چگونگی مواجهه با **داده‌های گم‌شده**

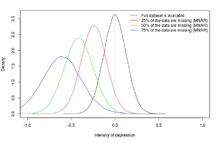
در [آمار](https://fa.wikipedia.org/wiki/%D8%A2%D9%85%D8%A7%D8%B1)، **داده‌های گم‌شده** یا **مقدارهای گم‌شده**زمانی رخ می‌دهد؛ که هیچ مقدار داده‌ای برای یک متغیر در یک مشاهده ذخیره نشده باشد. داده‌های گم‌شده اتفاقی رایج است و می‌تواند اثر معنی‌داری از نتایج حاصل از داده‌ها داشته باشد.

داده‌های گم‌شده می‌توانند به‌دلیل [بی‌پاسخی آماری](https://fa.wikipedia.org/w/index.php?title=%D8%A8%DB%8C%E2%80%8C%D9%BE%D8%A7%D8%B3%D8%AE%DB%8C_%D8%A2%D9%85%D8%A7%D8%B1%DB%8C&action=edit&redlink=1) رخ‌دهند: یعنی هیچ اطلاعاتی برای یک یا چند مورد در کل واحد («موضوع») به‌دست نیامده باشد. بعضی موارد به احتمال بیشتری بی‌پاسخی تولید می‌کنند: برای مثال مواردی که در مورد مسائل خصوصی مانند درآمد هستند. [ساییدگی](https://fa.wikipedia.org/w/index.php?title=%D8%B3%D8%A7%DB%8C%DB%8C%D8%AF%DA%AF%DB%8C&action=edit&redlink=1) نوعی از دست‌رفتن که در مطالعات جغرافیایی رخ دهد. برای نمونه مطالعهٔ توسعه که اندازه‌گیری بعد از دوره‌ای از زمان تکرارشده‌است. ازدست‌رفتگی زمانی که یکی از شرکت‌کنندگان قبل از پایان مطالعه را ترک می‌کند، یا یک یا چند اندازه‌گیری از دست می‌رود، رخ می‌دهد.

داده‌ها اغلب در تحقیقات اقتصادی، جامعه‌شناسی و [علوم سیاسی](https://fa.wikipedia.org/wiki/%D8%B9%D9%84%D9%88%D9%85_%D8%B3%DB%8C%D8%A7%D8%B3%DB%8C) از بین می‌روند، زیرا دولت یا نهادهای خصوصی گزارش‌های حساس را ناقص ارائه دهند، یا کلاً ارائه ندهند،یا به این علت که داده‌ها در دسترس نیستند ارائه داده نمی‌شود. بعضی مواقع گم‌شده بودن داده‌ها به‌خاطر خود محقق است، برای مثال، هنگامی که جمع‌آوری داده‌ها نامناسب است یا (در سیستم‌های اطلاعاتی) به اشتباه وارد می‌شوند

این شکل‌های بسته به نوع اثری که بر صحت نتیجه‌گیری‌های یک تحقیق می‌گذارند، انواع مختلفی دارند که عبارت‌اند از: به‌طور کامل ازدست‌رفته و به‌طور تصادفی ازدست‌رفته. می‌توان با داده‌های گم‌شده همانند [داده‌های سانسورشده](https://fa.wikipedia.org/w/index.php?title=%D8%AF%D8%A7%D8%AF%D9%87%E2%80%8C%D9%87%D8%A7%DB%8C_%D8%B3%D8%A7%D9%86%D8%B3%D9%88%D8%B1%D8%B4%D8%AF%D9%87&action=edit&redlink=1) رفتار کرد.

فهمیدن دلیل این‌که چرا داده‌ها گم‌شده‌اند برای مدیریت صحیح سایر داده‌ها مهم است. اگر داده‌ها کاملاً به‌تصادف گم‌شده باشند، نمونهٔ داده‌ها احتمالاً هنوز نماینده‌ای جامعه خواهد بود. اما اگه داده‌ها به شکل سیستماتیک گم‌شده باشند. تحلیل ممکن است اریب باشد. برای مثال، در مطالعهٔ ارتباط بین IQ و درآمد، اگر شرکت‌کنندگان با IQ متوسط به بالا تمایل به پاسخ ندادن به سؤال «درآمد شما چقدر است؟» نداشته باشند، تحلیلی که این مسئله را به‌عنوان یک الگوی گم‌شدن به‌صورت تصادفی الگوی MAR در نظر نمی‌گیرد، ممکن است در پیدا کردن همبستگی بین حقوق و IQ شکست بخورد. به‌خاطر این مشکلات، روش‌شناسان به‌طور معمول به محققان توصیه می‌کنند مطالعات را طراحی کنند تا وقوع داده‌های گم‌شده را به حداقل برسانند. مدل‌های گرافیکی می‌توانند با جزییات، سازوکار داده‌های گم‌شده را توصیف کنند

[](https://commons.wikimedia.org/wiki/File:Missing_not_at_random.png?uselang=fa)

این نمودار توزیع احتمال برآورد شدت مورد انتظار افسردگی در جامعه را نشان می‌دهد. تعداد موارد ۶۰ است. فرض کنید جامعه اصلی یک توزیه نرمال باشد و احتمال بی‌پاسخی تابعی لجستیک از شدت افسردگی باشد. نتیجه این است که: هر چه دادهٔ بیشتری گم‌شده باشد (MNAR)، برآوردها [اریب‌تر](https://fa.wikipedia.org/w/index.php?title=%D8%A7%D8%B1%DB%8C%D8%A8%E2%80%8C&action=edit&redlink=1) هستند و ما شدت افسردگی در جامعه را دست‌کم می‌گیریم.

**گم‌شده به‌صورت کاملاً تصادفی**

مقدارهای درون مجموعه‌داده‌ها به‌صورت کاملاً تصادفی گم‌شده (MCAR) هستند، اگر رویدادهایی که منجر به گم‌شدن هر نوع مورد-داده‌ای می‌شوند از متغیرهای قابل مشاهده و پارامترهای مورد علاقهٔ غیرقابل مشاهده مستقل باشند و به‌شکل کاملاً تصادفی رخ دهند  هنگامی که داده‌ها MCAR هستند، تحلیل انجام‌شده روی داده‌ها نااریب است، گرچه داده‌ها به ندرت MCAR هستند

در مورد MCAR، از دست‌رفتگی داده‌ها، از هر متغیر دیگر مستقل است؛ بنابراین، شرکت‌کنندگان با داده‌های کامل مشاهده‌شده در اثر، یک نمونه تصادفی از همه شرکت‌کنندگان مربوط یک مداخلهٔ خاص هستند. با MCAR، تخصیص تصادفی تیمارها، محفوظ فرض می‌شود، اما معمولاً در عمل این فرضی قوی و غیرواقع‌بینانه است.

**گم‌شده به‌طور تصادفی  
  
گم‌شده به‌طور تصادفی (MAR)** هنگامی رخ می‌دهد که از دست‌رفتگی تصادفی نباشد، اما ازدست‌رفتگی می‌تواند به حساب متغیری که در آن اطلاعات کاملی باشد اطلاق شود MAR فرضی است که تأیید آماری آن غیرممکن است، و بر فرض معقول بودن آن اعتماد می‌شود. یک مثال این است که مردها کمتر احتمال دارد که در یک مطالعه در مورد [افسردگی](https://fa.wikipedia.org/wiki/%D8%A7%D9%81%D8%B3%D8%B1%D8%AF%DA%AF%DB%8C) شرکت کنند، حال آن‌که این مسئله هیچ ربطی به سطح افسردگیشان ندارد. بسته به روش تحلیل، این داده‌ها به‌دلیل خالی‌بودن شرطی سلول‌ها منجر به [اریبی](https://fa.wikipedia.org/wiki/%D8%A7%D8%B1%DB%8C%D8%A8%DB%8C_%DB%8C%DA%A9_%D8%A8%D8%B1%D8%A2%D9%88%D8%B1%D8%AF%DA%AF%D8%B1) در تحلیل شوند (در مردها افسردگی خیلی زیاد ممکن است هیچ ورودی‌ای نداشته باشد). به‌هرحال، اگر پارامتر با [برآورد درستنمایی بیشینه](https://fa.wikipedia.org/w/index.php?title=%D8%A8%D8%B1%D8%A2%D9%88%D8%B1%D8%AF_%D8%AF%D8%B1%D8%B3%D8%AA%D9%86%D9%85%D8%A7%DB%8C%DB%8C_%D8%A8%DB%8C%D8%B4%DB%8C%D9%86%D9%87&action=edit&redlink=1) اطلاعات کامل برآورد شود، MAR برآوردهایی به‌طور [مجانبی:نظریه مجانبی](https://fa.wikipedia.org/w/index.php?title=%D9%85%D8%AC%D8%A7%D9%86%D8%A8%DB%8C:%D9%86%D8%B8%D8%B1%DB%8C%D9%87_%D9%85%D8%AC%D8%A7%D9%86%D8%A8%DB%8C&action=edit&redlink=1) نااریب ارائه می‌دهد.

**گم‌شده به‌صورت غیرتصادفی  
  
گم‌شده به‌صورت غیرتصادفی (MNAR)**(که به‌عنوان بی‌پاسخی غیرقابل چشم‌پوشی هم شناخته می‌شود) داده‌هایی هستند که نه MAR و نه MCAR باشند (برای مثال مقداری از متغیر که به دلیل مشخص گم‌شده باشد). برای توسیع مثال قبل، اگر مردان به‌خاطر سطح افسردگیشان نتوانند پرسشنامهٔ افسردگی را پرکنند.

روش‌های کار با داده‌های گم‌شده

داده‌های گم‌شده میزان نمایندهٔ جامعه بودنِ نمونه را کاهش می‌دهند، و بنابراین می‌توانند استنباط در مورد جامعه دچار مشکل کنند. به‌طور کلی، سه رویکرد کلی برای مقابله با داده‌های گم‌شده وجود دارد: [جانهی](https://fa.wikipedia.org/wiki/%D8%AC%D8%A7%D9%86%D9%87%DB%8C) -که به پرکردن جای داده‌های گم‌شده می‌گویند *حذف*که در آن نمونه‌ها با داده‌های نامعتبر از تحلیل دور انداخته می‌شوند*و تحلیلی*-به‌کار بردند روش‌هایی بی‌تأثیر از داده‌های گم‌شده.

در برخی موارد کاربردی، آزمایش‌گرها می‌تواند سطح گم‌شدگی داده‌ها را کنترل کنند و از دست‌رفتگی داده‌ها قبل از جمع‌آوری داده‌ها پیش‌گیری کنند. برای مثال، در [پرسشنامه‌های](https://fa.wikipedia.org/w/index.php?title=%D9%BE%D8%B1%D8%B3%D8%B4%D9%86%D8%A7%D9%85%D9%87%E2%80%8C&action=edit&redlink=1) کامپیوتری، معمولاً جواب ندادن سؤال‌ها غیرممکن است. به این شکل که یک سؤال باید جواب داده‌شود وگرنه نمی‌توان سؤال بعدی را جواب داد؛ بنابراین گم‌شدگی داده‌ها توسط شرکت‌کنندگان با این نوع پرسشنامه ممکن نیست، اگرچه این نوع پرسشنامه ممکن است این روش در مشی اخلاقی هیئت نظارت بر آن پژوهش مجاز نباشد. در تحقیقات، چندین تلاش برای تماس با شرکت‌کنندگان، اغلب از طریق ارسال نامه، برای ترغیب افرادی که تصمیم به شرکت نکردن در مطالعه گرفته‌اند، برای تغیر تصمیمشان، در مطالعه امری مرسوم است. بااین‌حال، این روش‌ها هم می‌تواند در کاهش آثار مخرب استنباطی داده‌های گم‌شده کمک کند و هم آسیب بزند، زیرا نوع افرادی که قرار است بعد از رد اولیه شرکتشان در مطالعه برای شرکت‌کردن مجدد متقاعد شوند، به‌طور معنی‌دار متفاوت از نوع افرادیست که پس از درخواست مجدد از آن‌ها برای شرکت در مطالعه، درخواست را رد می‌کنند

در شرایطی که وقوع داده‌های گم‌شده محتمل است، محققان معمولاً به استفاده از روش‌های آماری مقاوم در برابر گم‌شدگی داده‌ها توصیه می‌شوند. یک تحلیل وقتی مقاوم است که ما مطمئن باشیم تغییرات خفیف تا متوسط فرضیات تکنیکی کلیدی یا اریبی ایجاد نمی‌کند یا مقدار کمی اریبی ایجاد می‌کند، یا تحریف در نتایج در مورد جمعیت است.

**جانهی**

برخی روش‌های [تجزیه و تحلیل داده‌ها](https://fa.wikipedia.org/wiki/%D8%AA%D8%AD%D9%84%DB%8C%D9%84_%D8%AF%D8%A7%D8%AF%D9%87%E2%80%8C%D9%87%D8%A7) در برابر داده‌های گم‌شده مقاوم نیستند و نیاز به «پر کردن»، یا جانهی مقدار گم‌شده دارند. روبین (۱۹۸۷) استدلال کرد که تکرار جانهی حتی برای تعداد اندک (۵ یا کمتر) تا حد زیادی کیفیت برآورد را بهبود می‌بخشد. برای اهداف زیادی، ۲ یا ۳ جانهی برای بسیاری از اهداف عملی ۲ یا 3 imputations تصرف بسیاری از نسبی بهره‌وری است که می‌تواند گرفته شده با تعداد زیادی از imputations. اما بیش از حد تعداد کمی از imputations می‌تواند منجر به از دست دادن قابل توجهی از [آماری قدرت](https://fa.wikipedia.org/wiki/%D8%AA%D9%88%D8%A7%D9%86_%D8%A2%D9%85%D8%A7%D8%B1%DB%8C)و برخی از محققان در حال حاضر توصیه می‌کنند ۲۰ تا ۱۰۰ یا بیشتر هر ضرب-منسوب تجزیه و تحلیل داده‌ها باید تکرار برای هر یک از منسوب [مجموعه داده](https://fa.wikipedia.org/wiki/%D9%85%D8%AC%D9%85%D9%88%D8%B9%D9%87_%D8%AF%D8%A7%D8%AF%D9%87) و در برخی موارد مربوط به آمار باید ترکیب نسبتاً پیچیده در راه است.

[الگوریتم امید ریاضی–بیشینه کردن](https://fa.wikipedia.org/wiki/%D8%A7%D9%84%DA%AF%D9%88%D8%B1%DB%8C%D8%AA%D9%85_%D8%A7%D9%85%DB%8C%D8%AF_%D8%B1%DB%8C%D8%A7%D8%B6%DB%8C%E2%80%93%D8%A8%DB%8C%D8%B4%DB%8C%D9%86%D9%87_%DA%A9%D8%B1%D8%AF%D9%86) (EM) رویکردی است که مقادیر آماره‌ها که در صورت کامل بودن یک مجموعه‌داده برآورد می‌شود (جانهی می‌شود)، با توجه به الگوی داده‌های گم‌شده. در این روی‌کرد، مقدار داده‌های گم‌شده تکی به‌طور معمول جانهی نمی‌شوند.

**درون‌یابی**

در شاخهٔ ریاضیاتی [آنالیز عددی](https://fa.wikipedia.org/wiki/%D8%A2%D9%86%D8%A7%D9%84%DB%8C%D8%B2_%D8%B9%D8%AF%D8%AF%DB%8C)، [درون‌یابی](https://fa.wikipedia.org/wiki/%D8%AF%D8%B1%D9%88%D9%86%E2%80%8C%DB%8C%D8%A7%D8%A8%DB%8C) روش ساخت روش ساخت نقطه داده‌های جدید در یک بازهٔ مجموعه‌ای گسسته از نقاط داده‌ای شناخته شده‌است.

در مقایسه دو نمونهٔ زوجی شامل داده‌های گم‌شده، یک [آماره](https://fa.wikipedia.org/wiki/%D8%A2%D9%85%D8%A7%D8%B1%D9%87) آزمون که از همه داده‌های موجود بدون نیاز به جانهی استفاده می‌کند، [آزمون تی استیودنت](https://fa.wikipedia.org/wiki/%D8%A2%D8%B2%D9%85%D9%88%D9%86_%D8%AA%DB%8C_%D8%A7%D8%B3%D8%AA%DB%8C%D9%88%D8%AF%D9%86%D8%AA) برای نمونه‌های به‌طور جزئی دارای هم‌پوشانی است. این آزمون تحت شرط نرمال بودن و فرض MCAR معتبر است.

**حذف جزئی**

روش‌هایی که شامل محدود کردن یک مجموعه‌داده به داده‌هایی بدون داده‌های گم‌شدگی که شامل این موارد می‌شوند:

* حذف لیستی/حذف موردی
* حذف روجی

**تحلیل کامل**

روش‌هایی که از همهٔ اطلاعات موجود، بدون اعوجاج نتایج حاصل مقدارهای جانهی شده به‌عنوان حالتی که در واقع مشاهده می‌شوند، به‌کارگیری می‌کند.

* رویکردهای مولد:
  + الگوریتم امید ریاضی–بیشینه کردن
  + برآورد بیشینهٔ درستنماییی اطلاعات کامل
* رویکردهای افتراقی:
  + طبقه‌بندی حداکثر-حاشیه داده‌ها با ویژگی‌ها ناموجود

روش‌های مبتنی بر مدل

روش‌های مبتنی بر مدل، غالباً با استفاده از نمودار، ابزارهای دیگری برای تست داده‌های گم‌شده و برای برآورد پارامترها تحت شرایطی که داده‌های گم‌شده وجود دارندپیشنهاد می‌دهند (MCAR, MAR و MNAR). برای مثال، یک تست برای رد کردن MAR/MCAR به‌شرح زیر است:

برای هرکدام از سه‌متغیر X, Y و Z که Z کاملاً مشاهده شده‌است و X و Y به‌طور جزئی مشاهده شده‌اند، دادها باید در شرط {\displaystyle }

به بیان دیگر، مشاهدهٔ جزئی X باید مستقل از وضعیت گم‌شدگی Y، به شرط هر مقدار Z باشد. برقرار نشدن این شرط بیان‌گر این است که مسئله متعلق به دستهٔ MNAR است

نکته: این آزمون‌ها برای MAR مبتنی بر متغیر که تغییر کمی نسبت به MAR مبتنی بر رویداد دارند لازم‌اند.

هنگامی که داده‌ها در دستهٔ MNAR قرار می‌گیرند، روش‌هایی برای برآورد سازگار پارامترها در شرایط خاص مدل در دسترس هستند. برای مثال اگر Y که بیانگر دلیل گم‌شدگی در X, Y است، خود دارای داده‌های گم‌شده باشد، هنوز هم [توزیع احتمال توأم](https://fa.wikipedia.org/wiki/%D8%AA%D9%88%D8%B2%DB%8C%D8%B9_%D8%A7%D8%AD%D8%AA%D9%85%D8%A7%D9%84_%D8%AA%D9%88%D8%A3%D9%85) X و Y در صورتی که گم‌شدگی Y تصادفی باشد می‌تواند برآورد شود. برآورد در این حالت به این شکل خواهد بود:

{\displaystyle }

که در آن {\displaystyle } و {\displaystyle }ساختارهای متفاوت هر زمان که برآوردهای سازگار ممکن است، منجر به برآوردهای متفاوت و فرایندهای متفاوت برآورد شود. مورد برآورد قبل اول با برآورد {\displaystyle } از اطلاعات کامل و ضرب آن در {\displaystyle } {\displaystyle } {\displaystyle }

در بسیاری از موارد روش‌های مبتنی بر مدل به ساختار مدل این اجازه را می‌دهد تا تحت آزمون‌های نادرستی قرار گیرد. هر مدلی که استفلال بین متغیر به‌طور جزئی مشاهده شدهٔ X و شاخص گم‌شدگی متغیر دیگر Y را ایجاب کند، یعنی {\displaystyle }به شرط {\displaystyle }{\displaystyle }

در نهایت، برآوردشده‌هایی که از این روش‌ها می‌آیند از فرم بسته می‌آیند و نیاز به فراینده‌های تکراری مانند الگوریتم امید ریاضی–بیشینه کردن ندارند

یک کلاس خاص از مشکلات هنگامی که احتمال گم‌شدگی به زمان وابسته است ظاهر می‌شود. برای مثال، در پایگاه دادهٔ تروما، احتمال از دست‌رفتن داده‌های مربوط به تروما وابسته به زمان روز بعد از تروماست. در این موارد مدل‌های [زنجیره ای مارکوف](https://fa.wikipedia.org/wiki/%D8%B2%D9%86%D8%AC%DB%8C%D8%B1%D9%87_%D9%85%D8%A7%D8%B1%DA%A9%D9%88%D9%81) ناایستای متنوعی اعمال می‌شود