تحلیل اکتشافی داده‌ها در مطالعات کمی علوم ورزشی

با کمک نرم افزار SPSS

حسین پورسلطانی زرندی
دانشیار دانشگاه پیام نور

ناشر نمونه کشور در سال ۹۳

فهرســت

فصل اول : تحلیل اکتشافی داده ها **۶**

تک متغیره غیر گرافیکی ۷

داده‌های طبقه بندی شده ۷

ناهمگونی متغیرهای جمعیت شناختی ۸

مشکل(ترکیب) سیب‌ها و پرتقال‌ها ۱۰

ویژگی‌های داده‌های کمی ۱۰

فصل دوم : داده‌های مفقود و نحوه حذف یا جایگزین کردن مقادیر به جای آنها **۱۱**

فصل سوم : تشخیص داده‌های پرت از طریق نمرات استاندارد Z **۲۰**

نرمال بودن بر اساس ضرایب کجی و کشیدگی ۲۲

کجی منفی ۲۴

کجی مثبت ۲۶

کشیدگی منفی ۲۷

آزمون‌های کلموگروف اسمیرنف و شاپیرو – ویلک ۲۹

فصل چهارم : نرمال بودن چند متغیره **۳۷**

شناسایی پرت بودن داده‌های چند متغیره ۴۰

ارزیابی گرافیکی نرمال بودن چند متغیره ۴۴

فصل پنجم : روش‌های نرمال کردن متغیرها **۴۷**

مجذور کردن یا به توان دو رساندن ۵۴

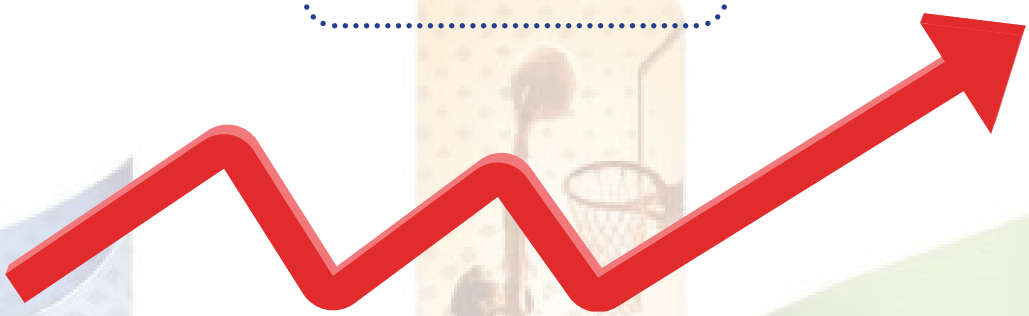
فصل ششم: میانگین حسابی، میانگین پیراسته و برآوردهای‌های میانگین **۶۱**
فصل هفتم : همگنی واریانس ها **۶۴**
فصل هشتم : نمودار هیستوگرام **۶۷**
نمودار جعبه ای ۷۰
نمودار ساقه و برگ ۷۸
نمودار چندک-چندک یا نمودار کیو-کیو ۸۲
نمودارهای پراکنش یا پراکندگی ۸۴
نمودار پراکنش نقطه ای ساده ۸۸
نمودار پراکنش ماتریسی ۹۰
نمودار پراکنش سه بعدی ۹۲
نمودار پراکنش پوششی ۹۳
پیوست الف: جدول بحرانی مقادیر آزمون خی - دو ۹۵
منابع: ۹۶

پیشگفتار

جهت انجام تحقیقات و به خصوص تحقیقات علوم ورزشی، لزوم بکارگیری آزمون‌های آماری پارامتریک و غیر پارامتریک مورد توجه قرار گرفت. انجام هر کدام از آزمون‌های آماری مستلزم وجود داده هاست. داده‌ها به دو صورت کمی و یا کیفی جمع آوری می‌شوند. در مطالعات با رویکرد کمی، نکات مهمی همچون ابزارهای دقیق، آزمونگرهای حرفه ای، داده‌های کافی، ورود صحیح و کامل داده‌ها به نرم افزارهای آماری از ضروریات است. برای بررسی داده‌ها از جنبه‌های مختلف، رویکرد تحلیل اکتشافی داده‌ها به صورت عددی، گرافیکی و یا آزمون‌های نرمال بودن بکار گرفته می‌شود. در این کتاب موضوعاتی همانند داده‌های مفقود و نحوه حذف یا جایگزین کردن مقادیر به جای آنها، ناهمگونی متغیرهای جمعیت شناختی، تشخیص داده‌های پرت از طریق نمرات استاندارد Z ، نرمال بودن بر اساس ضرایب کجی و کشیدگی، آزمون نرمال بودن(کلموگروف اسمیرنف، شاپیرو – ویلک، اندرسون دارلینگ)؛ نرمال بودن چند متغیره، شناسایی پرت بودن داده‌های چند متغیره، ارزیابی گرافیکی نرمال بودن چند متغیره، روش‌های نرمال کردن متغیرها(متغیر بودن یا ثابت ماندن شاخص‌های توصیفی میانگین و انحراف استاندارد(بر اساس تبدیل به رتبه))؛ انواع میانگین ها(حسابی، پیراسته، اندرو، توکی، همپل و هوبر)، همگنی واریانس بر اساس آزمون لون و با چهار رویکرد، نمودارهای(هیستوگرام، جعبه ای، ساقه و برگ، چندک–چندک یا نمودار کیو–کیو و پراکنش یا پراکندگی) توضیح داده شده اند. امید بر این است تا محققین با بکارگیری موارد مطروحه، نسبت به انجام درست تحقیق خود، موفق باشند.

دکتر حسین پورسلطانی زرندی

فصل اول



تحلیل اکتشافی داده ها

تحلیل اکتشافی داده ها[1] به بررسی داده‌ها قبل از انجام روش‌های آماری مربوط به آزمون فرضیه‌ها می‌پردازد. اولین بار در سال ۱۹۷۷ جان توکی[2] در مورد آن مطالبی را ارایه کرد.

اهداف

- کنترل داده‌ها از منظر نداشتن داده‌های مفقود[3]
- کنترل داده‌ها از منظر نداشتن داده‌های پرت و یا انتهایی
- انتخاب مناسب ترین روش یا روش‌های آماری
- فراهم کردن پایه ای برای جمع آوری احتمالی داده‌های بیشتر(تعداد داده ناکافی)
- تشخیص تدوین بهترین فرضیه‌های آماری
- کنترل داده‌ها به هنگام ورود به نرم افزارهای آماری

تحلیل اکتشافی داده‌ها به روش‌های مختلفی انجام می‌گیرد. روش‌های غیر گرافیکی، معمولاً شامل محاسبه آمار می‌باشند، در حالی که روش‌های گرافیکی، آشکارا داده‌ها را به صورت نموداری یا تصویری خلاصه می‌کنند. روش‌های تک متغیره در یک زمان به یک متغیر (ستون داده) نگاه می‌کنند، در حالی که روش‌های چند متغیره برای بررسی روابط به دو یا چند متغیر در یک زمان نگاه می‌کنند. معمولاً تحلیل اکتشافی داده‌های چند متغیره، شامل دو متغیره خواهد بود (دقیقاً به دو متغیر نگاه می‌شود)، اما گاهی اوقات شامل سه یا چند متغیر هم می‌شود. تقریباً همیشه ایده خوبی است که قبل از اجرای تحلیل اکتشافی داده‌های چند متغیره، تحلیل اکتشافی داده‌ها به صورت تک متغیره روی هر یک از متغیرها انجام گیرد. به طور کلی تحلیل اکتشافی داده‌ها از زوایای؛ تک متغیره غیر گرافیکی، چند متغیره غیرگرافیکی، گرافیکی تک متغیره، گرافیکی چند متغیره و ارزیابی مبتنی بر نظریه(آزمون کردن نرمال بودن داده ها) قابل بررسی است.

تک متغیره غیر گرافیکی

داده‌هایی که از انجام یک اندازه‌گیری خاص بر روی همه آزمودنی‌های یک نمونه به دست می‌آیند، مشاهدات محقق را برای یک ویژگی واحد مانند سن، جنسیت، سرعت در انجام یک کار یا پاسخ به یک محرک نشان می‌دهند. محقق باید این اندازه‌گیری‌ها را به‌عنوان نشان‌دهنده یک توزیع نمونه از متغیر در نظر بگیرد، که به نوبه خود کم و بیش نشان‌دهنده توزیع جامعه متغیر است. هدف معمول از تحلیل اکتشافی داده‌های غیر گرافیکی تک متغیره، درک بهتر توزیع نمونه است و همچنین نتیجه گیری آزمایشی در مورد اینکه چه توزیع(های) جامعه با توزیع نمونه سازگار است. تشخیص داده‌های پرت نیز بخشی از این تحلیل است.

داده‌های طبقه بندی شده

ویژگی‌های مورد علاقه برای یک متغیر طبقه بندی به محدوده مقادیر یا فراوانی، فراوانی نسبی یا درصدی وقوع برای هر مقدار است. (برای متغیرهای رتبه ای گاهی اوقات مناسب است که با استفاده از تکنیک هایی، آنها را به عنوان متغیرهای کمی در نظر گرفت). بنابراین تنها شیوه غیر گرافیکی تک متغیره مفید برای متغیرهای طبقه‌بندی، نوعی جدول‌بندی فراوانی است، معمولاً همراه با محاسبه نسبت یا درصد است.فرض کنید از تعدادی آزمودنی وضعیت مقطع تحصیلی آنان سوال شده است. توزیع فراوانی و درصدی در جدول(۱-۱)ارایه گردیده است:

1. Exploratory Data Analysis 2. John Tukey 3. Missing

جدول۱-۱.توزیع فراوانی مقاطع تحصیلی

شاخص مقطع تحصیلی	فراوانی	درصد
دانشجوی کارشناسی	۶۰	۶۰
دانشجوی کارشناسی ارشد	۳۰	۳۰
دانشجوی دکتری	۱۰	۱۰
مجموع	۱۰۰	۱۰۰

بر طبق جدول(۱-۱)درصدهای مربوط به مقاطع تحصیلی دانشجوی کارشناسی، دانشجوی کارشناسی ارشد و دانشجوی دکتری به ترتیب برابر با ۶۰، ۳۰ و ۱۰ می باشد. توجه داشته باشید که داشتن تعداد کل فراوانی مفید است تا بررسی شود که برای هر موضوعی که انتخاب شده، یک مشاهده وجود دارد(از دست دادن داده‌ها یک اشتباه رایج است و تحلیل اکتشافی داده‌ها برای یافتن اشتباهات، بسیار مفید است). همچنین، باید انتظار داشت که اگر به درستی آنها را محاسبه کنیم مجموع نسبت‌ها به ۱ و درصد به ۱۰۰می رسد. معمولا بعد از بیان فراوانی ها، گزارش یکی از آنها(نسبت یا درصد) کفایت می‌کند.

ناهمگونی[1] متغیرهای جمعیت شناختی

یکی از نکات مهم در نمونه گیری، رعایت نسبت متغیرهای جمعیت شناختی[2] در نمونه است. در واقع اگر محقق به روی فرضا دانشجویان تربیت بدنی یک دانشگاه(اعم از دانشجویان کارشناسی و ارشد و دکتری) تحقیق می‌کند، باید نسبت سه مقطع، در نمونه انتخابی را هم توجه کند. بی توجهی به وجود همه نوع از متغیرهای جمعیت شناختی در نمونه، تعمیم پذیری یافته‌ها را با مشکل مواجه می‌سازد. در این زمینه محقق می‌تواند از شاخص ناهمگونی بلاو[3] استفاده کند. فرض کنید تعداد نمونه به شرح زیر است:

جدول۲-۱.

	دانشجویان کارشناسی	دانشجویان ارشد	دانشجویان دکتری	تعداد کل
تعداد	۶۰	۳۰	۱۰	۱۰۰

برای محاسبه شاخص ناهمگونی بلاو از مراحل زیر کمک بگیرید:

نسبت هرکدام از ۳ مقطع تحصیلی را محاسبه کنید(تعداد هر مقطع تقسیم بر مجموع):

جدول۳-۱.

	دانشجویان کارشناسی	دانشجویان ارشد	دانشجویان دکتری	مجموع
تعداد	۶۰	۳۰	۱۰	۱۰۰
نسبت	۰/۶۰	۰/۳۰	۰/۱۰	۱

محاسبه مجذور هر نسبت:

1. Heterogeneity
2. Demography
3. Blau›s index of Heterogeneity

جدول۴-۱.

	دانشجویان کارشناسی	دانشجویان ارشد	دانشجویان دکتری	مجموع
تعداد	۶۰	۳۰	۱۰	۱۰۰
نسبت	۰/۶۰	۰/۳۰	۰/۱۰	۱
مجذور نسبت	۰/۳۶	۰/۰۹	۰/۰۱	۰/۴۶

برای محاسبه شاخص ناهمگونی بلاو، مجموع مجذور نسبت‌ها را از مقدار ۱ کسر کنید:

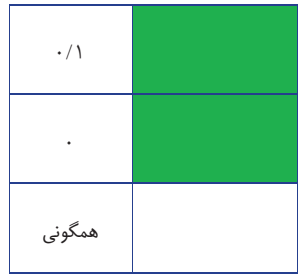
- تفسیر شاخص ناهمگونی بلاو محاسبه شده:
- مقدار بدست آمده هرچه به مقدار صفر نزدیک تر باشد، همگونی[1] بیشتر است.
- مقدار بدست آمده هرچه به مقدار ۱ نزدیک تر باشد، ناهمگونی بیشتر است.
- در خصوص مقدار ۰/۵۴ می‌توان گفت که در محدوده مرزی به سمت ناهمگونی بودن است.

می‌توانید از راهنمای زیر برای تفسیر بهتر کمک بگیرید(۱):

جدول۵-۱.راهنمای تفسیرشاخص ناهمگونی بلاو

	ناهمگونی
	۱
	۰/۹
	۰/۸
	۰/۷
	۰/۶
	۰/۵
	۰/۴
	۰/۳
	۰/۲

1. Homogeneity



در مطالعات فرا تحلیل، رعایت همین موضوع می‌تواند به صورت زیر مد توجه محققین باشد.

مشکل(ترکیب) سیب‌ها و پرتقال‌ها[1]

یکی از انتقادها به فراتحلیل[2] اسمیت و گلاس[3] این بود که در مطالعات فراتحلیل خود، انواع مختلف روش‌های درمانی را مورد استفاده قرار دادند. انتقاد اصلی این بود که اینگونه فراتحلیل منجر به نتایج بی معنی می‌شود. به عبارتی اگر می‌خواهید تا نتیجه گیری تان در باره یک جامعه آماری(مثلاً سیب ها)باشد، و سپس در نمونه شما، پرتقال‌ها هم باشند؛ در آن صورت به دلیل ناهمگن و مختلف بودن نمونه، نتیجه گیری شما صرفاً برای سیب‌ها مخدوش است. اگر عنوان تحقیقتان میوه‌ها باشد، می‌توانید انواع میوه ها(مثلاً سیب هاو پرتقال‌ها و ...) را در نمونه داشته باشید(۲).

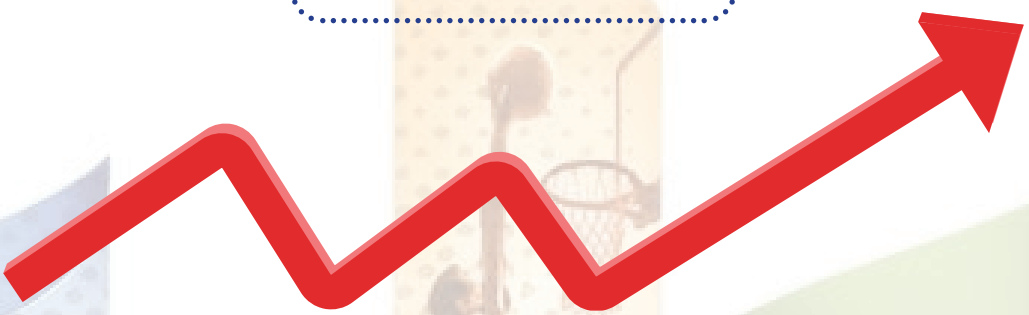
ویژگی‌های داده‌های کمی

تحلیل اکتشافی داده‌های تک متغیره برای یک متغیر کمی، راهی برای ارزیابی اولیه در مورد توزیع جامعه متغیر با استفاده از داده‌های نمونه مشاهده شده است. ویژگی‌های توزیع جامعه یک متغیر کمی شامل شاخص‌های مرکزی(میانگین، مد یا نما، میانه)، شاخص‌های پراکندگی(واریانس، انحراف استاندارد و دامنه تغییرات) است. مثلا رایج ترین معیار گرایش مرکزی، میانگین است. برای توزیع اریب یا زمانی که نگرانی در مورد نقاط پرت وجود دارد، میانه ممکن است ترجیح داده شود.

برای ارزیابی گرافیکی تک متغیره، می‌توانید از نمودارهای هیستوگرام، ساقه و برگ، جعبه ای، چندک – چندک و پراکنش‌ها استفاده کرد. برای آزمون نرمال بودن داده ها، می‌توانید از آزمون‌های کلموگروف اسمیرنف، شاپیرو – ویلک، اندرسون دارلینگ کمک گرفت.

1. Mixing Apples and Oranges
2. Meta-Analysis
3. Smith and Glass(1977)

فصل دوم



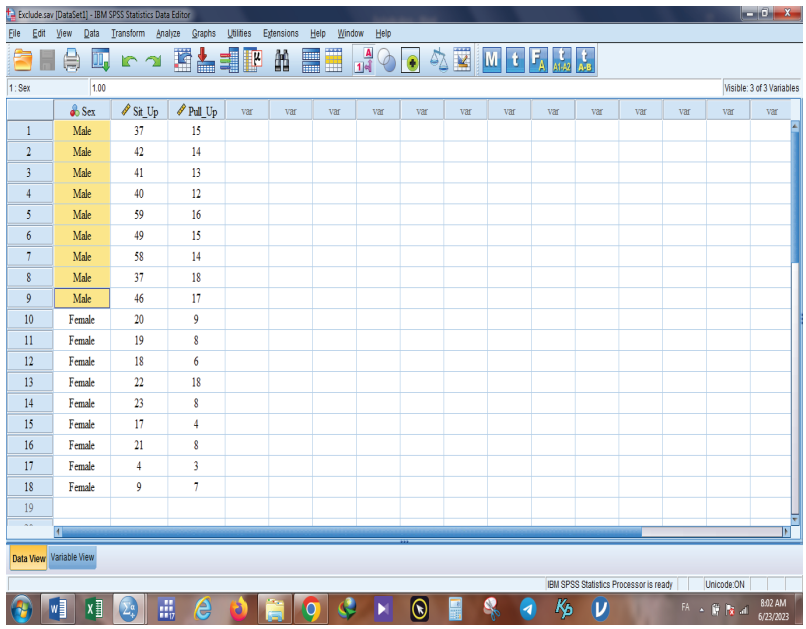
داده‌های مفقود و نحوه حذف یا جایگزین کردن مقادیر به جای آنها

همواره در جریان تحقیق، محقق با آزمودنی‌هایی مواجه می‌شود که به بعضی از سوالات پرسشنامه پاسخ نداده اند و یا آزمون‌های مهارتی را ناتمام گذاشته اند. از اینرو با داده‌های مفقود(نبودن داده برای آزمودنی) مواجه می‌شود. در این نوشتار نحوه رفتار با داده‌های مفقود[1] از منظر حذف کردن[2] و یا جایگزین کردن مورد بررسی قرار خواهد گرفت. فرض کنید از تعداد ۱۸ آزمودنی(۹ مرد و ۹ زن)آزمون‌های بارفیکس و درازونشست اندازه گیری شده است.

جدول۱-۲.توضیحات متغیرها در فایل داده ها

نام فایل	متغیر	گزینه ها	کد
Exclude.sav	جنسیت Sex	مرد Man	۱
		زن Woman	۲
	بارفیکس Pull-up	کمی است(بدون کدبندی)	
	درازونشست Sit_Up	کمی است(بدون کدبندی)	

توجه: فایل Exclude. sav باز کنید و تمرین را روی این فایل انجام دهید.



شکل ۱-۲. داده‌ها در محیط SPSS

1. Analysis
2. Descriptive Statistics
3. Explore…

۴. متغیرهای بارفیکس و درازونشست به کادر Dependent List انتقال یابد.

1. Missing Value
2. Exclude

۵. متغیر جنسیت به کادر Factor List انتقال یابد.
۶. دکمه Plots... تایید شود.
۷. گزینه Normality Plots With Test تایید شود.
۸. دکمه Options... تایید شود.
۹. گزینه Exclude Cases Listwise در حالت تایید باشد.

10. Continue
11. Ok

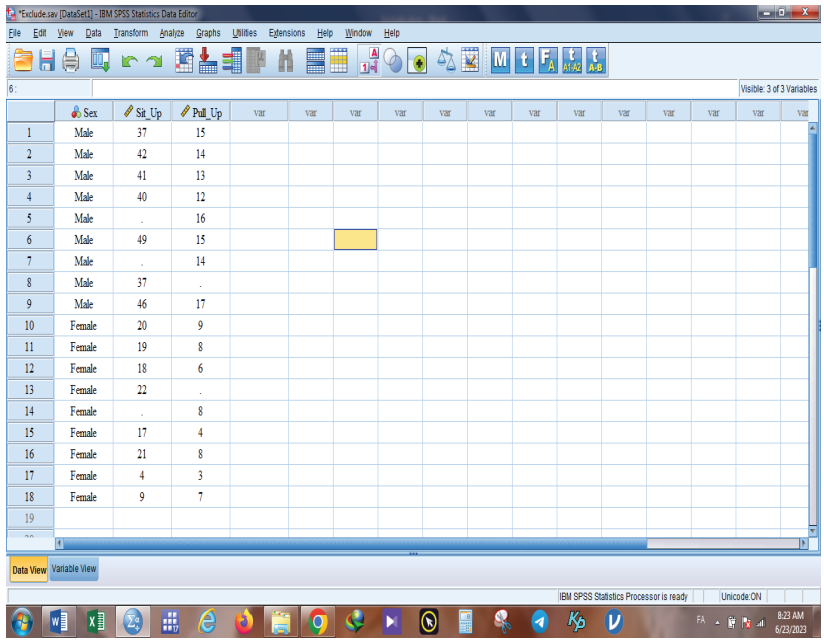
جدول۲-۲.تعداد آزمودنی‌ها در آزمون‌های کلموگروف اسمیرنف و شاپیرو-ویلک

Tests of Normality							
		Kolmogorov-Smirnova			Shapiro-Wilk		
	Sex	Statistic	df	Sig.	Statistic	df	Sig.
Sit_Up	Male	.216	9	.200*	.869	9	.119
	Female	.278	9	.044	.828	9	.043
Pull_Up	Male	.143	9	.200*	.980	9	.964
	Female	.287	9	.032	.815	9	.030
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

مشاهده می‌کنید که تعداد آزمودنی‌های مرد و زن با هم(هر کدام ۹ نفر) برابر است.

نکته: اگر مراحل بالا را مجددا انجام دهید و اینبار در مرحله نهم به جای Exclude Cases Listwise گزینه Exclude Cases Pairwise انتخاب شود، خروجی همانند جدول() بالا است.

حال داده‌های ردیف‌های ۵ و ۷ دراوزنشست مردان و ردیف ۱۴ زنان از داده‌ها و همچنین داده ردیف ۸ بارفیکس مردان و ردیف ۱۳ زنان از داده‌ها حذف می‌شوند(شکل). از اینرو در متغیر درازونشست مردان تعداد ۷ رکورد و برای زنان ۸ رکورد وجود دارد. ضمنا در متغیر بارفیکس مردان تعداد ۸ رکورد و برای زنان هم ۸ رکورد وجود دارد.



شکل ۲-۲. حذف تعدادی از داده‌ها

مراحل گفته شده را انجام دهید و فقط در مرحله نهم، Exclude Cases Listwise گزینه اختیار شود، خروجی به شرح زیر است:

جدول۳-۲. تعداد آزمودنی‌ها در آزمون‌های کلموگروف اسمیرنف و شاپیرو-ویلک با حذف تعدادی داده

Tests of Normality							
	Sex	Kolmogorov-Smirnova			Shapiro-Wilk		
		Statistic	df	Sig.	Statis-tic	df	Sig.
Sit_Up	Male	.213	6	.200*	.963	6	.841
	Female	.311	7	.039	.822	7	.067
Pull_Up	Male	.185	6	.200*	.974	6	.918
	Female	.189	7	.200*	.922	7	.482
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

مشاهده می‌شود که تعداد آزمودنی‌های مرد در متغیرهای درازونشست و بارفیکس ۶ نفر و برای زنان ۷ نفر می‌باشد.

مجددا مراحل گفته شده را انجام دهید و فقط در مرحله نهم، گزینه Exclude Cases Pairwise اختیار شود، خروجی به شرح زیر است:

جدول۴-۲.تعداد آزمودنی‌ها در آزمون‌های کلموگروف اسمیرنف و شاپیرو–ویلک با حذف تعدادی داده

Tests of Normality							
	Sex	Kolmogorov-Smirnova			Shapiro-Wilk		
		Statistic	df	Sig.	Statis-tic	df	Sig.
Sit_Up	Male	.189	7	.200*	.921	7	.476
	Female	.297	8	.036	.822	8	.049
Pull_Up	Male	.128	8	.200*	.983	8	.975
	Female	.240	8	.194	.882	8	.197
*. This is a lower bound of the true significance.							
a. Lilliefors Significance Correction							

مشاهده می‌شود که تعداد آزمودنی‌های مرد در متغیر درازونشست ۷ نفر و برای زنان ۸ نفر می‌باشد. همچنین تعداد آزمودنی‌های مرد در متغیر بارفیکس ۸ نفر و برای زنان ۸ نفر می‌باشد. نرم افزار در روش Exclude Cases Listwise در مورد مردان، تعداد داده‌های مفقود(برابر با ۳) برای هر دو متغیر درازونشست و بارفیکس و در مورد زنان، تعداد داده‌های مفقود(برابر با ۲) برای هر دو متغیر درازونشست و بارفیکس را ملاک عمل قرار می‌دهد. در روش Exclude Cases Pairwise به صورت زوجی یا جفتی عمل می‌کند. مثلا در مورد مردان، تعداد جفت داده مفقود(برابر با ۲) برای هر دو متغیر درازونشست و بارفیکس و در مورد زنان، تعداد داده جفت داده مفقود(برابر با ۱) برای هر دو متغیر درازونشست و بارفیکس را ملاک عمل قرار گرفته است(۳).

اگر بخواهید(و به خصوص وقتی تعداد داده‌ها زیاد است)، داده‌های مفقود را از داده‌های خود حذف کنید و یا داده دیگری را جایگزین آنها کنید، به توضیحات زیر توجه کنید:

حذف داده ها:

برای حذف داده‌های مفقود، در ابتدا باید آنها را شناسایی کرد و سپس اقدام به حذف نمود.

شناسایی داده‌های مفقود:

1. Data
2. Select Cases

۳. انتخاب If Condition is Satisfied

۴. تایید دکمه If...

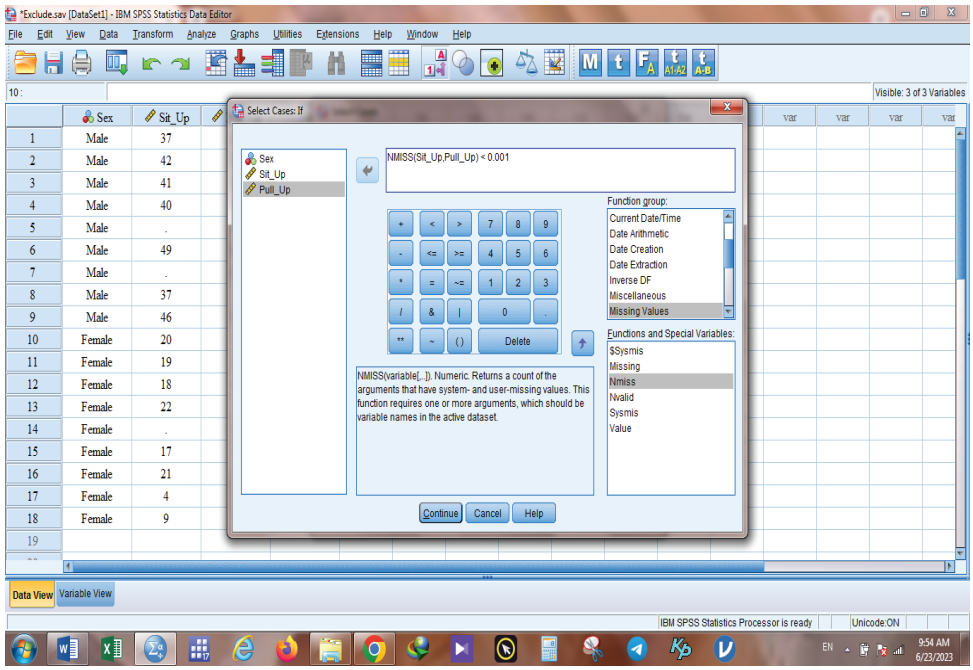
۵. از بخش Function Group گروه توابع محاسباتی Missing Values را انتخاب کنید

۶. از بخش Function and Special Variables گزینه NMISS را انتخاب و دو بار روی آن کلیک کنید تا به کادر خالی بالا منتقل شود. حال آن را به صورت NMISS(?) مشاهده می‌نمایید.

۷. متغیر درازونشست را به جای علامت بگذارید(انتقال دهید)، سپس کاما گذاشته و در نهایت متغیر بارفیکس را انتقال دهید.

۸. بعد از پرانتز علامت کوچکتر بگذارید.

۹. بعد از علامت کوچکتر، یک عدد خیلی کوچک مثلا ۰/۰۰۱ را بنویسید(شکل۱-۳).

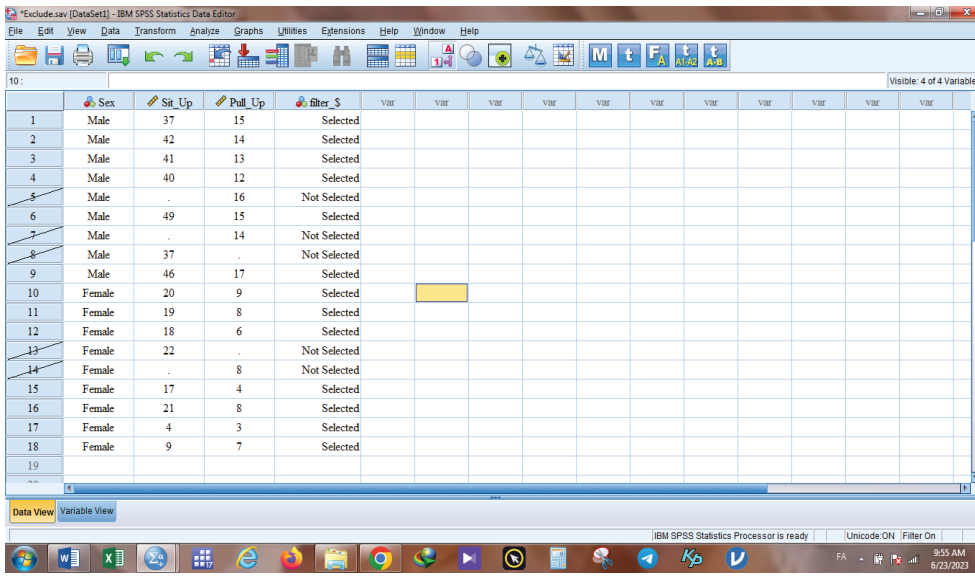


شکل ۳-۲.

۱۰. دکمه Continue را تایید کنید

۱۱. دکمه Ok را تایید کنید.

در فایل داده ها، روی ردیف‌های با داده مفقود، علامت انتخاب می‌بینید(شکل ۱-۴).



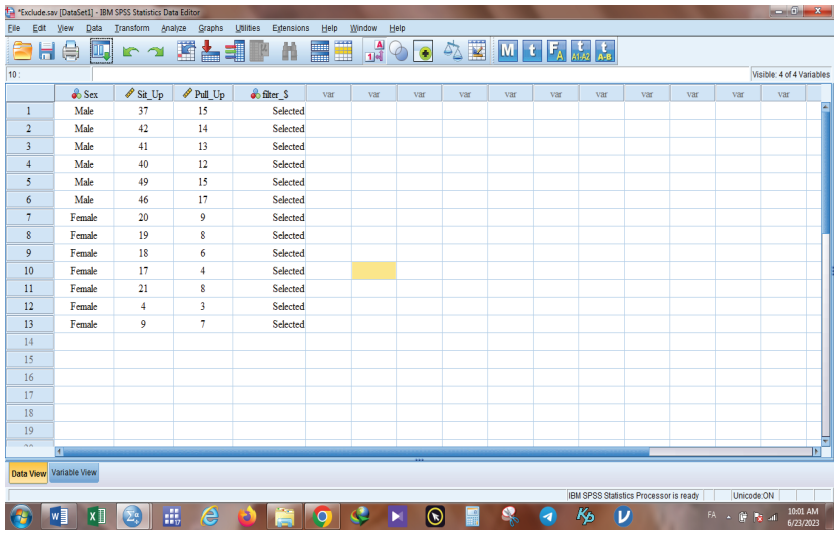
شکل ۴-۲.

اگر بخواهید داده‌های انتخاب را حذف کنید، از مراحل زیر کمک بگیرید:

1. Data
2. Select Cases

۳. انتخاب گزینه Delete Unselected Cases

۴. تایید دکمه Ok



شکل ۵-۲.

مشاهده می‌کنید که ردیف‌های با داده مفقود از مجموعه داده‌ها حذف شده اند(۴).

اگر بخواهید به جای داده‌های مفقود، داده ای جایگزین کنید، از مراحل زیر(که یکی از روش‌ها است)کمک بگیرید:

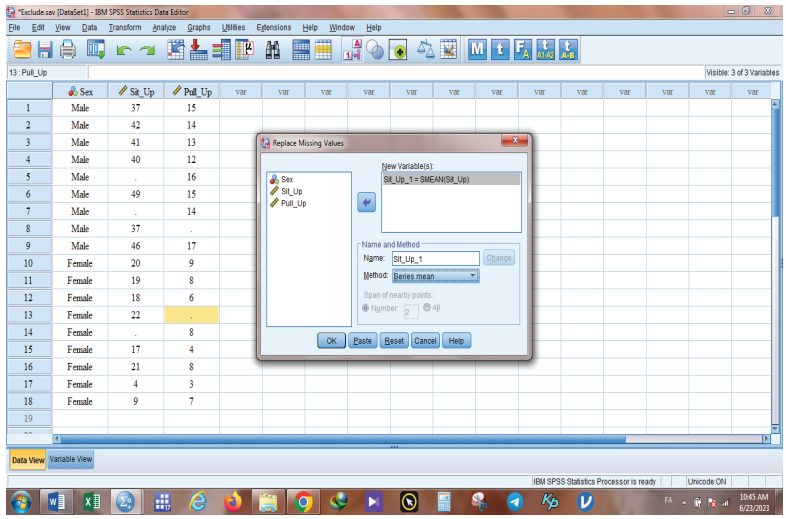
1. Transform
2. Replace Missing Values…

۳. متغیر درازونشست را به کادر New Variable(s) انتقال دهید.

۴. در کادر Name یک نام جدید بگذارید. در این مثال همان نام پیشنهادی نرم افزار بکاربرده شده است.

۵. از کادر Method روش بر پایه میانگین[1] را انتخاب کنید(شکل ۱-۶).

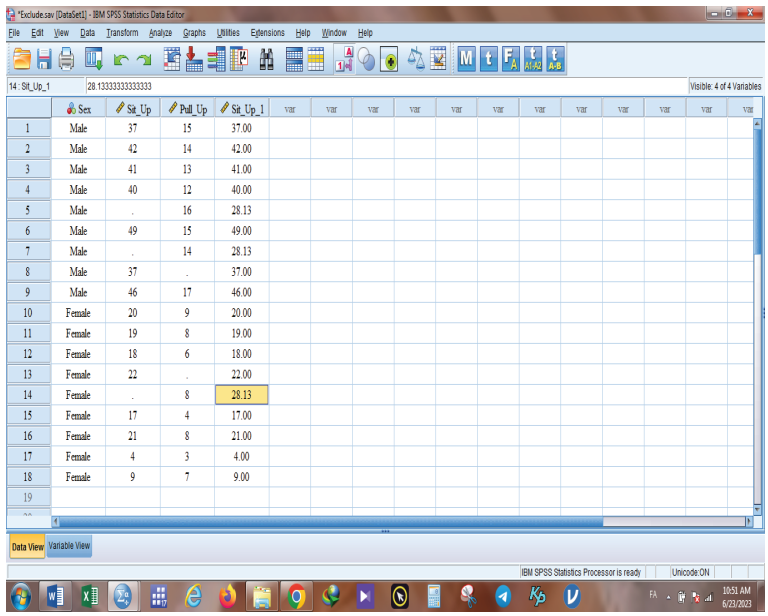
1. Series Mean



شکل ۶-۲.

۶. دکمه Ok را تایید کنید.

در داده‌ها متغیر جدید ساخته شده است. اعشار متغیر جدید را به ۲ رقم افزایش دهید. مقدار ۲۸/۱۳ جای مقادیر مفقود (در ردیف‌های ۵ و ۷ و ۱۴)را گرفته است(شکل ۱-۷).



شکل ۷-۲.

این عدد، میانگین متغیر درازونشست می‌باشد. برای اطمینان میانگین متغیر مذکور از نرم افزار خواسته شد:

جدول۵-۲.میانگین متغیر درازونشست

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Sit_Up	15	4	49	28.13	14.202
Valid N (listwise)	15				

توجه: فرآیند جایگزین مقادیر به جای داده‌های مفقود را می‌توانید برای متغیر بارفیکس هم انجام دهید(۵).

فصل سوم

تشخیص داده‌های پرت از طریق نمرات استاندارد Z