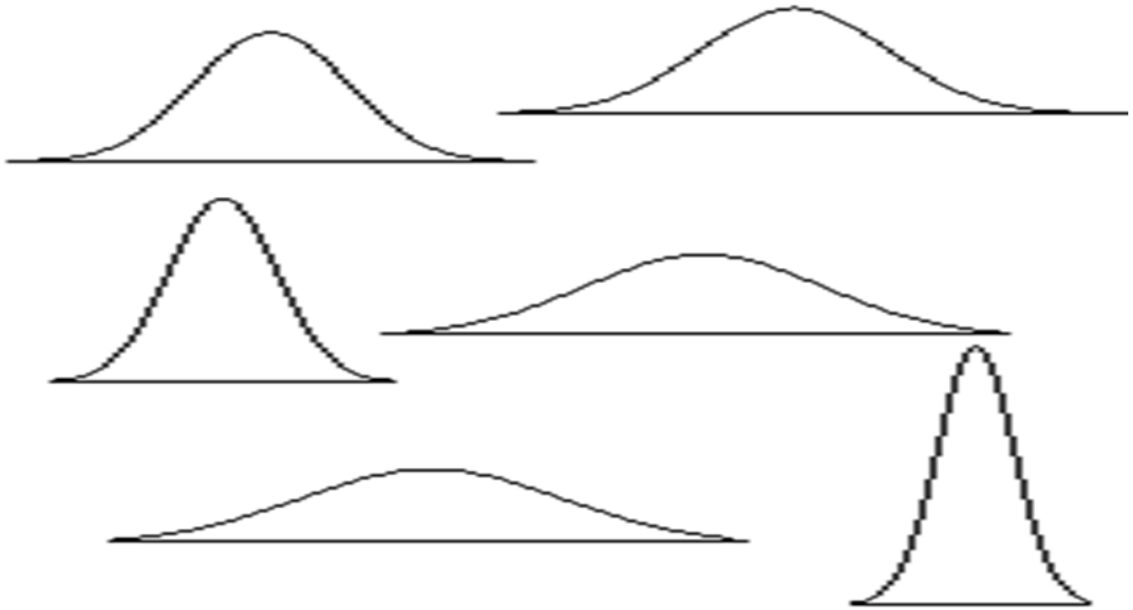
پاسخ

x	$(x-\bar{x})$	$\Sigma(x-\bar{x})^{۲}$
۸۳	۲	۴
۷۵	-۶	۳۶
۸۱	۰	-
۷۹	-۲	۴
۷۱	-۱۰	۱۰۰
۹۵	۱۴	۱۹۶
۷۵	-۶	۳۶
۷۷	-۴	۱۶
۸۴	۳	۹
۹۰	۹	۸۱
$\bar{x}=۸۱$ $\eta=۱۰$		جمع =۴۸۲

$$S.D.=\sqrt{\frac{\Sigma(x-\bar{x})^{۲}}{\eta-۱}}=\sqrt{\frac{۴۸۲}{۱۰-۱}}=\sqrt{\frac{۴۸۲}{۹}}=\sqrt{۵۳/۵۵}=۷/۳۱$$

توزیع نرمال

توزیع نرمال یا «منحنی نرمال» یکی از مفاهیم مهم در تئوری آمار به شمار می‌آید. بیایید فرض کنیم مقدار هموگلوبین شمار زیادی از مردم را گردآوری کرده و شکل توزیع فراوانی آن را در حالی که میان طبقات فاصله‌ی کمی وجود دارد، ترسیم نموده‌ایم. در این صورت به احتمال زیاد خواهیم دید یک منحنی متقارن و هموار به دست آمده است. چنین منحنی را توزیع یا منحنی نرمال می‌نامند. شکل منحنی بستگی به میانگین و انحراف معیار داده‌ها دارد که خود بستگی به شمار و طبیعت مشاهده‌ها دارد؛ بنابراین می‌توان نتیجه گرفت که شمار بی‌شماری منحنی نرمال وجود دارد.



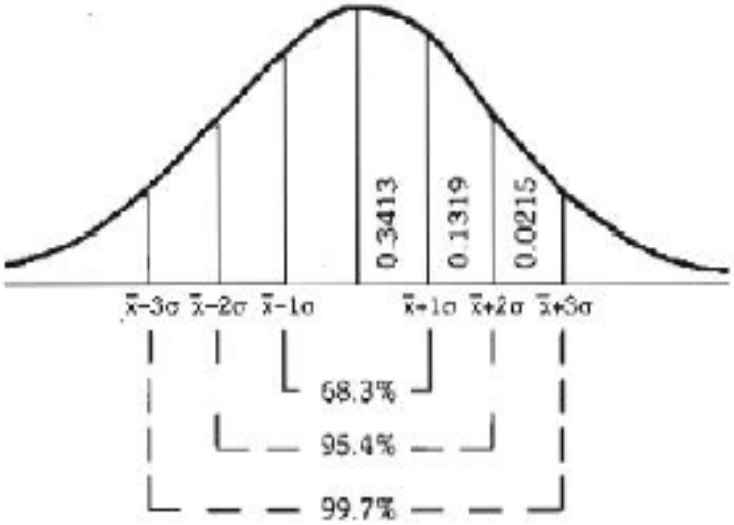
چند نکته در خصوص توزیع نرمال

دانستن برخی نکته‌ها درباره‌ی منحنی نرمال مفید است (شکل ۱۵):

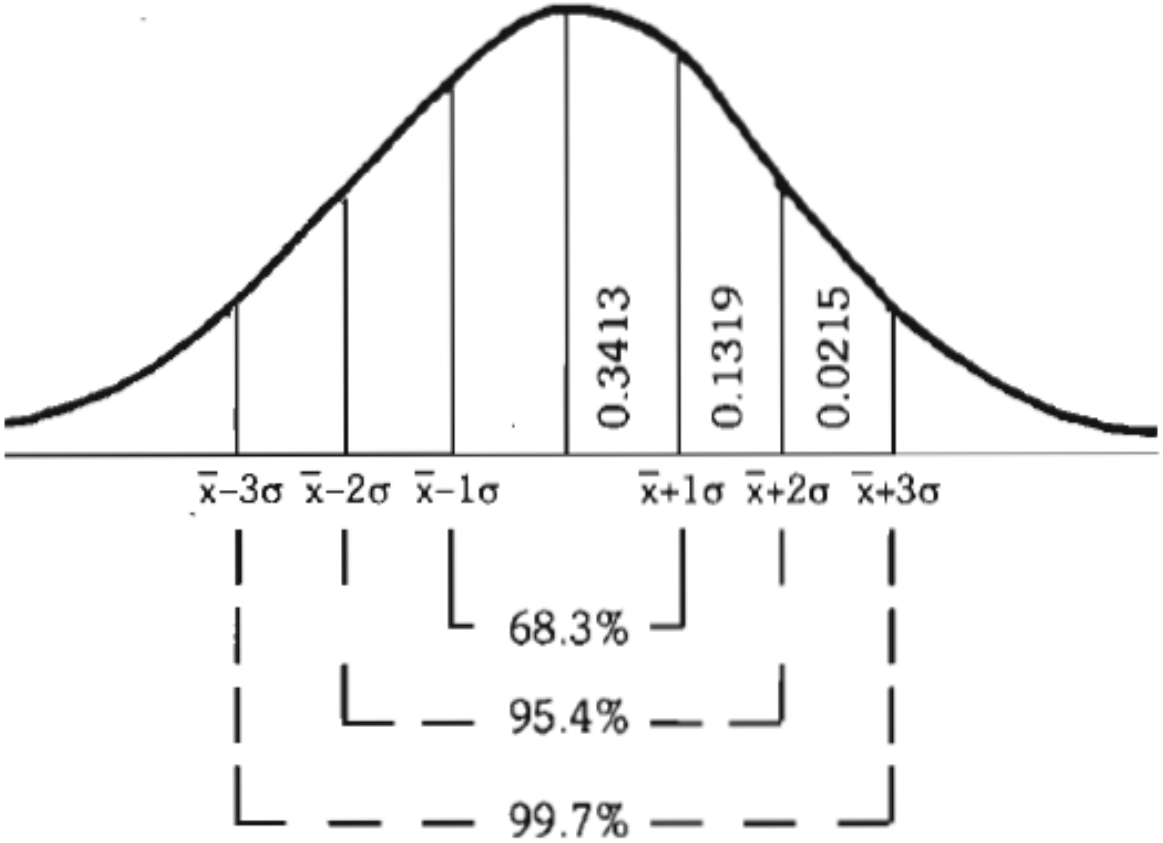
آ) مساحت محدود بین یک انحراف معیار $(\bar{x} \pm ۱\sigma)$ از طرفین میانگین تقریبا ٪۶۸ از مقادیر را در توزیع شامل می‌شود.

ب) مساحت محدود بین دو انحراف معیار $(\bar{x} \pm ۲\sigma)$ از طرفین میانگین اغلب مقادیر، یعنی تقریباً ٪۹۵ آن‌ها را در توزیع شامل می‌شود.

پ) مساحت محدود بین سه انحراف معیار $(\bar{x} \pm ۳\sigma)$ از طرفین میانگین ٪۹۹/۷ از مقادیر را شامل می‌شود. حدود مذکور در طرفین میانگین را «حدود اطمینان»[1] می‌نامند که در شکل ۱۵ نشان داده شده است.



شکل۱۵. منحنی نرمال



حدود اطمینان ٪۹۵ $(\bar{x} \pm ۲\sigma)$ را در نظر بگیرید. این عبارت به ٪۹۵ از کل مساحت زیر منحنی نرمال اشاره می‌کند و به این معناست که ٪۹۵ از کل مشاهده‌ها در فاصله $(\bar{x} \pm ۲\sigma)$ قرار می‌گیرند. پس احتمال این‌که یکی از مشاهده‌ها خارج از حدود اطمینان ٪۹۵ قرار گیرد، یک بیستم است(P=۰/۰۵).

منحنی نرمال استاندارد

اگر چه با توجه به تفاوت در میانگین و انحراف معیار، بی‌شمار منحنی نرمال وجود دارد، اما فقط یک منحنی نرمال استاندارد (standard) می‌توان رسم کرد که توسط آمارشناسان تعریف شده است، به‌گونه‌ای که می‌توان به راحتی مساحت بین هر دو نقطه از سطح زیر منحنی نرمال را محاسبه کرد. منحنی نرمال استاندارد یک منحنی هموار، زنگوله‌ای شکل و کاملا قرینه است که بر اساس شمار بی‌شماری از مشاهده‌های به‌دست می‌آید. مساحت کل زیر منحنی برابر با یک، میانگین آن صفر و انحراف معیار آن یک است. در منحنی نرمال میانگین، میانه و نما روی هم می‌افتند. در این منحنی، فاصله‌ی میانگین $(\overline{X})$ از یک مقدار (X) تقسیم بر انحراف‌معیار «انحراف نسبی» یا «متغیر نرمال استاندارد» نامیده می‌شود که آن را با z نشان می‌دهند. متغیر نرمال استاندارد یا z از فرمول زیر محاسبه می‌شود:

$$z = \frac{(x - \overline{x})}{\sigma}$$

یک متغیر تصادفی (x) وقتی استاندارد خوانده می‌شود که میانگین ان صفر و انحراف معیار ان یک باشد. برای استانداردکردن ابتدا میانگین x را از خود x کم کرده و حاصل تفاضل را بر σ یعنی انحراف معیار x تقسیم می‌کنیم. بنابراین $\frac{(x-\bar{x})}{\sigma}$ یک متغیر استاندارد شده است. از مفاهیم مهم آمار ریاضی این است که متغیر جدید z هم مثل متغیر x از توزیع نرمال پیروی می‌کند. میانگین توزیع متغیر جدید صفر و انحراف معیار (σ) آن یک است.

آگاهی از مقدارهای سطح زیر منحنی نرمال استاندارد معمولا نیاز است، که بر حسب مقدارهای مختلف، انحراف نسبی محاسبه می‌شود. نتایج محاسبه این مقادیر در جدول ۴ آمده است.

جدول۴، مقادیر سطح زیر منحنی نرمال استاندارد با میانگین صفر و انحراف‌معیار یک، منبع (۲۰)

سطح زیر منحنی نرمال از مقدار میانگین تا نقطه مورد بررسی	انحراف نسبی(Z) $\frac{(x-\bar{x})}{\sigma}$
۰.	۰
۰.۱۹۱۵	۰.۵
۰.۳۴۱۳	۱
۰.۴۳۳۲	۱.۵
۰.۴۷۷۲	۲
۰.۴۹۸۷	۳
۰.۴۹۹۹۷	۴
۰.۴۹۹۹۹۹۸	۵

برآورد احتمال (نمونه)

فرض کنید ضربان قلب گروهی از مردان سالم ۷۲ با انحراف معیار ۲ باشد. احتمال این‌که ضربان قلب مردی که شانسی از این جمعیت انتخاب شده است، ۸۰ یا بیشتر از ۸۰ باشد، چقدر است؟

$$\text{انحراف نسبی } (z) = \frac{(x - \bar{x})}{\sigma}$$

$$= \frac{۸۰ - ۷۲}{۲} = ۴$$

سطح زیر منحنی نرمال متناظر با انحراف نسبی ۴، برابر است با ۰/۴۹۹۹۷ (جدول۴). از آن‌جایی که فقط با نصف سطح کل سروکار داریم (یعنی ۰/۵) بقیه سطح یعنی ۰/۴۹۹۹۷–۰/۵ یا ۰/۰۰۰۰۳ میزان احتمال مذکور را بیان می‌کند. به عبارت دیگر فقط ۳ نفر از ۱۰۰.۰۰۰ نفر احتمال دارد ضربان قلبی برابر با ۸۰ یا بیشتر داشته باشند.

جدول۴، مقادیر سطح زیر منحنی نرمال استاندارد با میانگین صفر و انحراف‌معیار یک، منبع (۲۰)

سطح زیر منحنی نرمال از مقدار میانگین تا نقطه مورد بررسی	انحراف نسبی(Z) $\frac{(x - \bar{x})}{\sigma}$
۰.	۰
۰.۱۹۱۵	۰.۵
۰.۳۴۱۳	۱
۰.۴۳۳۲	۱.۵
۰.۴۷۷۲	۲
۰.۴۹۸۷	۳
۰.۴۹۹۹۷	۴
۰.۴۹۹۹۹۹۸	۵

خطای معیار (Standard error)

اگر از یک جمعیت نمونه‌ای به حجم (η) انتخاب و نمونه‌گیری به‌گونه‌ای مشابهی بارها تکرار شود، هر نمونه میانگین $(\overline{x})$ خاص خود را خواهد داشت. چنانچه منحنی توزیع فراوانی میانگین‌های نمونه‌ای را رسم کنیم، می‌بینیم که منحنی دارای توزیع تقریبا نرمال خواهد بود و عملا میانگین‌های نمونه‌ای برابر با میانگین خود جمعیت (μ) است. این مشاهده که میانگین‌های نمونه‌ای دارای توزیع نرمال در اطراف میانگین جمعیت هستند، موضوع بسیار مهمی است. انحراف معیار میانگین‌های نمونه‌ای که معیاری برای سنجش خطای نمونه‌گیری است و با فرمول $\sigma / \sqrt{\eta}$ محاسبه می‌شود، **خطای معیار یا خطای معیار میانگین** نامیده می‌شود. از آن‌جایی که میانگین نمونه‌ها از توزیع نرمال پیروی می‌کند به سادگی مشاهده‌پذیر است که ٪۹۵ از میانگین‌های نمونه‌ای در محدوده دو انحراف معیار $\left[\mu \pm ۲\left(\sigma / \sqrt{\eta}\right)\right]$ از طرفین میانگین جمعیت قرار می‌گیرند. بنابراین خطای معیار (S.E) شاخصی است که با استفاده از آن می‌توانیم قضاوت نماییم که آیا میانگین یک نمونه‌ی خاص درون حدود اطمینان واقع شده است یا نه.

آزمون‌های معنی‌داری

- درواقع خطای‌معیار نشان می‌دهد تا چه حد برآورد یک میانگین ممکن است درست باشد. لذا با کمک آن می توان آزمون های معنی داری یا همان (Tests of significance) را انجام داد.
- مفهوم خطای‌معیار با استفاده از فرمول‌های مناسب برای همه‌ی آماره‌ها (statistics) کاربرد دارد که برخی از پرکاربرد ترین آنها عبارتند از:

(الف) خطای‌معیار میانگین

(ب) خطای‌معیار نسبت

(ج) خطای‌معیار تفاوت دو میانگین

(د) خطای‌معیار تفاوت دو نسبت

خطای معیار میانگین

- پیش از این درباره‌ی معنای "خطای‌معیار میانگین" که به آن به طور ساده *خطای‌معیار* هم گفته می‌شود و نیز در خصوص توزیع میانگین‌های نمونه‌ای در اطراف میانگین واقعی جمعیت مطالبی بیان شد.
- در عمل معمولا نمونه‌گیری‌های تکراری از جمعیت انجام نمی دهیم ، بلکه با گرفتن فقط یک نمونه از جمعیت ، میانگین و خطای‌معیار را محاسبه می کنیم.

- در این‌جا سوالاتی به این شرح مطرح می شود : میانگین نمونه‌ی گرفته‌شده تا چه اندازه صحیح است؟ چه اظهار نظری درباره‌ی میانگین واقعی جمعیت می‌توان کرد؟
- برای پاسخ به این سوالات باید خطای‌معیار میانگین محاسبه شود و بر اساس آن "حدود اطمینانی"که میانگین واقعی جمعیت (که فقط نمونه‌ای از آن مورد مطالعه قرار گرفته‌است) محتمل است در آن قرار گیرد، تعیین شود.

مثال از محاسبه خطای معیار میانگین

- *مثال:* فرض کنید در مطالعه‌ی نمونه‌ای تصادفی شامل ۲۵ مرد ۲۰ تا ۲۴ ساله ، میانگین درجه حرارت بدن آن‌ها ۹۸/۱۴درجه‌ی فارنهایت با انحراف معیار ۰/۶به‌دست آمده باشد. در مورد میانگین درجه‌ی حرارت بدن افراد متعلق به جامعه‌ای که این نمونه از آن انتخاب شده‌است چه اظهار نظری می‌توان کرد؟

برای پاسخ به این پرسش از خطای معیار کمک می‌گیریم: $S.E.\bar{x} = S/\sqrt{n} = ۰/۶/\sqrt{(۲۵)} = ۰/۱۲$

حال می‌توانیم حدود اطمینان را بر اساس منحنی توزیع نرمال به دست آوریم. اگر حدود اطمینان در فاصله‌ی دو خطای معیار از میانگین محاسبه شود (یعنی حدود اطمینان٪۹۵) دامنه‌ی میانگین جمعیت بین ۹۸/۱۴ − (۲ ×۰/۱۲) = ۹۷/۹۰ تا ۹۸/۱۴ +(۲×۰/۱۲) =۹۸/۳۸ °F خواهد بود. احتمال این‌که میانگین جامعه خارج از این محدوده باشد، فقط ۱ در ۲۰ (P=۰/۰۵) است.

خیلی وقت‌ها در متون علمی به واژه‌ی **معنی‌دار** (significant) برخورد می‌کنیم. زمانی می‌گوییم «این تفاوت معنی‌دار است» منظورمان این است که بعید است این تفاوت شانسی رخ داده باشد. به‌عنوان یک قرارداد، هنگامی از کلمه‌ی معنی‌دار استفاده می‌کنیم که P <۰/۰۵ (۱ در ۲۰) باشد و هنگامی که P <۰/۰۱ (۱ در ۱۰۰) باشد، یعنی معنی‌داری بیشتری وجود دارد.

حدود معنی داری

حدود معنی‌داری	$(N.D.) = \frac{x-\mu}{\sigma/\sqrt{n}}$ انحراف نرمال	حدود اطمینان
$P < ۰/۰۵$ معنی‌دار در سطح کمتر از ٪۵	$N.D. > ۲$	μ بیرون از حدود اطمینان ٪۹۵ قرار دارد
$P = ۰/۰۵$ معنی‌دار درست در سطح ٪۵	$N.D. = ۲$	μ درست روی حدود اطمینان ٪۹۵ قرار دارد
$P > ۰/۰۵$ در سطح ٪۵ معنی‌دار نیست	$N.D. < ۲$	μ درون حدود اطمینان ٪۹۵ قرار دارد

خطای معیار نسبت

- فرض کنید نسبت مردان در جمعیت ساکن در یک روستا ۵۲ درصد باشد. از این روستا نمونه‌ای تصادفی شامل ۱۰۰ نفر گرفته‌ایم و نسبت مردان در این نمونه ۴۰ درصد است. از این نمونه چه نتیجه‌ای می‌گیریم؟ از یک نمونه ۱۰۰ تایی، دامنه‌ی احتمالی قابل انتظار برای نسبت مردان با حدود اطمینان ۹۵ درصد چقدر است؟

- در این نوع مثال‌ها دیگر نه با میانگین‌ها ، بلکه با نسبت‌ها در نمونه و در جامعه اصلی سروکار داریم. این نسبت‌ها را می‌توانیم با علامت p و q نشان دهیم و انحراف‌معیار را حول‌وحوش نسبت موردانتظار، یعنی ۵۲ درصد محاسبه نماییم که به آن *خطای‌معیار نسبت* می‌گویند و از فرمول زیر محاسبه می‌شود:

$$\text{خطای معیار نسبت} = \sqrt{\frac{pq}{n}}$$

که در آن: حجم نمونه= n؛ نسبت زنان= q؛ نسبت مردان =p

$$\text{معیار خطای نسبت} = \sqrt{\frac{۵۲ \times ۴۸}{۱۰۰}} = ۵$$

- دو خطای‌معیار را در طرفین نسبت ۵۲ درصد به عنوان ملاک در نظر می‌گیریم. به این معنی که اگر واقعا نمونه‌ی انتخاب شده نماینده‌ی خوبی برای جامعه اصلی باشد انتظار داریم نسبت برآورد شده از آن یک عدد تصادفی در محدوده‌ی ۶۲=(۵)۲+۵۲ تا ۴۲=(۵)۲-۵۲ درصد باشد.
- از آن‌جایی که نسبت مردان در نمونه مورد مطالعه فقط ۴۰ درصد و کاملا خارج از حدود اطمینان بوده‌است، تفاوت بین مقادیر نسبت مشاهده شده و نسبت موردانتظار "معنی‌دار" تلقی می‌شود و این تفاوت نمی‌تواند شانسی باشد. در این مورد خاص، *انحراف‌نسبی* برحسب خطای معیار ، یعنی ۵ برابر خواهد بود با:

$$\text{انحراف نسبی} = \frac{۵۲ - ۴۰}{۵} = ۲/۴$$

حدود اطمینان	$(N.D.) = \frac{x-\mu}{\sigma/\sqrt{n}}$ انحراف نرمال	حدود معنی‌داری
μ بیرون از حدود اطمینان ٪۹۵ قرار دارد	N.D. > ۲	P < ۰/۰۵ معنی‌دار در سطح کمتر از ٪۵
μ درست روی حدود اطمینان ٪۹۵ قرار دارد	N.D. = ۲	P = ۰/۰۵ معنی‌دار درست در سطح ٪۵
μ درون حدود اطمینان ٪۹۵ قرار دارد	N.D. < ۲	P > ۰/۰۵ در سطح ٪۵ معنی‌دار نیست

- چون مقدار انحراف نسبی از عدد ۲ بیشتر شده است ، لذا انحراف معنی‌دار می‌باشد.
- چنین آزمونی در مواردی کاربرد پیدا می‌کند که جمعیت فقط شامل دو گروه یا دو نسبت باشد ، برای مثال زنان و مردان، بیماران و افراد سالم ، موفقیت‌ها و شکست‌ها و مشابه آن.

خطای‌معیار تفاوت دو میانگین

- در پژوهش‌های زیست‌شناختی زیاد پیش می‌آید که پژوهشگر با مشکل مقایسه میان دو گروه شاهد و تجربه مواجه شود. سوالی که در این موارد مطرح است این است که آیا تفاوت بین دو گروه به اندازه‌ای معنی‌دار هست که نشان دهد این دو گروه متعلق به دو جمعیت متفاوت می باشند. پاسخ این سوال با محاسبه‌ی *خطای‌معیار تفاوت دو میانگین* داده خواهد شد.

فرض کنید می‌خواهیم تأثیر یک داروی خاص را بر موش‌ها آزمایش کنیم. ۲۴ موش که از هر لحاظ با هم مقایسه‌پذیر هستند شانسی به دو گروه تقسیم می‌شوند. گروه (آ) را به‌عنوان گروه شاهد در نظر می‌گیریم به این معنی که موش‌های این گروه هیچ درمانی دریافت نمی‌کنند. ولی موش‌های گروه (ب) که همان گروه آزمایش‌اند، برای مدت مشخصی زیر درمان با داروی مورد بررسی قرار می‌گیرند. در پایان، همه‌ی موش‌ها کشته شده، و وزن کلیه‌ی هر یک از آن‌ها بر حسب میلی‌گرم اندازه‌گیری می‌شود. تأثیر درمان بر وزن کلیه‌ها به شرح زیر بوده است:

گروه	شمار	میانگین	انحراف‌معیار
گروه شاهد	۱۲	۳۱۸	۱۰/۲
گروه تجربه	۱۲	۳۷۰	۲۴/۱

حال بیایید ببینیم تفاوت دیده شده بین وزن کلیه‌های دو گروه، معنی‌دار بوده است یا نه. برای محاسبه‌ی خطای معیار تفاوت دو میانگین از این فرمول به شرح زیر استفاده می‌شود:

$$\text{میان میانگین‌ها } S.E.(d) = \sqrt{\frac{\sigma_۱^۲}{n_۱} + \frac{\sigma_۲^۲}{n_۲}}$$

$$= \sqrt{\frac{(۱۰/۲)^۲}{۱۲} + \frac{(۲۴/۱)^۲}{۱۲}} = \sqrt{۸/۶۷ + ۴۵/۴} = \sqrt{۵۷/۷۰} = ۷/۵$$

به این ترتیب خطای‌معیار تفاوت دو میانگین ۷/۵ به‌دست می‌آید. تفاوت واقعی میان میانگین‌ها ۵۲=(۳۱۸-۳۷۰) می باشد که بیشتر از دو برابر خطای معیار تفاوت دو میانگین است و بنابراین "معنی‌دار" می باشد. پس نتیجه می گیریم درمان بر وزن کلیه‌ها موثر بوده است.

خطای‌معیار تفاوت دو نسبت

- گاهی اوقات به جای تفاوت دو میانگین ، معنی داری تفاوت بین دو نسبت یا کسر (ratio) آزمون می شود که آیا تفاوت قابل مشاهده بین دو نسبت یا کسر شانسی است یا نه. در این موارد *خطای‌معیار تفاوت دو نسبت* را محاسبه می کنیم.

فرض کنید قصد داریم در یک کارآزمایی میدانی دو نوع واکسن سیاه سرفه را با هم مقایسه کنیم. نتایج مطالعه بدین شرح است:

واکسن	شمار افراد واکسینه	شمار افراد مواجهه یافته	شمار موارد بیماری	میزان حمله[1]
آ	۲۴۰۰	۹۰	۲۲	۲۴/۴٪
ب	۲۳۰۰	۸۶	۱۴	۱۶/۲٪

از نتایج مزبور این گونه به نظر می‌رسد که واکسن (ب) بهتر از واکسن (آ) است. آیا این تفاوت، معنی‌دار است؟
برای پاسخ به این پرسش ابتدا خطای معیار تفاوت دو نسبت را به ترتیب زیر محاسبه می‌کنیم:

$$\text{تفاوت دو نسبت}\ S.E. = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$= \sqrt{\frac{۲۴/۴ \times ۷۵/۶}{۹۰} + \frac{۱۶/۲ \times ۸۳/۸}{۸۶}} = \sqrt{۲۰/۴۹ + ۱۵/۷۸} = \sqrt{۳۶/۲۷} = ۶/۰۲$$

خطای معیار تفاوت دو نسبت ۶ است در حالی‌که تفاوت مشاهده شده (۲۴/۴–۱۶/۲) یعنی ۸/۲ است. به عبارت دیگر تفاوت دیده شده بین دو گروه کمتر از دو برابر خطای معیار تفاوت بین دو نسبت یعنی ۶×۲ است. پس استنتاج می‌شود که شواهد قوی دال بر تفاوت میان تأثیر دو واکسن وجود ندارد. بنابراین تفاوت دیده شده می‌تواند شانسی باشد.

در این موردها از آزمون مجذورکای (χ^2) نیز می‌توان بهره‌برداری نمود.

آزمون مجذور کای (CHI-SQUARE)

آزمون مجذور کای (χ^2) روشی جایگزین برای آزمون تفاوت بین دو نسبت است. مزیتش این است که برای انجام مقایسه نسبت‌ها در بیش از دو گروه نیز قابل استفاده است.

همان نمونه پیش را در نظر بگیرید: دو واکسن سیاه سرفه را در یک کارآزمایی میدانی با هم مقایسه کرده‌ایم. نتایج بدین قرار هستند:

واکسن	شمار افرادی که دارای حمله‌ی بیماری بوده‌اند	شمار افرادی که بدون حمله‌ی بیماری بوده‌اند	جمع	میزان حمله
آ	۲۲	۶۸	۹۰	۲۴/۴٪
ب	۱۴	۷۲	۸۶	۱۶/۲٪
جمع	۳۶	۱۴۰	۱۷۶	–

در ظاهر ملاحظه می‌شود که واکسن (ب) بهتر از واکسن (آ) بوده‌است. اما این پرسش وجود دارد که آیا واکسن ب از واکسن آ بهتر است، یا تفاوت مشاهده‌پذیر فقط از روی شانس بوده‌است.

واکسن	شمار افراد واکسینه	شمار افراد مواجهه یافته	شمار موارد بیماری	میزان حمله[1]
آ	۲۴۰۰	۹۰	۲۲	۲۴/۴٪
ب	۲۳۰۰	۸۶	۱۴	۱۶/۲٪

۱. آزمون فرضیه‌ی صفر[1]

در ابتدا فرضیه‌ای را تنظیم می‌کنیم که به نام **فرضیه‌ی صفر** (Null Hypothesis) خوانده می‌شود. این فرضیه می‌گوید که هیچ تفاوتی بین دو واکسن وجود ندارد. فرضیه مذکور باید به‌گونه کمی آزمون شود. آزمون به شکل زیر اجرا می‌شود:

ابتدا نتایج مربوط به دو واکسن را گردآوری می‌کنیم. نسبت کسانی که مورد حمله بیماری قرار گرفته‌اند ۳۶÷۱۷۶ = ۰/۲۰۴ و نسبت کسانی که مورد حمله بیماری قرار نگرفته‌اند ۱۴۰÷۱۷۶= ۰/۷۹۵ است.

با بهره‌گیری از این نسبت‌ها، شمار مورد انتظار بیماران یا همان موارد حمله و نیز شمار مورد انتظار کسانی که دچار حمله‌ی بیماری نشده‌اند را می‌توان محاسبه کرد. به این ترتیب که برای واکسن آ، شمار مورد انتظار برای کسانی که دارای حمله‌ی بیماری بوده‌اند ۱۸/۳۶ =۰/۲۰۴ × ۹۰ و برای کسانی که حمله‌ی بیماری نداشتند ۷۱/۵۵ =۰/۷۹۵ × ۹۰ محاسبه می‌شود. به همین صورت برای واکسن ب، شمار بیماران مورد انتظار ۱۷/۵۴۴ =۰/۲۰۴ × ۸۶ و شمار کسانی که حمله‌ی بیماری نداشته‌اند ۶۸/۳۷ =۰/۷۹۵ × ۸۶ برآورد می‌شود. حال می‌توان جدول پیشین را دوباره نوشت. در این جدول شمار مشاهده‌ها (O)، موارد انتظار (E) و تفاضل این دو مقدار $(O-E)^2$ در هر خانه از جدول آمده است.

واکسن	افرادی که دارای حمله‌ی بیماری بوده‌اند	افرادی که حمله‌ی بیماری نداشتند
آ	O = ۲۲	O =۶۸
	E =۱۸/۳۶	E = ۷۱/۵۵
	O - E =+۳/۶۴	O-E = −۳/۵۵
ب	O =۱۴	O =۷۲
	E = ۱۷/۵۴	E = ۶۸/۳۷
	O-E = −۳/۵۴	O-E = +۳/۶۳

۲. به‌کارگیری آزمون مجذور کای

$$\chi^{۲} = \frac{\Sigma(O-E)^{۲}}{E}$$

$$= \frac{(۳/۶۴)^{۲}}{۱۸/۳۶} + \frac{(۳/۵۵)^{۲}}{۷۱/۵۵} + \frac{(۳/۵۴)^{۲}}{۱۷/۵۴} + \frac{(۳/۶۳)^{۲}}{۶۸/۳۷}$$

$$= ۰/۷۲ + ۰/۱۷ + ۰/۷۱ + ۰/۱۹ = ۱/۷۹$$

۳. یافتن درجه‌ی آزادی[۱]

گام بعدی، محاسبه درجه‌ی آزادی (d.f.) است. درجه‌ی آزادی (Degree freedom) به شمار سطرها و ستون‌های جدول اصلی بستگی دارد و از فرمول زیر به‌دست می‌آید:

$$d.f.(c-۱)(r-۱)$$

شمار ستون‌ها = C　　شمار سطرها = r

جدول مزبور دو ستون (افرادی که دارای حمله‌ی بیماری بوده‌اند و افرادی که بدون حمله‌ی بیماری بوده‌اند) و دو سطر (واکسن آ و واکسن ب) دارد. این جدول یک جدول توافقی ۲×۲ است و درجه‌ی آزادی آن برابر است با:

$$d.f. = (c-۱)(r-۱) = (۲-۱)(۲-۱) = ۱$$

استفاده از جدول مجذور کای

۳. جداول احتمال

مرحله‌ی بعدی مراجعه به جدول احتمال است که برای این منظور از قبل آماده شده و نمونه‌ای از آن در آخر همین فصل آورده شده است. با مراجعه به این جدول می‌بینیم در صورتی که درجه‌ی آزادی ۱ باشد، مقدار χ^2 برای احتمال ۰/۰۵، برابر با ۳/۸۴ است. از آن جایی که مقدار مشاهده شده (۱/۷۹) بسیار کوچک‌تر از ۳/۸۴ است، چنین نتیجه‌گیری می‌شود که فرضیه صفر درست است و واکسن ب از واکسن آ بهتر نیست. این آزمون فقط هنگامی معتبر است که هیچ‌یک از اعداد مورد انتظار در خانه‌های جدول از عدد ۲ کمتر نباشد.

Probability (P)

D.F.	.50	.10	.05	.02	.01	.005	.001
1.	0.45	2.71	3.84	5.41	6.64	7.88	10.83
2.	1.39	4.61	5.99	7.82	9.21	10.60	13.82
3.	2.37	6.25	7.82	9.84	11.34	12.64	16.27
4.	3.36	7.78	9.49	11.67	13.28	14.86	18.47
5.	4.35	9.24	11.07	13.39	15.09	16.75	20.51
6.	5.35	10.65	12.59	15.03	16.81	18.55	22.46
7.	6.35	12.02	14.07	16.62	18.48	20.28	24.32
8.	7.34	13.36	15.51	18.17	20.09	21.96	26.13
9.	8.34	14.68	16.92	19.68	21.67	23.59	27.88
10.	9.34	15.99	18.31	21.16	23.21	25.19	29.59

مفهوم همبستگی

- گاهی اوقات می‌خواهیم بدانیم آیا بین دو متغیر مثلا وزن و قد ، دمای بدن و ضربان نبض، سن و ظرفیت حیاتی ریه و مواردی مانند این‌ها ارتباط خطی وجود دارد یا خیر.

- برای پاسخ به این سوال که آیا بین دو متغیر (مثلا x و y) ارتباط معنی‌داری وجود دارد یا نه ، باید ***ضریب همبستگی (Coefficient Correlation)*** را محاسبه کنیم که با حرف "r" نمایش داده می شود.

- فرض کنید دو متغیر به‌نام‌های x و y داریم که هر دوی آن‌ها برای افراد یک نمونه به حجم nبررسی شده‌اند. ضریب همبستگی با استفاده از فرمول ذیل محاسبه می‌شود:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2(y-\bar{y})^2}}$$

تفسیر ضریب همبستگی

- مقدار ضریب همبستگی بین ۱+ و ۱- تغییر می کند .
- اگر ضریب همبستگی به ۱+ نزدیک‌تر باشد، نشان می‌دهد رابطه‌ی بین x و y مثبت و قوی است، به‌این معنی که اگر یکی از متغیرها افزایش یابد دیگری هم افزایش می‌یابد.
- اگر نزدیک به ۱- باشد ، نشان دهنده‌ی رابطه‌ی منفی و قوی بین دو متغیر است ؛ یعنی با افزایش یکی دیگری کاهش می‌یابد.
- اگر r=۰ باشد ، نشان‌دهنده‌ی این است که بین x و y رابطه‌ی خطی وجود ندارد.
- آزمون‌هایی وجود دارند که با استفاده از آن‌ها می‌توان نشان داد آیا همبستگی مشاهده شده شانسی است یا نه.

لازم است اشاره شود وجود همبستگی الزاما علیت (Causation) را ثابت نمی کند.

مفهوم رگرسیون

در مورد یک فرد اگر بخواهیم به مقدار نامعلوم یک متغیر [متغیر وابسته]، از روی مقدار معلوم یک متغیر دیگر [متغیر مستقل] پی ببریم، باید ضریب رگرسیون[۲] یکی را نسبت به دیگری حساب کنیم. به‌گونه قراردادی متغیر مستقل را با x و متغیر وابسته را با y نشان می‌دهند. فرمول محاسبه ضریب رگرسیون به شرح زیر است:

$$y = \bar{y} + b(x - \bar{x})$$

که در آن:

$\bar{y} =$ میانگین $y_1 + y_2 + y_3 \ldots\ldots y_n$

$\bar{x} =$ میانگین $x_1 + x_2 + x_3 \ldots\ldots x_n$

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

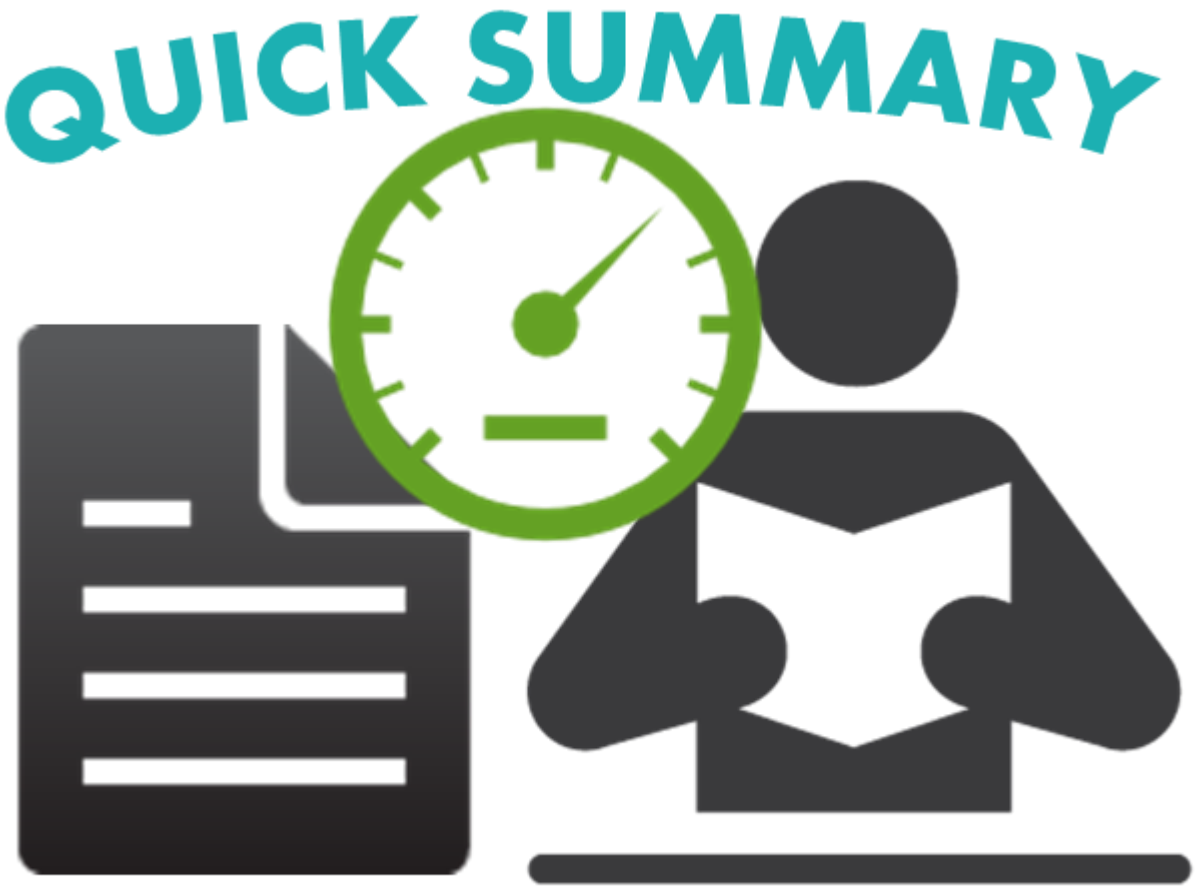
مقدار b به نام ضریب رگرسیون y نسبت به x خوانده می‌شود رگرسیون x نسبت به y را نیز با همین روش می‌توان حساب کرد:

$$x = \bar{x} + b^1(y - \bar{y})$$

$$b^1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

که درآن b^1 ضریب رگرسیون x نسبت به y است. کاربرد رگرسیون، محاسبه‌ی میانگین برآوردهای مقدار یک متغیر از روی مقدار متغیر دیگر است.

خلاصه مطالب



- در هر پژوهش، ابتدا داده‌ها جمع‌اوری می‌شوند. داده‌هایی که مستقیم از افراد مورد بررسی به‌دست می‌اینـد، داده‌های ابتدایی و آن‌هایی که از منابع خارجی به‌دست می‌آیند، داده‌های بعدی(دومین داده‌ها) نامیده می‌شوند. نخستین داده‌ها معمولاً از داده‌های بعدی دقیق‌تر هستند. پس از جمع‌آوری داده‌ها، بایستی آن‌ها را هدفمندانه مرتب و خلاصه نمود تا استخراج نکته‌های مهم به‌گونه‌ی شفاف و کامل از آن‌ها ممکن شود.

- با جدول می‌توان به‌سادگی انبوهی از داده‌های آماری را به نمایش گذارد. در یک جدول توزیع فراوانی، ابتدا داده‌ها به گروه‌های مناسب تقسیم شده و شمار افراد هر گروه، در ستون کنار آن نمایش داده می‌شود. نمودارها و طرح‌ها در مقایسه با جدول‌ها ساده‌ترند و تأثیر نیرومندتری بر ذهن می‌گذارند. از میان آن‌ها می‌توان به انواع نمودارهای ستونی، دایره‌ای، هیستوگرام، چندگوشه‌ی فراوانی، طرح نقاط پراکنده و طرح‌های خطی و تصویری اشاره کرد. وقتی داده‌های آماری مربوط به مناطق جغرافیایی باشند، می‌توان آن‌ها را با بهره‌برداری از نقشه‌های هاشور زده یا نقطه‌گذاری شده نیز نشان داد.

- انواع گوناگونی از معدل‌های آماری وجود دارد که رایج‌ترین آن‌ها میانگین، میانه و نما هستند. برای محاسبه‌ی میانگین ابتدا اندازه‌ی همه مشاهده‌ها با هم جمع و سپس مجموع به‌دست آمده بر کل مشاهدات بخش می‌شود. برای به‌دست آوردن میانه، ابتدا داده‌ها بر اساس مقدارشان مرتب می‌شوند و سپس مشاهده‌ای که از نظر مقدار در وسط قرار گرفته‌است، میانه نام می‌گیرد. نما به پرتکرارترین مشاهده در یک توزیع داده‌ای نسبت داده می‌شود.

- چندین مقدار برای اندازه‌گیری پراکندگی وجود دارد که شناخته شده‌ترین آن‌ها دامنه، میانگین انحرافات و انحراف معیار هستند. دامنه عبارت است از اختلاف بین بیشترین و کمترین مقدارها در یک نمونه. میانگین انحرافات، معدل مقادیر انحرافات هر مشاهده از میانگین داده‌هاست. انحراف معیار که رایج‌ترین شاخص پراکندگی به شمار می‌آید، ریشه‌ی دوم میانگین مجذور انحراف‌ها است.

- توزیع یا منحنی نرمال با شکل متقارن و زنگوله مانند خود یکی از مفاهیم مهم و پرکاربرد آماری به شمار می‌رود. در این منحنی میانگین، میانه و نما روی هم می‌افتند. مساحت محدود بین یک، دو و سه انحراف معیار از طرفین میانگین به ترتیب تقریبا ۶۸، ۹۵ و ۹۹/۷٪ از مقادیر را در توزیع شامل می‌شود. حدود مزبور در طرفین میانگین را حدود اطمینان می‌نامند.

- منحنی نرمال استاندارد نوع خاصی از توزیع نرمال است که در آن مساحت کل زیر منحنی برابر با یک، میانگین آن صفر و انحراف معیار آن یک است. در این منحنی، فاصله‌ی میانگین از یک مقدار بخش بر انحراف معیار، متغیر نرمال استاندارد نامیده می‌شود.

- وقتی قرار باشد شمار زیادی از مردم یا اشیا یا هر واحد دیگر مورد بررسی قرار گیرند، نمونه‌گیری می‌کنیم. چارچوب نمونه‌گیری فهرستی از افراد جمعیت مورد نظر است که قرار است نمونه از بین آن‌ها انتخاب شود. روش نمونه‌گیری باید به‌گونه‌ای باشد که در آن نمونه‌ی انتخاب شده نماینده خوبی از کل جمعیت مورد مطالعه باشد. رایج‌ترین روش‌های نمونه‌گیری عبارت‌اند از تصادفی ساده، تصادفی منظم و تصادفی طبقه‌بندی شده.

- اگر از یک جمعیت همواره نمونه‌گیری شود، نتیجه‌ی بررسی بر پایه هر نمونه متفاوت خواهد بود. به این تفاوت خطای نمونه‌گیری می‌گویند که تحت تأثیر حجم نمونه و تفاوت‌های طبیعی موجود میان اندازه‌گیری‌ها است. اگر از یک جمعیت با یک حجم معین بارها نمونه‌گیری انجام شود، هر نمونه میانگین خاص خود را خواهد داشت که توزیع آن‌ها نرمال و میانگین آن‌ها برابر با میانگین خود جمعیت خواهد بود.
- انحراف معیار میانگین‌های نمونه‌ای که معیاری برای سنجش خطای نمونه‌گیری است، خطای معیار نامیده می‌شود. با برآورد خطای معیار برای میانگین، نسبت، تفاوت دو میانگین و یا تفاوت دو نسبت، می‌توان برای این مقدارها آزمون آماری انجام داد. آزمون مجذورکای روشی جایگزین برای آزمون تفاوت بین دو نسبت است که برای انجام مقایسه نسبت‌ها در بیش از دو گروه نیز قابل بهره‌برداری است.
- برای بررسی ارتباط خطی میان دو متغیر باید ضریب همبستگی آن‌ها را محاسبه کنیم که مقدار آن بین ۱+ و ۱- تغییر می‌کند. اگر ضریب همبستگی به ۱+ نزدیک‌تر باشد، نشان می‌دهد رابطه‌ی بین دو متغیر مثبت و توانمندتر است، به این معنی که اگر یکی از متغیرها افزایش یابد دیگری هم افزایش می‌یابد. اگر نزدیک به ۱- باشد، نشان دهنده‌ی رابطه‌ی منفی و توانمند بین دو متغیر است. ضریب همبستگی صفر به معنای نبود همبستگی خطی میان آن دو است.
- در مورد یک فرد اگر بخواهیم به مقدار نامعلوم یک متغیر، از روی مقدار معلوم یک متغیر دیگر پی ببریم، باید ضریب رگرسیون یکی را نسبت به دیگری حساب کنیم. کاربرد رگرسیون، محاسبه‌ی میانگین برآوردهای مقدار یک متغیر از روی مقدار متغیر دیگر است.

منبع:

درس نامه پزشکی پیشگیری و پزشکی اجتماعی

✓تالیف ک. پارک

- ویراست ۲۱
- فصل سوم جلد اول با عنوان:

"اطلاعات سلامت و آمار پایه پزشکی"