بسم الله الرحمن الرحیم

تحلیل داده‌های گمشده در SPSS

مرکز تحقیقات صداوسیما

اداره کل پژوهش‌های اجتماعی
و سنجش برنامه‌ای
گروه روش‌شناسی
پژوهشگر: علیرضا خوشگویان‌فرد

داده‌های گمشده در SPSS

معمولاً داده‌های گمشده در مجموعه داده‌ها، با خالی‌گذاشتن خانه‌هایی از این مجموعه مشخص می‌شوند. به عبارت دیگر، وقتی داده‌ای از تعدادی از پاسخگویان در برخی از متغیرها وجود ندارد، هنگام ورود داده‌ها نیز در خانه‌های مربوط به این متغیرها که به این پاسخگویان تعلق دارند، هیچ مقداری وارد نمی‌شود. در SPSS این خانه‌های خالی‌مانده، برای متغیرهای عددی با نقطه (ممیز انگلیسی) نشان‌داده می‌شوند.

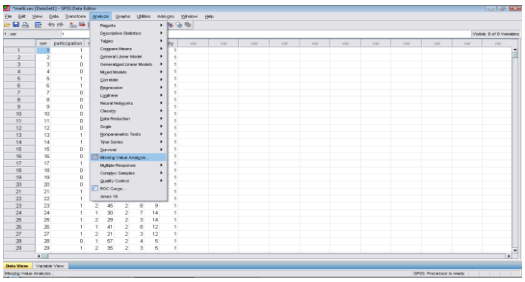
اغلب دو وضعیت باعث می‌شود تا مجموعه داده‌ای دارای متغیرهایی با خانه‌های خالی باشد. بخشی از این خانه‌های خالی ناشی از بی‌پاسخی هستند؛ یعنی پاسخگو از پاسخ دادن به پرسشی (متغیری) خودداری می‌کند و به این ترتیب، داده‌ای برای او در آن متغیر ثبت نمی‌شود. گاهی نیز خانه‌های خالی ناشی از پرش‌هایی هستند که در پرسشنامه وجود دارند. برای مثال، هیچ‌گاه از فردی که بیننده تلویزیون نیست، میزان رضایت از تلویزیون پرسیده نمی‌شود تا داده مربوط به متغیر رضایت برای او وجود داشته باشد. بنابراین، داده‌ای نیز برای رضایت غیربینندگان در مجموعه داده‌ها ثبت نمی‌شود.

گاهی نیز داده‌های گمشده نتیجه انجام محاسبات و ایجاد متغیرهای جدید از سوی کاربر است. برای مثال، ممکن است متغیری از جمع چند متغیر دیگر محاسبه شود. در SPSS این محاسبه را می‌توان از دو طریق انجام داد: استفاده از تابع sum یا نوشتن فرمول جمع‌بستن چند متغیر. اگر کاربر از شیوه فرمول‌نویسی استفاده کند، وجود حداقل یک متغیر با داده گمشده باعث می‌شود هیچ مقداری برای مجموع متغیرها محاسبه نشود و خانه مربوط به آن خالی بماند، در حالی که استفاده از تابع sum چنین محدودیتی ندارد (تنها در صورتی تابع sum دارای هیچ مقداری نخواهد بود که تمام متغیرهای آن بدون مقدار باشند).

علاوه بر شیوه فوق، این امکان در SPSS فراهم‌شده است تا کاربر نیز قادر باشد بعضی از مقادیر را به عنوان داده گمشده تعریف کند. به سخن دیگر، خانه‌هایی از مجموعه داده‌ها، با وجود داشتن مقدار، به دلخواه کاربر از طریق SPSS گمشده قلمداد می‌شوند. برای مثال، ممکن است در پرسشی، گزینه «نمی‌دانم» وجودداشته و کدی نیز برای آن انتخاب‌شده باشد. بنابراین، هنگام ورود داده‌ها برای این پاسخ از این کد استفاده خواهد شد و خانه‌ای خالی نمی‌ماند. حال اگر پژوهشگر بخواهد در بخشی از پردازش‌های خود،

پاسخ «نمی‌دانم» را نیز به عنوان داده گمشده درنظر بگیرد، بدون آنکه در کد مربوط به آن تغییری ایجاد کند، این امر با تعریف این کد به عنوان داده گمشده امکان‌پذیر است.

برای تعریف داده گمشده برای یک متغیر از سوی کاربر، در پنجره variable view روی خانه missing مقابل آن متغیر کلیک می‌کنیم تا پنجره missing values باز شود. کاربر می‌تواند از طریق این پنجره تک مقدارها را در بخش discrete missing values به عنوان داده گمشده تعریف کند و حتی می‌تواند فاصله‌ای از داده‌ها را به عنوان گمشده تعریف کند. به این ترتیب، این داده‌ها با وجود آنکه مقداری برای آنها در مجموعه داده‌ها ثبت‌شده است، به عنوان گمشده قلمداد می‌شوند.

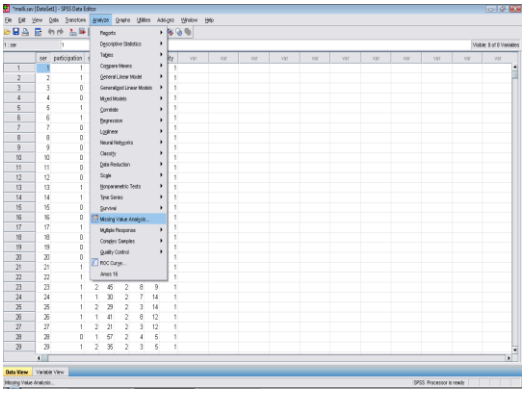


تحلیل داده‌های گمشده

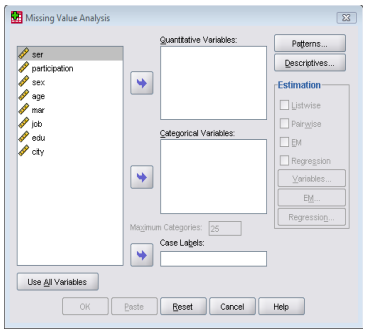
در منوی analyze گزینه‌ای با عنوان missing values analysis وجود دارد که برای تحلیل داده‌های گمشده و جانهی از آن استفاده می‌شود. با کلیک بر روی این گزینه، پنجره‌ای با همین نام باز می‌گردد که دستورات کاربر برای تحلیل داده‌های گمشده از طریق آن به SPSS داده می‌شود.

کاربر می‌تواند با امکانات موجود در این پنجره، تحلیل‌های زیر را بر روی داده‌های گمشده موجود در مجموعه داده خود اعمال کند.

- الگوی حاکم بر داده‌های گمشده را بررسی کند. از این طریق گزارشی از خانه‌هایی با داده‌های گمشده در اختیار کاربر قرار می‌گیرد. این گزارش، بررسی تصادفی‌بودن بروز داده‌های گمشده، محل قرارگرفتن داده گمشده یا این امر را که آیا یک پاسخگو در دو متغیر دارای داده گمشده است، ممکن می‌سازد.
- برآورد آماره‌هایی نظیر میانگین، انحراف معیار یا ضرایب همبستگی مربوط به متغیرهای دارای داده‌های گمشده براساس روش‌های مختلف تحلیل داده‌های گمشده.
- جانهی داده‌های گمشده؛ یعنی قراردادن مقداری برآوردشده در خانه‌های فاقد داده.



در سمت چپ پنجره missing values analysis، تمام متغیرهای موجود در مجموعه داده‌ها فهرست شده‌اند. کاربر می‌تواند هر یک از این متغیرها را از این قسمت انتخاب و به مستطیل‌های میان پنجره منتقل کند. اگر متغیر فاصله‌ای یا نسبتی باشد، به قسمت quantitative variables و اگر اسمی یا ترتیبی باشد، به قسمت categorical variables منتقل می‌شود. در صورتی که ستونی در مجموعه داده‌ها وجود دارد که شامل برچسب برای تک‌تک سطرها (پاسخگویان) است، این متغیر می‌تواند به case labels منتقل شود.

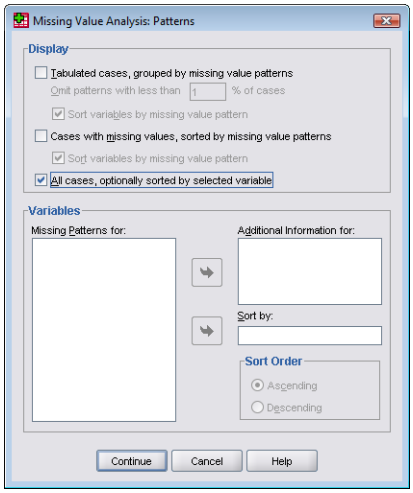


برای تحلیل داده‌های گمشده می‌توان کار را با کلیک بر روی لبه Patterns آغاز کرد که به بازشدن پنجره missing value analysis: patterns منجر می‌شود. امکانات موجود در این پنجره آگاهی از وضعیت داده‌های گمشده در تک‌تک متغیرها را ممکن می‌سازد. به این ترتیب پژوهشگر قادر خواهد بود به پرسش‌های زیر پاسخ دهد:

- داده‌های گمشده در کجای مجموعه داده‌ها قرار دارند (کدام خانه‌های ماتریس داده‌ها)؟
- آیا ممکن است داده‌های مربوط به دو متغیر، همزمان روی تعدادی از پاسخگویان گمشده باشد؟

- آیا مقادیر دورافتاده در داده‌ها وجود دارد؟

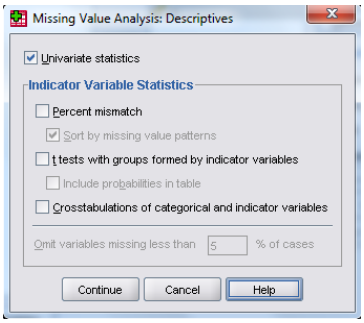
در این پنجره سه گزینه برای نمایش گزارش خروجی وجود دارد که کاربر می‌تواند همگی یا برخی از آنها را به‌طور همزمان انتخاب کند. گزینه اول، وضعیت داده‌های گمشده و معلوم را برای هر متغیر در خروجی نمایش می‌دهد، در حالی که گزینه دوم، تنها محدود به موارد گمشده است. گزینه سوم، وضعیت گمشده‌بودن داده‌های تک‌تک پاسخگویان را برای متغیرهای انتخابی نشان می‌دهد.



بنابراین، با انتخاب گزینه اول در خروجی تعداد موارد گمشده و معلوم ارائه می‌شود و با انتخاب گزینه دوم، وضعیت تک‌تک پاسخگویان دارای داده گمشده مشخص می‌گردد. با انتخاب گزینه سوم نیز، گزارشی از وضعیت هر پاسخگو، خواه دارای داده گمشده باشد یا نباشد، ارائه خواهد شد. توجه کنید که ممکن است در این خروجی از نمادهای جدول زیر استفاده شود:

+	مقدار دورافتاده خیلی بزرگ	S	داده‌های گمشده سیستمی	B	نوع دوم داده‌های گمشده تعریف‌شده از سوی کاربر
–	مقدار دورافتاده خیلی کوچک	A	نوع اول داده‌های گمشده تعریف‌شده از سوی کاربر	C	نوع سوم داده‌های گمشده تعریف‌شده از سوی کاربر

لبه دیگری که درپنجره missing values analysis وجود دارد Descriptives است که با کلیک بر روی آن پنجره زیر باز می‌شود. انتخاب گزینه Univariate statistics در این پنجره، آماره‌های توصیفی را برای تک‌تک متغیرهای انتخاب‌شده ارائه می‌کند. این آماره‌ها عبارتند از: تعداد موارد معلوم، تعداد و درصد موارد گمشده، میانگین، انحراف معیار و تعداد داده‌های دورافتاده.



بخش دوم این پنجره با عنوان Indicator Variable Statistics مشخص شده است. منظور از متغیر نشانگر (Indicator)، متغیری دو حالتی با مقادیر صفر و یک است؛ مقدار صفر برای پاسخگو با داده گمشده و مقدار یک برای پاسخگو با داده معلوم درنظر گرفته می‌شود. به این ترتیب، پاسخگویان بر حسب این متغیر به دو گروه دارای داده‌های گمشده و دارای داده‌های معلوم تفکیک می‌شوند. از این متغیر می‌توان برای تحلیل‌های آماری زیر برای مقایسه این دو گروه با یکدیگر استفاده کرد:

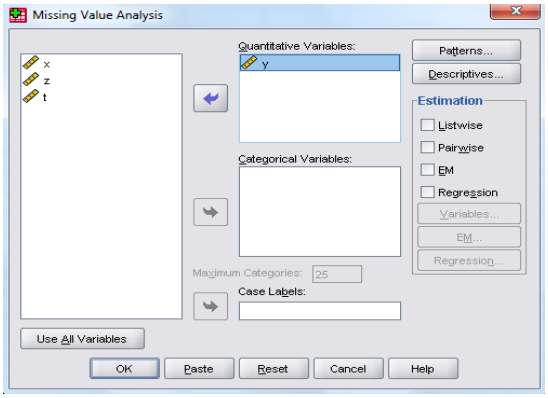
- گزینه Percent mismatch برای هر زوج از متغیرها، درصد پاسخگویانی را نشان می‌دهد که در یک متغیر دارای داده گمشده و در متغیر دیگر دارای داده معلوم هستند. درواقع، ابتدا برای هر متغیر، متغیر نشانگری ساخته می‌شود و سپس با تقاطع این دو متغیر نشانگر، این درصد محاسبه می‌گردد (این درصد بر حسب مواردی محاسبه می‌شود که یک متغیر نشانگر دارای مقدار صفر و متغیر نشانگر دیگر دارای مقدار یک است).
- گزینه t tests میانگین یک متغیر کمّی را در دو گروه متغیر نشانگر با آزمون t مقایسه می‌کند.
- گزینه Crosstabulations وضعیت داده‌های گمشدهٔ متغیرهای موجود در مجموعه داده‌ها را به‌تفکیک رده‌های متغیرهای اسمی یا ترتیبی انتخاب‌شده گزارش می‌دهد.

برآورد یا جانهی داده‌های گمشده

برآورد آماره‌ها یا جانهی موارد گمشده برای متغیرهای کمّی صورت می‌گیرد. استفاده از روش‌های برآورد یا جانهی تنها در شرایطی منطقی است که الگوی داده‌های گمشده کاملاً تصادفی باشد. به عبارت دیگر، احتمال گمشده‌بودن برخی از مقادیر بیشتر از مقادیر دیگر نباشد[1].

۱. گاه احتمال گمشدگی برای برخی از مقادیر متغیر بیش از مقادیر دیگر است. برای مثال، وقوع گمشدگی برای متغیر درآمد در مقادیر درآمدی بالا بیش از مقادیر درآمدی دیگر است (افراد پردرآمد بیش از سایرین از گفتن درآمد خود امتناع می‌کنند). به این ترتیب، گمشدگی در متغیر به مقادیر آن بستگی دارد و کاملاً تصادفی نیست.

در سمت راست پنجره missing values analysis، چهار گزینه ذیل عنوان Estimation وجود دارند. این چهار گزینه هنگامی فعال خواهند شد که متغیری در مستطیل Quantitative Variables انتخاب شده باشد.



دو گزینه Listwise و Pairwise به SPSS می‌گویند که چگونه پاسخگویان را با توجه به داده‌های گمشده آنها در محاسبه میانگین، انحراف معیار، همبستگی و کوواریانس دخالت دهد. با انتخاب گزینه Listwise، پاسخگویانی در محاسبه این آماره‌ها حضور خواهند داشت که در تمام متغیرهای انتخاب‌شده دارای داده معلوم باشند. گزینه Pairwise بر روی محاسبات آماره‌های میانگین و انحراف معیار تأثیری ندارد، ولی محاسبات مربوط به همبستگی و کوواریانس را محدود به پاسخگویانی می‌کند که در دو متغیرِ تحت محاسبه، دارای مقادیر معلوم هستند.

برای روشن‌شدن موضوع به مجموعه داده زیر توجه کنید. اگر هدف، محاسبه آماره‌های میانگین، انحراف معیار، همبستگی و کوواریانس برای سه متغیر y، z و t در این مجموعه داده ۲۵ تایی باشد، گزینه Listwise باعث می‌شود پاسخگویان ۳، ۹، ۱۶، ۱۹ و ۲۲ در محاسبه هیچ یک از آماره‌ها حضور نداشته باشند و محاسبات بر اساس ۲۰ پاسخگویی صورت گیرد که در هر سه متغیر دارای داده‌های معلوم هستند. بنابراین، حتی میانگین متغیر z نیز بدون حضور پاسخگوی ۹ محاسبه می‌شود، هر چند مقدار متغیر z برای این پاسخگو معلوم است.

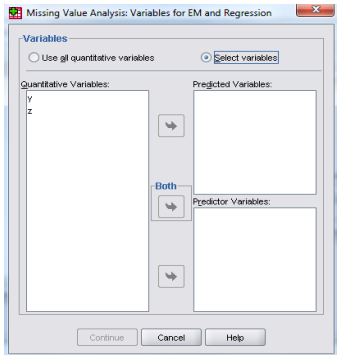
انتخاب گزینه Pairwise باعث می‌شود میانگین و انحراف معیار برای هر یک از متغیرها بر اساس داده‌های معلوم آنها صورت گیرد. به این ترتیب، میانگین و انحراف معیار متغیر z بر اساس ۲۲ پاسخگو یعنی با حذف پاسخگویان ۱۶، ۱۹ و ۲۲ و میانگین و انحراف معیار متغیر t بر اساس ۲۰ پاسخگو یعنی با حذف پاسخگویان ۳، ۹، ۱۶، ۱۹ و ۲۲ محاسبه می‌شود. از سوی دیگر، همبستگی و کوواریانس دو متغیر y و t بر اساس داده‌هایی که برای هر دو متغیر معلوم هستند ـ یعنی با حذف پاسخگویان ۹، ۱۶ ۱۹ و ۲۲ ـ محاسبه می‌گردد، در حالی که همبستگی y و t با حذف پاسخگویان ۳، ۹، ۱۶، ۱۹ و ۲۲ محاسبه می‌شود.

	x	y	z	t
1	1.00	23.00	45.00	55.50
2	1.00	45.00	89.00	110.50
3	2.00	61.00	118.00	.
4	1.00	23.00	45.00	55.50
5	2.00	49.00	94.00	110.50
6	3.00	80.00	156.00	188.00
7	2.00	90.00	176.00	213.00
8	1.00	67.00	133.00	165.50
9	1.00	.	70.00	.
10	1.00	11.00	21.00	25.50
11	3.00	53.00	105.00	130.50
12	1.00	59.00	117.00	145.50
13	1.00	70.00	139.00	173.00
14	2.00	99.00	194.00	235.50
15	2.00	10.00	16.00	13.00
16	2.00	.	.	.
17	2.00	100.00	196.00	238.00
18	2.00	34.00	64.00	73.00
19	.	.	.	.
20	1.00	82.00	163.00	203.00
21	2.00	20.00	36.00	38.00
22	.	18.00	.	.
23	2.00	33.00	62.00	70.50
24	3.00	92.00	180.00	218.00
25	1.00	90.00	179.00	223.00

روش رگرسیونی

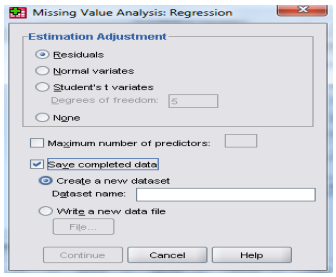
روش رگرسیونی، داده‌های گمشده را به کمک رگرسیون خطی چندگانه برآورد و سپس میانگین، انحراف معیار، همبستگی و کوواریانس متغیرها را بر اساس مجموعه داده کامل‌شده محاسبه می‌کند. به عنوان نمونه، برای متغیر y در مجموعه داده قبل، ابتدا مقدارهایی برای پاسخگویان ۹، ۱۶ و ۱۹ برآورد می‌شود تا داده‌های مربوط به این متغیر کامل گردد و سپس میانگین و انحراف معیار این متغیر بر اساس ۲۵ داده (۲۲ داده معلوم و ۳ داده برآوردشده) محاسبه می‌شود. برای محاسبه همبستگی و کوواریانس دو متغیر y و z نیز ابتدا مقادیر گمشده هر یک از این دو متغیر برآورد می‌شود و سپس همبستگی و کوواریانس دو متغیر بر اساس ۲۵ داده موجود محاسبه می‌گردد.

مدل رگرسیون برای برآورد داده‌های گمشده بر اساس تمام متغیرهای کمّی انتخاب‌شده صورت می‌گیرد، ولی کاربر می‌تواند خود تعیین کند کدام متغیرها به عنوان پیشگو (متغیرهای مستقل مدل) و کدامیک به عنوان پیش‌بینی‌شونده (متغیر وابسته مدل) باشند. برای این منظور باید بر روی لبه Variables در پنجره missing values analysis کلیک کنید تا پنجره مقابل باز شود (توجه کنید که لبه Variables هنگامی فعال می‌شود



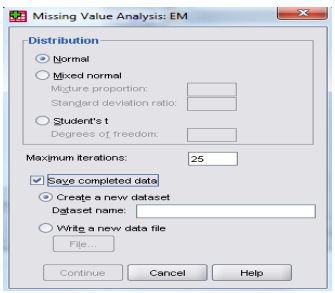
که گزینه Regression در پنجره missing values analysis انتخاب‌شده باشد). اکنون می‌توانید با کلیک بر روی گزینه Select variables از مستطیل سمت راست که حاوی متغیرهای انتخاب‌شده برای تحلیل رگرسیونی است، تعدادی را به عنوان پیشگو و تعدادی را به عنوان پیش‌بینی‌شونده انتخاب کنید. البته این امکان نیز وجود دارد که متغیری هم پیشگو و هم پیش‌بینی‌شونده باشد.

لبه دیگری که در پنجره missing values analysis برای روش رگرسیونی فعال می‌شود، Regression است که منجر به باز شدن پنجره مقابل می‌گردد. با انتخاب گزینه Save complete data در این پنجره از SPSS می‌خواهیم تا مقادیری را که با روش رگرسیونی برای داده‌های گمشده برآورد کرده است، در مجموعه داده دیگری ذخیره کند. به این ترتیب، مجموعه داده کاملی برای تحلیل‌های آماری در اختیار خواهیم داشت.



روش EM

این روش، تابع توزیعی را برای داده‌های گمشده درنظر می‌گیرد و بر اساس احتمالات حاصل از این توزیع در دو گام، فرایند جانهی را صورت می‌دهد. در گام نخست بر اساس مقادیر معلوم، پارامترهای مدل‌های آماری خاصی برآورد می‌شوند تا میانگین‌های (امیدهای شرطی) داده‌های گمشده محاسبه شوند و جایگزین این داده‌ها گردند. اکنون مجموعه کاملی از داده‌ها در دست است و می‌توان همان پارامترها را بر اساس داده‌های کامل برآورد و بار دیگر میانگین‌های (امیدهای شرطی) داده‌های گمشده را محاسبه کرد تا جایگزین آنها شوند. به این ترتیب، برای بار دوم، مجموعه داده کاملی ایجاد می‌شود و می‌توان این فرایند را تکرار



کرد. تکرار تا آنجا ادامه می‌یابد که نسبت به دقیق بودن برآورد خود اطمینان حاصل کنیم[۱].

انتخاب گزینه EM در پنجره missing values analysis باعث فعال‌شدن دو لبه Variables و EM می‌شود. پنجره اول همان کاربردی را دارد که در روش رگرسیونی داشت، لذا از توضیح مجدد آن خودداری می‌شود. کلیک بر روی لبه EM موجب باز شدن پنجره مقابل می‌گردد. قسمتی از این پنجره که با عنوان Distribution مشخص‌شده است، به کاربر این امکان را می‌دهد تا توزیعی را برای داده‌های گمشده تعیین کند. توزیع پیش‌فرض در SPSS، توزیع نرمال است. قسمت دوم این پنجره برای ذخیره کردن مقادیر جانهی شده در مجموعه داده دیگری است تا کاربر بتواند از یک مجموعه داده کامل در تحلیل‌های آماری دیگر خود استفاده کند.

منابع:

1. Miller, D., and Vivien, C. (2006). "Imputation Methods Documents", Available at http://help.pop.psu.edu/help-by-statistical-method/missing-data-impuaimputation/imputation_methods_document.doc
2. SPSS 16 Help for Missing Value Analysis.
3. PASW Statistics 18 Manual (2009). "Build better models when you fill in the blanks". Available at www.spss.com/media/collateral/statistics/SMV1702SPC-0209.pdf.

۱. پیشتر اشاره شد که شرط استفاده از روش‌های جانهی، کاملاً تصادفی‌بودن الگوی گمشدگی است. روش EM به شرط ساده‌تری یعنی گمشدگی تصادفی (Missing at Random یا MAR) نیاز دارد. گاه ممکن است احتمال وقوع داده گمشده برای متغیری در سطوح متغیر دیگری بیشتر شود. برای مثال به دلیل محافظه‌کار بودن افراد مسن، آنان بیشتر از گروه‌های سنی دیگر از گفتن مدت زمان تماشای ماهواره امتناع می‌کنند. بنابراین، احتمال اینکه متغیر مدت زمان تماشای ماهواره بی‌پاسخ بماند، برای این گروه سنی بیش از گروه‌های سنی دیگر است، ولی احتمال بی‌پاسخی در این متغیر برای تمام افراد مسن یکسان است. به عبارت دیگر، رفتار گروه سنی مسن با سایر گروه‌ها فرق می‌کند، ولی رفتار افراد مسن در داخل گروه خود با یکدیگر تفاوتی ندارد. چنین داده‌های گمشده‌ای را که احتمال وقوعشان با تغییر سطح متغیر دیگر تغییر می‌کند، ولی در هر سطح کاملاً تصادفی است، گمشدگی تصادفی می‌نامند.