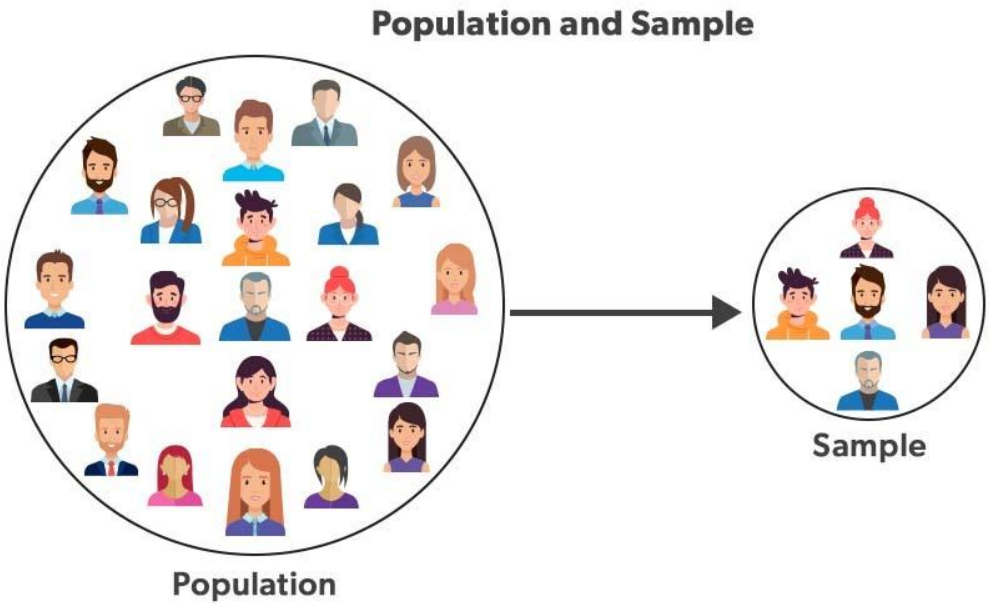
آمار از پایه تا حرفه‌ای؛ مفاهیم، اصول، احتمال و قانون‌هایی که باید بشناسی

آمار یه زبونه؛ زبونی که با اون می‌تونیم دنیای پر از عدد و داده‌ی اطرافمون رو بفهمیم. خیلی‌ها فکر می‌کنن آمار یعنی حفظ کردن فرمول و جدول، ولی واقعیت اینه که آمار یه **طرز فکره**؛ یه ابزار برای تصمیم‌گیری بهتر توی دنیایی که پر از عدم قطعیته.

توی این مقاله می‌خوایم قدم‌به‌قدم، با مثال‌های عددی واقعی و محاسبه‌شده جلو بریم: از مفاهیم پایه شروع می‌کنیم، بعد معیارهای مرکزی و پراکندگی رو با اعداد واقعی حساب می‌کنیم، بعد وارد دنیای احتمال می‌شیم، به قوانین بزرگ آماری می‌رسیم، و در آخر یه فهرست کامل از اشتباهات رایجی که حتی خیلی از آدم‌های باهوش هم توش می‌افتن رو بررسی می‌کنیم. حواستون باشه، این مقاله طولانیه ولی هر بخشش لازمه—پس یه چای بریز و بیا شروع کنیم.

بخش اول: مفاهیم پایه

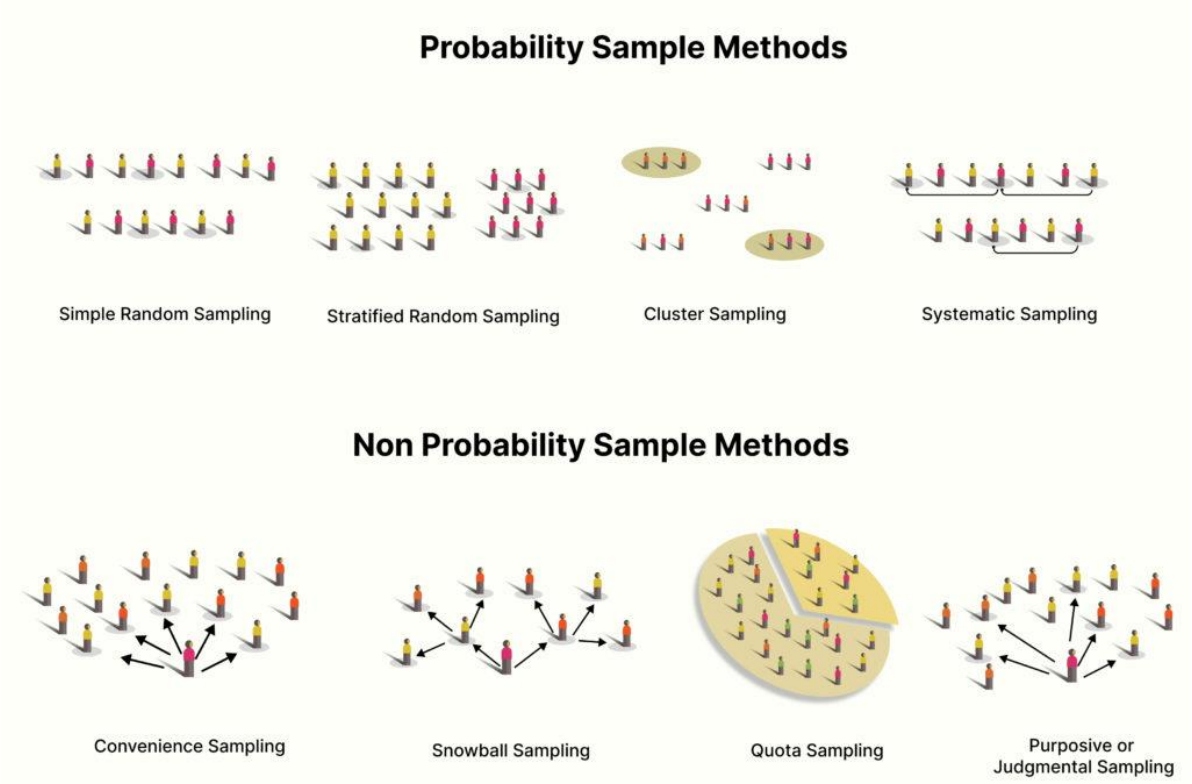


۱.جامعه‌ی آماری و نمونه

جامعه‌ی آماری (Population) یعنی کل گروهی که می‌خوایم دربارش اطلاعات جمع کنیم. مثلاً اگه بخوایم بفهمیم مردم ایران به طور میانگین چقدر قهوه می‌خورن، «همه‌ی ۸۵ میلیون نفر جمعیت ایران» جامعه‌ی آماری‌ست.

چون بررسی کل جامعه معمولاً غیرممکن یا خیلی پرهزینه‌ست (فکرش رو بکن بخوای از ۸۵ میلیون نفر تک‌تک بپرسی!)، یه بخش کوچیک‌تر و نماینده‌ی اون رو انتخاب می‌کنیم که بهش **نمونه (Sample)** می‌گیم. مثلاً پرسیدن از ۲۰۰۰ نفر به جای کل جمعیت.

مثال عددی :یه شرکت نظرسنجی می‌خواد بفهمه چند درصد مردم تهران از یه اپلیکیشن خاص استفاده می‌کنن. جامعه‌ی آماری، حدود ۹.۵ میلیون نفر جمعیت تهرانه. اگه از ۱۰۰۰ نفر بپرسن و ۴۲۰ نفرشون بگن «بله»، تخمین می‌زنن که ۴۲% از کل جمعیت تهران (یعنی تقریباً ۳.۹۹ میلیون نفر) از اون اپ استفاده می‌کنن. این تخمینه، نه عدد دقیق—و دقیقاً همینجا بحث خطای نمونه‌گیری پیش میاد که بعداً بهش می‌رسیم.



۲ .روش‌های نمونه‌گیری

انتخاب نمونه فقط «هر کسی که دم دستمون بود» نیست؛ چند روش اصلی داریم:

- **نمونه‌گیری تصادفی ساده :**هر عضو جامعه شانس مساوی برای انتخاب شدن داره. مثل قرعه‌کشی.
- **نمونه‌گیری طبقه‌ای :**جامعه رو به گروه‌ها (طبقات) تقسیم می‌کنیم، مثلاً بر اساس سن یا شهر، بعد از هر طبقه به نسبت جمعیتش نمونه می‌گیریم. مثلاً اگه ۳۰% جمعیت تهران زیر ۲۵ سال باشن، ۳۰% نمونه‌مون هم باید از این گروه سنی باشه.

- **نمونه‌گیری خوشه‌ای :**جامعه رو به خوشه‌ها (مثل محله‌ها) تقسیم می‌کنیم، چندتا خوشه رو تصادفی انتخاب می‌کنیم و همه‌ی اعضای اون خوشه‌ها رو بررسی می‌کنیم.
- **نمونه‌گیری در دسترس :**ساده‌ترین ولی ضعیف‌ترین روش؛ فقط از کسایی که راحت بهشون دسترسی داریم می‌پرسیم (مثل پرسیدن از دوستان توی اینستاگرام). این روش معمولاً سوگیری داره چون نماینده‌ی واقعی جامعه نیست.

نکته‌ی مهم: هرچقدر نمونه بهتر و تصادفی‌تر انتخاب بشه، نتیجه‌ای که می‌گیریم به واقعیت جامعه نزدیک‌تره. یه نمونه‌ی ۱۰۰۰ نفری تصادفی خیلی معتبرتر از یه نمونه‌ی ۵۰۰۰ نفری غیرتصادفیه (مثلاً فقط کسایی که توی یه صفحه‌ی خاص اینستاگرام کامنت گذاشتن).

۳ .متغیرها و سطوح اندازه‌گیری

هر چیزی که اندازه‌گیری یا ثبت می‌کنیم یه **متغیر** حساب می‌شه. دو نوع اصلی داره:

- **متغیر کمّی (عددی) :**مثل قد، وزن، سن، درآمد. خودش هم به دو دسته تقسیم می‌شه:
  - **پیوسته :**هر مقداری بین دو عدد ممکنه (مثل قد ۱۷۵.۳۲ سانتی‌متر، یا وزن ۶۸.۷ کیلوگرم)
  - **گسسته :**فقط مقادیر صحیح و مشخص (مثل تعداد فرزندان: ۰، ۱، ۲، ۳، ...)
- **متغیر کیفی (توصیفی) :**مثل رنگ چشم، جنسیت، شهر محل سکونت. عدد نیستن، دسته‌بندی‌ان.

فراتر از این تقسیم‌بندی، آمارگرها چهار **سطح اندازه‌گیری** رو هم تعریف می‌کنن که نشون می‌ده با یه داده چه عملیاتی می‌شه انجام داد:

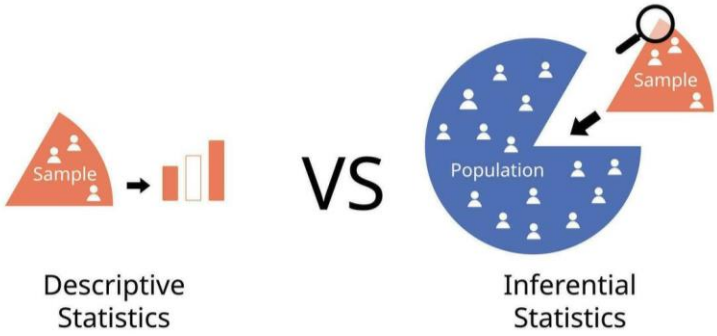
۱. **اسمی :(Nominal)** فقط دسته‌بندیه، ترتیب نداره. مثل: رنگ مو (مشکی، قهوه‌ای، بلوند).

۲. **ترتیبی :(Ordinal)** دسته‌بندی با ترتیب، ولی فاصله‌ها مشخص نیست. مثل: رضایت مشتری (خیلی ناراضی، ناراضی، خنثی، راضی، خیلی راضی). نمی‌تونیم بگیم فاصله‌ی «راضی» تا «خیلی راضی» دقیقاً برابر فاصله‌ی «ناراضی» تا «خنثی»ه.

۳. **فاصله‌ای :(Interval)** فاصله‌ها معنی‌دارن ولی صفر مطلق نداره. مثل: دما بر حسب سلسیوس. صفر درجه به معنی «نبود دما» نیست.

٤. **نسبتی :(Ratio)** فاصله‌ها معنی‌دارن و صفر مطلقه. مثل: وزن، قد، درآمد. صفر کیلوگرم یعنی واقعاً هیچی.

چرا این تقسیم‌بندی مهمه؟ چون خیلی از عملیات آماری (مثل میانگین گرفتن) فقط روی داده‌های فاصله‌ای و نسبتی معنی‌دارن. میانگین گرفتن از «رنگ مو» (مثلاً کد کردن مشکی=۱، قهوه‌ای=۲، بلوند=۳ و میانگین گرفتن) کاری بی‌معنیه، ولی خیلی‌ها این اشتباه رو مرتکب می‌شن.



۴ .آمار توصیفی در برابر آمار استنباطی

- **آمار توصیفی :(Descriptive Statistics)** فقط داده‌هایی که داریم رو خلاصه و توصیف می‌کنه. مثل میانگین نمرات یه کلاس ۳۰ نفره.
- **آمار استنباطی :(Inferential Statistics)** از روی یه نمونه، نتیجه‌گیری کلی درباره‌ی کل جامعه می‌کنه، همراه با یه سطح عدم قطعیت. مثل نظرسنجی‌های انتخاباتی که از ۱۵۰۰ نفر، نتیجه‌ی رأی ۶۰ میلیون نفر رو با «حاشیه‌ی خطای ۳%» پیش‌بینی می‌کنن.

مثال ملموس :اگه من میانگین قد ۳۰ نفر توی کلاسمون رو حساب کنم و بگم ۱۷۲ سانتی‌متره، این آمار توصیفیه— فقط دارم همون داده‌ای که دارم رو خلاصه می‌کنم. ولی اگه بگم «با ۹۵% اطمینان، میانگین قد کل دانش‌آموزان این مدرسه بین ۱۷۰ تا ۱۷۴ سانتی‌متره»، این آمار استنباطیه؛ دارم از روی یه نمونه، درباره‌ی یه جامعه‌ی بزرگ‌تر نتیجه می‌گیرم.

بخش دوم: معیارهای مرکزی و پراکندگی (با محاسبه‌ی کامل)

۵ .سه‌گانه‌ی مرکزیت: میانگین، میانه، نما

بیایم با یه مثال عددی واقعی کار کنیم. فرض کن حقوق ماهانه‌ی ۱۰ کارمند یه شرکت (به میلیون تومان) این‌طوریه:

۸، ۹، ۹، ۱۰، ۱۰، ۱۰، ۱۱، ۱۲، ۱۳، ۷۰

(نفر دهم مدیرعامله با حقوق ۷۰ میلیون)

میانگین :(Mean) جمع همه‌ی داده‌ها تقسیم بر تعدادشون.

جمع = ۷۰+۱۳+۱۲+۱۱+۱۰+۱۰+۱۰+۹+۹+۸ = ۱۶۲ میانگین = ۱۶۲ ÷ ۱۰ = **۱۶.۲ میلیون تومان**

میانه :(Median) عددی که وقتی داده‌ها رو مرتب کنیم، دقیقاً وسط قرار می‌گیره. چون ۱۰ عدد داریم (زوج)، میانه میانگین دو عدد وسطیه.

داده‌های مرتب‌شده: ۸، ۹، ۹، ۱۰، ۱۰، ۱۰، ۱۱، ۱۲، ۱۳، ۷۰ دو عدد وسطی (پنجم و ششم): ۱۰ و ۱۰ میانه = (۱۰+۱۰) ÷ ۲ = **۱۰ میلیون تومان**

نما :(Mode) پرتکرارترین مقدار. اینجا عدد ۱۰ سه بار تکرار شده، پس **نما = ۱۰ میلیون تومان**

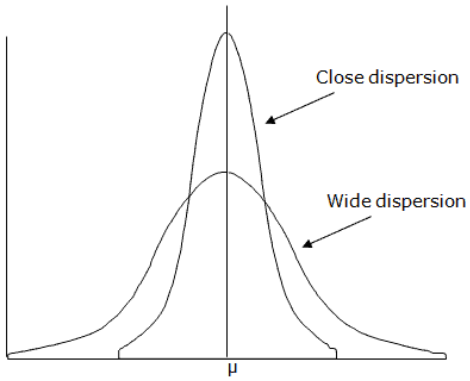
می‌بینی چه اتفاقی افتاد؟ میانگین (۱۶.۲) خیلی بالاتر از میانه و نما (۱۰) دراومد، چون حقوق مدیرعامل (۷۰) یه **داده‌ی پرت (Outlier)** بود که میانگین رو به شدت بالا کشید. اگه فقط میانگین رو ببینی، فکر می‌کنی وضعیت مالی کارمندا خوبه، در حالی که واقعیت اینه که ۹ نفر از ۱۰ نفر، حقوقی حدود ۸ تا ۱۳ میلیون دارن. برای همین توی گزارش‌های اقتصادی معتبر، معمولاً «میانه‌ی درآمد» به جای «میانگین درآمد» گزارش می‌شه—دقیق‌تر و کمتر گمراه‌کننده‌ست.

میانگین وزنی :(Weighted Mean) گاهی داده‌ها اهمیت یکسانی ندارن. مثلاً نمره‌ی درسی ممکنه از میان‌ترم (۳۰% وزن)، تکالیف (۲۰% وزن) و پایان‌ترم (۵۰% وزن) تشکیل بشه.

فرض کن نمره‌ی میان‌ترم ۱۵، تکالیف ۱۸، پایان‌ترم ۱۴ باشه:

میانگین وزنی = (۱۵×۰.۳) + (۱۸×۰.۲) + (۱۴×۰.۵) = ۴.۵ + ۳.۶ + ۷ = **۱۵.۱**

این با میانگین ساده‌ی این سه عدد (۱۵.۶۷) فرق داره، چون وزن‌ها یکسان نیستن.



Same mean and differnet dispersion

۶ .پراکندگی داده‌ها: دامنه، واریانس و انحراف معیار

میانگین به تنهایی کافی نیست. باید بدونیم داده‌ها چقدر دور یا نزدیک به میانگین پخش شدن.

دامنه :(Range) ساده‌ترین معیار؛ بیشترین منهای کمترین .توی مثال بالا: دامنه = ۷۰ − ۸ = ۶۲ میلیون تومان

واریانس و انحراف معیار :

واریانس :

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

انحراف معیار :

$$\sigma = \sqrt{\sigma^2}$$

بیایم با یه مثال کوچیک‌تر و بدون داده‌ی پرت کار کنیم تا محاسبه ساده‌تر بشه. فرض کن نمرات ۵ دانش‌آموز اینه :**۱۲، ۱۴، ۱۶، ۱۸، ۲۰**

قدم ۱: میانگین = (۲۰+۱۸+۱۶+۱۴+۱۲) ÷ ۵ = ۸۰ ÷ ۵ = **۱۶**

قدم ۲: فاصله‌ی هر داده از میانگین و مجذورش:

- ۱۲ − ۱۶− = ۴ → مجذور = ۱۶
- ۱۴ − ۱۶− = ۲ → مجذور = ۴
- ۱۶ − ۱۶ = ۰ → مجذور = ۰
- ۱۸ − ۱۶ = ۲ → مجذور = ۴
- ۲۰ − ۱۶ = ۴ → مجذور = ۱۶

قدم ۳: جمع مجذورها = ۱۶+۴+۰+۴+۱۶ = ۴۰

قدم ۴ :**واریانس** =۴۰ ÷ ۵ = **۸**

قدم ۵ :**انحراف معیار** =جذر ۸ ≈ = **۲.۸۳**

یعنی به طور میانگین، هر نمره حدود ۲.۸۳ واحد از میانگین (۱۶) فاصله داره. انحراف معیار چون واحدش مثل خود داده‌هاست (اینجا «نمره»)، فهمش خیلی راحت‌تر از واریانسه (که واحدش «نمره به توان ۲» می‌شه، یه چیز انتزاعی).

مثال مقایسه‌ای :دو کلاس ممکنه میانگین نمره‌شون ۱۵ باشه، ولی:

- کلاس الف: نمرات ۱۴، ۱۵، ۱۵، ۱۵، ۱۶ → انحراف معیار کم (حدود ۰.۷)
- کلاس ب: نمرات ۵، ۱۰، ۱۵، ۲۰، ۲۵ → انحراف معیار زیاد (حدود ۷.۱)

میانگین یکسانه ولی وضعیت واقعی خیلی فرق داره؛ کلاس الف یکدست و کلاس ب خیلی پراکنده‌ست. اگه معلم فقط میانگین رو گزارش بده، این تفاوت مهم دیده نمی‌شه.

۷. ضریب تغییرات (Coefficient of Variation)

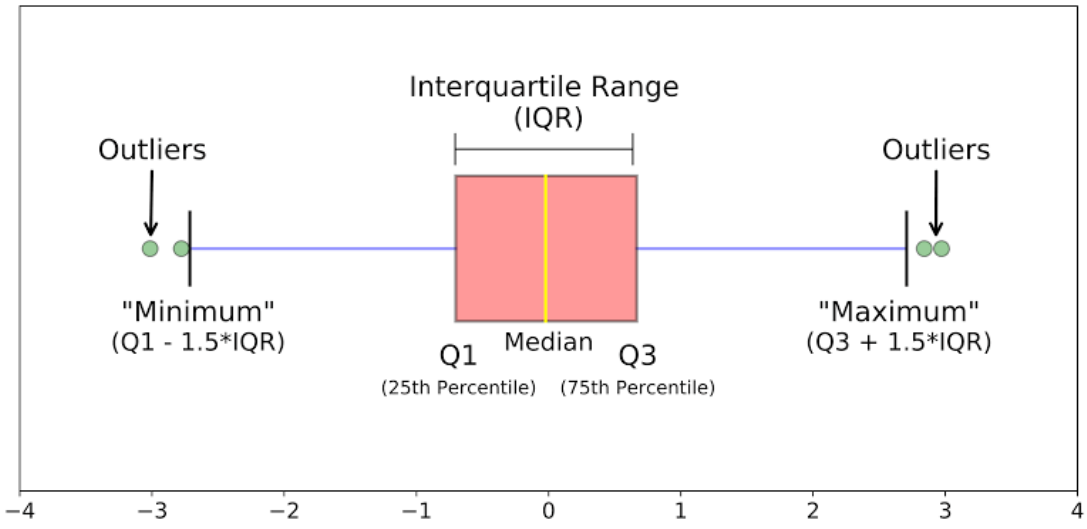
وقتی می‌خوایم پراکندگی دو مجموعه داده با واحد یا مقیاس متفاوت رو مقایسه کنیم، از **ضریب تغییرات** استفاده می‌کنیم:

CV = (انحراف معیار ÷ میانگین) × ۱۰۰

مثال: فرض کن می‌خوایم بدونیم قیمت خونه‌ها یا قیمت نون بیشتر نوسان داره.

- قیمت خونه: میانگین ۵ میلیارد تومان، انحراف معیار ۵۰۰ میلیون تومان) → CV = ۵۰۰÷۵۰۰۰ × (۱۰۰ = ۱۰%
- قیمت نون: میانگین ۱۵ هزار تومان، انحراف معیار ۳ هزار تومان) → CV = ۳÷۱۵ × (۱۰۰ = ۲۰%

با اینکه انحراف معیار قیمت خونه (۵۰۰ میلیون) به مراتب از انحراف معیار قیمت نون (۳ هزار تومان) بزرگ‌تره، اما نسبت به میانگینشون، قیمت نون نوسان بیشتری (۲۰% در برابر ۱۰%) داره. این یه نمونه از اینکه چرا مقایسه‌ی مستقیم اعداد خام گاهی گمراه‌کننده‌ست.



۸. چارک‌ها و دامنه‌ی بین‌چارکی (IQR)

داده‌ها رو می‌شه به چهار بخش مساوی تقسیم کرد:

- **چارک اول (Q1):** ۲۵% داده‌ها زیر این عدد قرار دارن.
- **چارک دوم (Q2):** همون میانه‌ست؛ ۵۰% داده‌ها زیرشن.
- **چارک سوم (Q3):** ۷۵% داده‌ها زیر این عدد قرار دارن.

دامنه‌ی بین‌چارکی (IQR) = Q3 − Q1

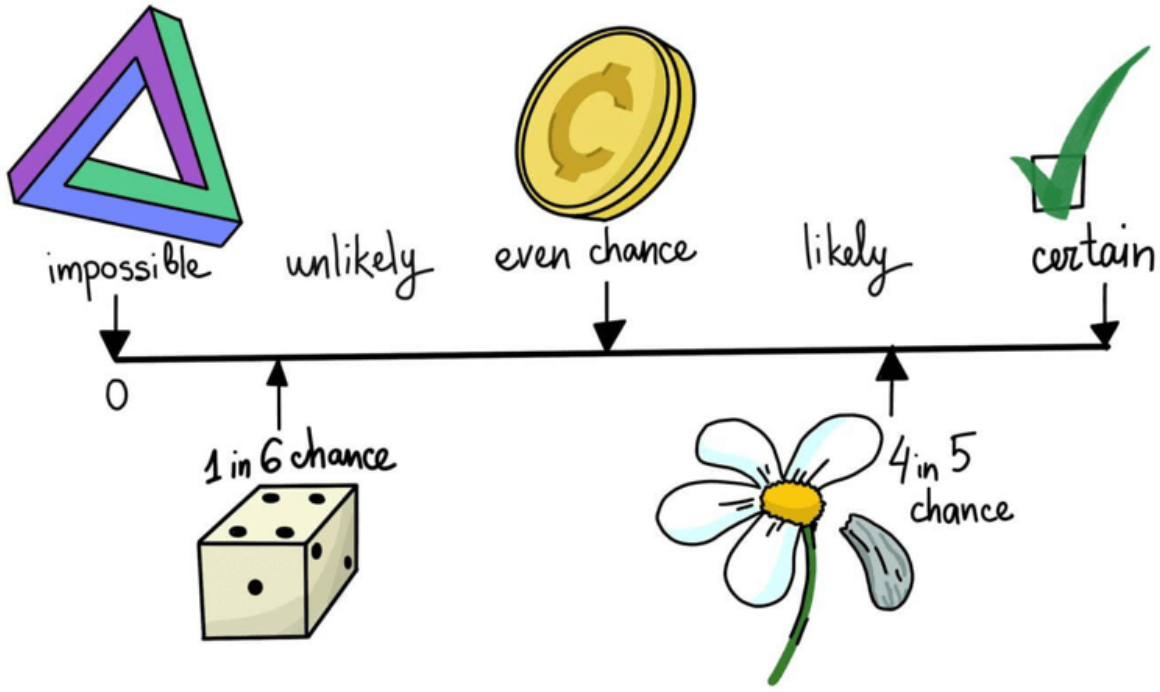
این معیار خیلی مفیده چون به داده‌های پرت حساس نیست (برخلاف دامنه‌ی معمولی). توی نمودارهای جعبه‌ای (Box Plot)که توی گزارش‌های علمی زیاد می‌بینیم، دقیقاً از همین معیارها استفاده می‌شه تا داده‌های پرت به‌راحتی شناسایی بشن.

۹. چولگی(Skewness)

وقتی توزیع داده‌ها متقارن نیست، بهش «چوله» می‌گیم:

- **چولگی مثبت (به راست) :**اکثر داده‌ها سمت چپ‌ان و یه دم بلند سمت راست هست. مثال کلاسیک: توزیع درآمد (اکثر مردم درآمد متوسط دارن، ولی چندتا فرد خیلی پردرآمد، میانگین رو به سمت راست می‌کشن).
- **چولگی منفی (به چپ) :**برعکس. مثال: نمرات یه امتحان خیلی آسون که اکثر دانش‌آموزان نمره‌ی بالا گرفتن ولی چندتا نمره‌ی خیلی پایین هم هست.

وقتی توزیع چوله باشه، میانگین و میانه با هم فاصله دارن—و همون‌طور که قبلاً دیدیم، توی این حالت‌ها میانه معمولاً معیار بهتری برای «حالت معمولی» جامعه‌ست.



بخش سوم: احتمال و قواعد پایه‌اش (با محاسبات کامل)

احتمال، پایه و اساس آمار استنباطیه. هر عددی که آمار بهمون می‌ده (پیش‌بینی، درصد اطمینان، ریسک) در واقع از دل نظریه‌ی احتمال بیرون میاد.

۱۰. احتمال چیه؟

احتمال یه عدد بین ۰ تا ۱ (یا ۰% تا ۱۰۰%) هست که نشون می‌ده یه اتفاق چقدر ممکنه رخ بده. ۰ یعنی غیرممکن، ۱ یعنی حتمی.

فرمول ساده‌ش (وقتی همه‌ی حالت‌ها هم‌شانس باشن):

احتمال یه اتفاق = تعداد حالت‌های مطلوب ÷ تعداد کل حالت‌ها

مثال ۱ :احتمال اومدن عدد ۳ توی پرتاب یه تاس شش‌وجهی: ۱ حالت مطلوب از ۶ حالت ممکن = $\frac{1}{6} \approx$ ۱۶.۷%

مثال ۲ :یه جعبه داره ۴ توپ قرمز، ۳ توپ آبی و ۳ توپ سبز (جمعاً ۱۰ توپ). احتمال برداشتن تصادفی یه توپ آبی چقدره؟ احتمال = ۳ ÷ ۱۰ = **۰.۳ = ۳۰%**

۱۱. رویدادهای مستقل و وابسته

- **مستقل :**وقوع یه اتفاق روی احتمال اتفاق دیگه تأثیری نمی‌ذاره. مثل پرتاب دوباره‌ی سکه؛ نتیجه‌ی پرتاب قبلی هیچ تأثیری روی پرتاب بعدی نداره.
- **وابسته :**وقوع یه اتفاق، احتمال اتفاق بعدی رو تغییر می‌ده. مثل کشیدن دو کارت پشت‌سرهم از یه دسته بدون برگردوندن کارت اول؛ احتمال کارت دوم به نتیجه‌ی کارت اول بستگی داره.

مثال عددی رویداد وابسته :توی همون جعبه‌ی ۱۰ توپی بالا (۴ قرمز، ۳ آبی، ۳ سبز)، اگه یه توپ قرمز برداری و **برنگردونی**، احتمال اینکه توپ دوم هم قرمز باشه چقدره؟

بعد از برداشتن توپ اول: ۹ توپ باقی مونده (۳ قرمز، ۳ آبی، ۳ سبز) احتمال توپ دوم قرمز = ۳ ÷ ۹ = **۳۳.۳%**

در مقابل، اگه توپ اول رو **برمی‌گردوندیم**، احتمال توپ دوم قرمز همون ۴ ÷ ۱۰ = ۴۰% می‌موند، چون رویدادها مستقل می‌شدن.

۱۲. مغالطه‌ی قمارباز(Gambler's Fallacy)

یکی از رایج‌ترین اشتباهات ذهنی انسان‌ها! خیلی‌ها فکر می‌کنن اگه سکه پنج بار پشت‌سرهم شیر بیاد، دفعه‌ی ششم «باید» خط بیاد تا «تعادل برقرار بشه .«**این کاملاً غلطه!**

چون پرتاب سکه مستقله، احتمال شیر یا خط توی هر پرتاب همیشه دقیقاً ۵۰% می‌مونه، مهم نیست قبلش چی اومده. سکه حافظه نداره! این مغالطه باعث میلیاردها تومان ضرر توی قمارخونه‌ها می‌شه؛ آدم‌ها فکر می‌کنن چون رولت چند دور قرمز اومده، «حتماً» دور بعدی سیاهه، در حالی که احتمال هر دور دقیقاً مثل قبله.

مثال عددی :احتمال اینکه سکه‌ای عادلانه، ۵ بار پشت‌سرهم شیر بیاره چقدره؟ $= \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) =$ $= \left(\frac{1}{2}\right)^5 = \left(\frac{1}{32}\right) =$ **۳.۱۲۵%**

این عدد کمه (فقط حدود ۳%)، ولی وقتی این اتفاق **افتاده باشه** و تو بخوای پرتاب ششم رو پیش‌بینی کنی، احتمال شیر اومدن دوباره دقیقاً **۵۰%** ئه—نه بیشتر، نه کمتر. این تفاوت بین «احتمال یه دنباله‌ی طولانی، قبل از وقوعش» و «احتمال پرتاب بعدی، بعد از وقوع دنباله» دقیقاً جایی‌ست که ذهن آدم‌ها گول می‌خوره.

۱۳. قانون جمع و قانون ضرب احتمال

- **قانون جمع** برای اتفاق‌های ناسازگار؛ یعنی هم‌زمان نمی‌تونن رخ بدن

- **P(A یا B) = P(A) + P(B)**

مثال: احتمال اومدن عدد ۱ یا ۲ توی تاس = $\frac{1}{6} + \frac{1}{6} = \frac{2}{6}$ = **۳۳.۳%**

- **قانون جمع عمومی** وقتی اتفاق‌ها می‌تونن هم‌زمان رخ بدن

- **P(A یا B) = P(A) + P(B) – P(A و B)**

مثال: توی یه دسته‌ی ۵۲ کارتی، احتمال کشیدن یه کارت که «دل» باشه یا «شاه» باشه چقدره؟ P(دل) = $\frac{13}{52}$، P(شاه) = $\frac{4}{52}$، P(دل و شاه با هم، یعنی شاه دل) = $\frac{1}{52}$P(دل یا شاه)

= $\frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} \approx$ **۳۰.۸%**

- **قانون ضرب** برای اتفاق‌های مستقل

- **P(A و B) = P(A) × P(B)**

مثال: احتمال اومدن شیر در دو پرتاب متوالی سکه = $\frac{1}{2} \times \frac{1}{2}$ = **۲۵%**

مثال بزرگ‌تر: احتمال اینکه سه نفر که تصادفی انتخاب می‌شن، هر سه در یه ماه خاص (مثلاً فروردین) متولد شده باشن (با فرض توزیع یکنواخت ۱۲ ماه

= $\frac{1}{12} \times \frac{1}{12} \times \frac{1}{12} = \frac{1}{1728} \approx$ **۰.۰۵۸%**

۱۴. احتمال شرطی

احتمال وقوع یه اتفاق، به شرط اینکه اتفاق دیگه‌ای از قبل رخ داده باشه. نمادش P(A|B) هست؛ یعنی «احتمال A به شرط اینکه B رخ داده»

مثال عددی :توی یه کلاس ۴۰ نفره، ۲۵ نفر دخترن و ۱۵ نفر پسر. از بین دخترها، ۱۰ نفر عینکی‌ان و از بین پسرها، ۶ نفر عینکی‌ان.

احتمال اینکه یه دانش‌آموز تصادفی، عینکی باشه به شرط اینکه دختر باشه P(عینکی|دختر) = (تعداد دخترهای عینکی) ÷ (تعداد کل دخترها) = ۱۰ ÷ ۲۵ = **۴۰%**

این با احتمال کلی عینکی بودن (۱۶ عینکی از ۴۰ نفر = ۴۰%) اینجا یکسان دراومد؛ یعنی جنسیت و عینکی بودن اینجا مستقل از هم بودن. ولی اگه اعداد فرق می‌کرد (مثلاً فقط ۵ نفر از دخترها عینکی بودن)، اون‌وقت P(عینکی|دختر) = ۲۵/۵ = ۲۰% می‌شد که خیلی متفاوت از احتمال کلیه—یعنی اونجا جنسیت و عینکی بودن با هم رابطه دارن.

۱۵. قضیه‌ی بیز (Bayes' Theorem)

یکی از قدرتمندترین و در عین حال گیج‌کننده‌ترین ابزارهای احتمال؛ کمک می‌کنه باورمون رو با رسیدن اطلاعات جدید، به‌روزرسانی کنیم.

فرمول کلی:

P(A|B) = [P(B|A) × P(A)] ÷ P(B)

مثال کاربردی و کامل (با اعداد دقیق):

فرض کن یه بیماری خاص شیوعش در جامعه ۱ در ۱۰۰۰ نفره) یعنیP (بیمار) = ۰.۰۰۱ .(یه تست پزشکی برای این بیماری داریم که:

- اگه فرد واقعاً بیمار باشه، تست با احتمال ۹۹% مثبت می‌شه (حساسیت تست)
- اگه فرد سالم باشه، تست با احتمال ۹۵% منفی می‌شه، یعنی ۵% مواقع اشتباهاً مثبت می‌شه (این ۵% یعنی نرخ مثبت کاذب)

حالا فرض کن تست یه نفر مثبت شده. احتمال اینکه این فرد **واقعاً** بیمار باشه چقدره؟

بیایم با فرض یه جمعیت ۱۰۰,۰۰۰ نفری حساب کنیم تا شهودی‌تر بشه:

- تعداد واقعاً بیمار = ۱۰۰,۰۰۰ × ۰.۰۰۱ = **۱۰۰ نفر**
- تعداد واقعاً سالم = ۹۹,۹۰۰ نفر

از بین ۱۰۰ نفر بیمار:

- تست مثبت واقعی = (True Positive) ۱۰۰ × ۰.۹۹ = **۹۹ نفر**

از بین ۹۹,۹۰۰ نفر سالم:

- تست مثبت کاذب = (False Positive) ۹۹,۹۰۰ × ۰.۰۵ = **۴,۹۹۵ نفر**

پس کل کسایی که تستشون مثبت شده = ۹۹ + ۴,۹۹۵ = **۵,۰۹۴ نفر**

احتمال اینکه یه نفر با تست مثبت، واقعاً بیمار باشهP :(بیمارامثبت) = ۹۹ ÷ ۵,۰۹۴ ≈ **۱.۹۴%**

فقط ۱.۹۴ درصد !با اینکه تست ۹۹% حساسیت داشت، به خاطر نادر بودن بیماری، اکثریت قریب‌به‌اتفاق نتایج مثبت، در واقع کاذب هستن. این پدیده رو **مغالطه‌ی نرخ پایه (Base Rate Fallacy)** می‌گن؛ خیلی از پزشکان و حتی محققان هم توی این محاسبه اشتباه می‌کنن و فقط به دقت تست توجه می‌کنن، بدون در نظر گرفتن شیوع واقعی بیماری. برای همین بعد از یه تست مثبت اولیه، همیشه تست‌های تکمیلی لازمه.

۱۶. متغیر تصادفی، توزیع احتمال و امید ریاضی

متغیر تصادفی یعنی خروجی یه اتفاق نامشخص که هنوز رخ نداده، مثل نتیجه‌ی پرتاب تاس یا تعداد مشتری‌هایی که فردا وارد یه مغازه می‌شن.

امید ریاضی (Expected Value) یعنی میانگین وزنی همه‌ی نتایج ممکن، با وزن احتمال هرکدوم:

E(X) = Σ [مقدار × احتمال آن مقدار]

مثال عددی :یه بلیت بخت‌آزمایی ۱۰ هزار تومانه. جایزه‌ها اینطوریه:

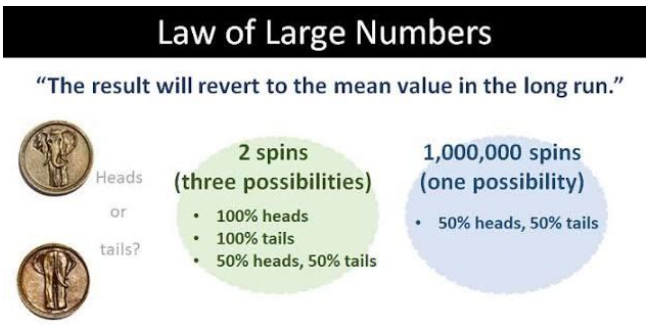
- ۱ نفر از هر ۱۰۰۰ نفر: ۵ میلیون تومان می‌بره
- ۵ نفر از هر ۱۰۰۰ نفر: ۵۰۰ هزار تومان می‌برن
- بقیه: هیچی نمی‌برن

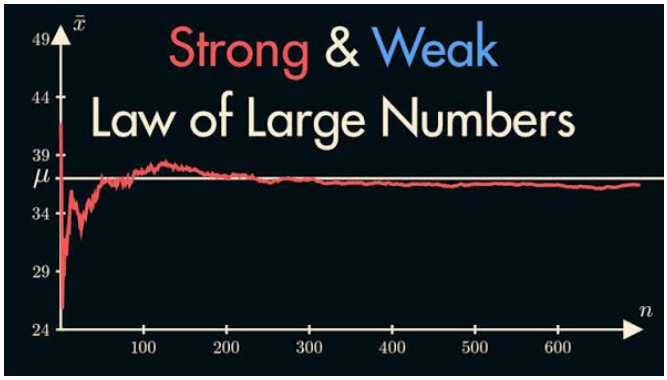
امید ریاضی جایزه =

$$\left(\frac{1}{1000}\times 5{,}000{,}000\right) + \left(\frac{5}{1000}\times 500{,}000\right) + \left(0\times\frac{994}{1000}\right) = 5000 + 2500 + 0 = \textbf{۷۵۰۰ تومان}$$

چون امید ریاضی جایزه (۷۵۰۰ تومان) کمتر از قیمت بلیت (۱۰,۰۰۰ تومان) هست، در بلندمدت هر کسی که مدام بلیت بخره، به طور میانگین ضرر می‌کنه (حدود ۲۵۰۰ تومان به ازای هر بلیت). این دقیقاً همون منطقیه که پشت هر بخت‌آزمایی، لاتاری و بازی کازینویی هست—طراحی شده که در بلندمدت، طرف مقابل (شرکت برگزارکننده) سود کنه.

بخش چهارم: اصول و قانون‌های مهم آماری





۱۷. قانون اعداد بزرگ(Law of Large Numbers)

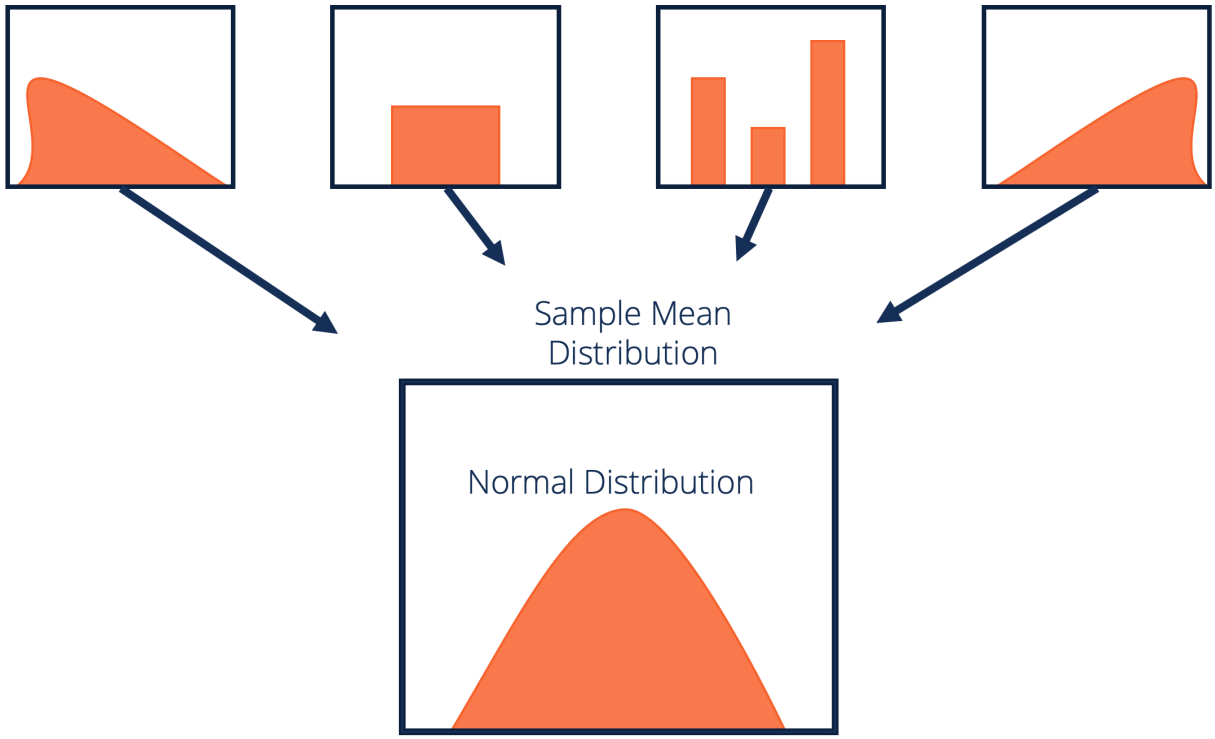
هرچقدر تعداد نمونه‌ها بیشتر بشه، میانگین به‌دست‌اومده به میانگین واقعی جامعه نزدیک‌تر می‌شه.

مثال زندگی و عددی :اگه یه سکه رو فقط ۵ بار بندازی، ممکنه ۴ بار شیر بیاد (نسبت ۸۰٪) و به نظر برسه سکه «تقلبیه». ولی:

- بعد از ۱۰۰ پرتاب، ممکنه نسبت شیر چیزی مثل ۵۶٪ باشه
- بعد از ۱۰,۰۰۰ پرتاب، نسبت شیر معمولاً چیزی مثل ۵۰.۳٪ می‌شه

- بعد از ۱,۰۰۰,۰۰۰ پرتاب، نسبت شیر تقریباً دقیقاً ۵۰.۰۰% می‌شه

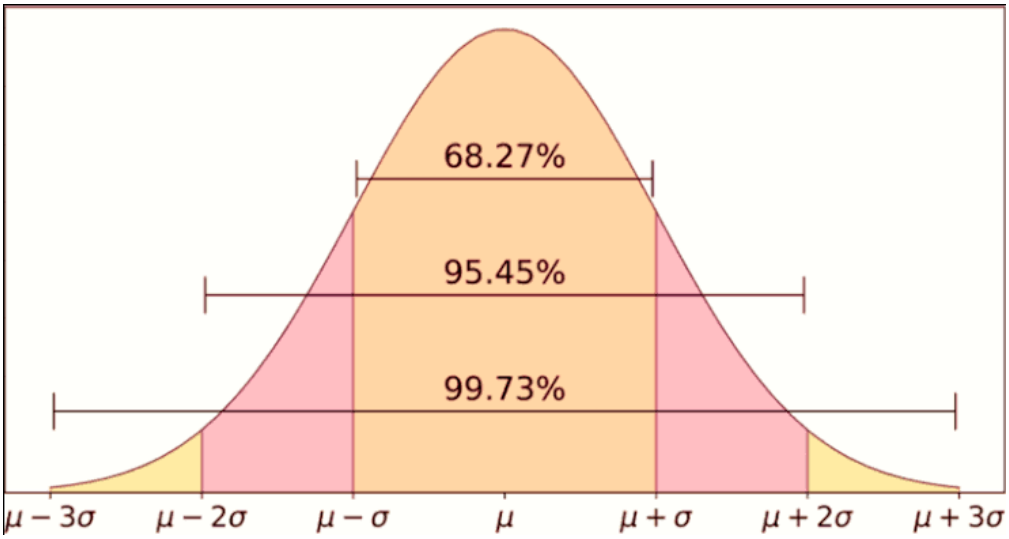
کازینوها دقیقاً روی همین اصل سود می‌کنن؛ تک‌تک بازی‌ها تصادفیه ولی روی هزاران بازی، آمار به نفع کازینو (که همیشه یه مزیت ریاضی کوچیک داره) ثابت می‌مونه.



۱۸. قضیه‌ی حد مرکزی (Central Limit Theorem)

اگه از یه جامعه (با هر شکل توزیعی، حتی کاملاً نامتقارن) نمونه‌های زیادی بگیریم و میانگین هرکدوم رو حساب کنیم، توزیع این میانگین‌ها به یه شکل خاص به اسم **توزیع نرمال** (زنگوله‌ای شکل) نزدیک می‌شه؛ حتی اگه خود داده‌های اصلی نرمال نباشن.

چرا این قضیه انقلابیه؟ چون بهمون اجازه می‌ده با نمونه‌های نسبتاً کوچیک (معمولاً از ۳۰ نمونه به بالا کافیه)، درباره‌ی جامعه‌های بزرگ نتیجه بگیریم و حتی حاشیه‌ی خطای دقیق هم محاسبه کنیم. این قضیه پایه‌ی نظرسنجی‌ها، آزمایش‌های بالینی دارو، و کنترل کیفیت کارخانه‌هاست.



۱۹. توزیع نرمال و قاعده‌ی ۶۸-۹۵-۹۹.۷

خیلی از پدیده‌های طبیعی (قد آدم‌ها، نمرات آزمون، خطای اندازه‌گیری، وزن نوزادان) از یه الگوی زنگوله‌ای شکل به اسم **توزیع نرمال** پیروی می‌کنن. قانون کلیش اینه:

- حدود **۶۸%** داده‌ها در فاصله‌ی یک انحراف معیار از میانگین قرار دارن
- حدود **۹۵%** در فاصله‌ی دو انحراف معیار
- حدود **۹۹.۷%** در فاصله‌ی سه انحراف معیار

مثال عددی :فرض کن قد میانگین مردان یه کشور ۱۷۵ سانتی‌متر با انحراف معیار ۷ سانتی‌متره:

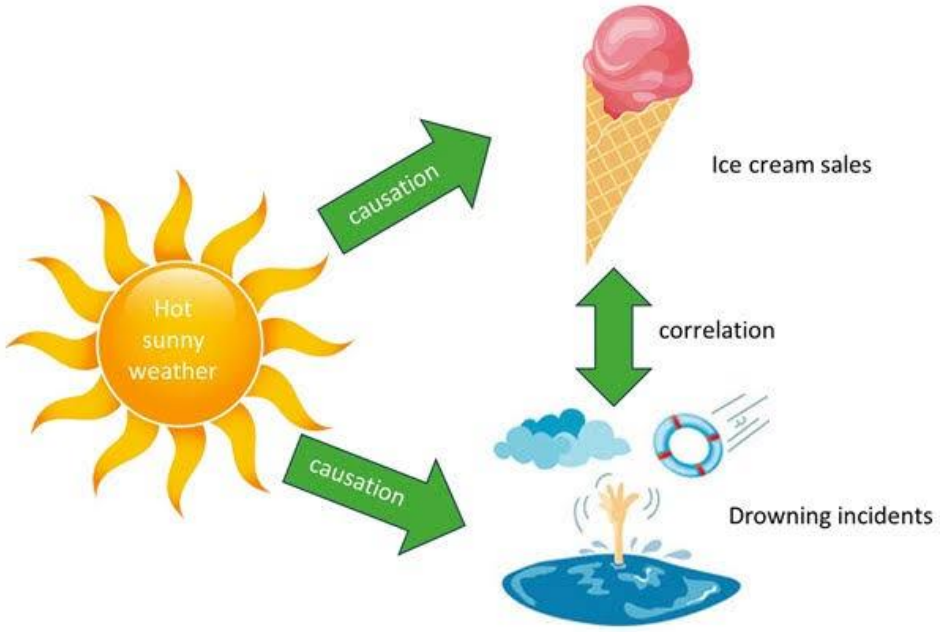
- ۶۸% مردان قدی بین ۱۶۸ تا ۱۸۲ سانتی‌متر دارن(۱۷۵ ± ۷)
- ۹۵% مردان قدی بین ۱۶۱ تا ۱۸۹ سانتی‌متر دارن(۱۷۵ ± ۱۴)
- ۹۹.۷% مردان قدی بین ۱۵۴ تا ۱۹۶ سانتی‌متر دارن(۱۷۵ ± ۲۱)

یعنی یه مرد با قد ۱۹۶ سانتی‌متر (۳ انحراف معیار بالاتر از میانگین)، جزو نادرترین ۰.۱۵% بلندترین مردان جامعه‌ست—یه آدم خیلی خیلی بلند نسبت به بقیه!

نمره‌ی استاندارد :(Z-Score) فاصله‌ی یه داده از میانگین رو بر حسب تعداد انحراف معیار نشون می‌ده:

Z = (مقدار داده − میانگین) ÷ انحراف معیار

مثلاً برای فردی با قد ۱۸۹ سانتی‌متر) = Z :۱۷۵−۱۸۹÷(۷ = ۲ .یعنی این فرد دقیقاً ۲ انحراف معیار بالاتر از میانگینه.



۲۰. همبستگی، نه علیت (Correlation is not Causation)

یکی از مهم‌ترین و پرکاربردترین اصول آماری. اینکه دو پدیده با هم رابطه یا همبستگی دارن، به این معنی نیست که یکی باعث اون یکی می‌شه. **ضریب همبستگی (r)** عددی بین ۱- تا ۱+ هست:

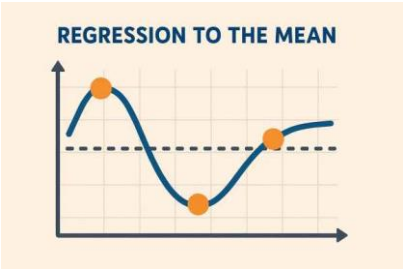
- نزدیک به ۱+: رابطه‌ی مستقیم قوی (وقتی یکی زیاد می‌شه، اون یکی هم زیاد می‌شه)
- نزدیک به ۱-: رابطه‌ی معکوس قوی
- نزدیک به ۰: رابطه‌ی خطی ضعیف یا وجود نداره

مثال‌های معروف همبستگی کاذب:

۱. فروش بستنی و تعداد غرق‌شدگی توی استخر هر دو توی تابستون بالا می‌رن. آیا بستنی باعث غرق شدن می‌شه؟ نه! یه **عامل سوم — (Confounding Variable)** یعنی گرمای هوا — هر دو رو تحت تأثیر قرار می‌ده.

۲. توی یه پژوهش دیده شده کشورهایی که مصرف شکلات سرانه‌شون بیشتره، تعداد برنده‌ی جایزه‌ی نوبل بیشتری هم دارن! آیا شکلات باعث نبوغ می‌شه؟ احتمالاً نه — هر دو با ثروت و توسعه‌ی اون کشور رابطه دارن.

۳. تعداد فیلم‌های نیکلاس کیج در یه سال با تعداد غرق‌شدگی در استخر توی آمریکا همبستگی بالایی داره (این یه مثال طنز مشهوره که نشون می‌ده گاهی همبستگی فقط تصادفیه، بدون هیچ رابطه‌ی منطقی).



۲۱ .بازگشت به میانگین(Regression to the Mean)

وقتی یه اتفاق خیلی خوب یا خیلی بد بیفته، احتمالاً اتفاق بعدی بهتر (یا بدتر) نمی‌شه، بلکه به حالت میانگین و عادی نزدیک‌تر می‌شه.

مثال عددی :فرض کن میانگین نمره‌ی یه کلاس توی امتحان‌ها ۱۵ از ۲۰ هست. یه دانش‌آموز توی یه امتحان نمره‌ی ۲۰ کامل می‌گیره (خیلی بالاتر از حد معمولش که مثلاً ۱۴ بوده). توی امتحان بعدی، به احتمال زیاد نمره‌اش به سمت ۱۵-۱۴ برمی‌گرده، نه اینکه دوباره ۲۰ بگیره. این افت، لزوماً به خاطر «تنبلی بعد از موفقیت» نیست؛ صرفاً یه پدیده‌ی آماریه.

مثال ورزشی :یه بازیکن فوتبال که یه فصل استثنایی و پرگل داشته (مثلاً ۳۵ گل، در حالی که میانگین کاریرش ۱۸ گله)، فصل بعد معمولاً عملکرد «عادی‌تری» داره؛ نه چون افت کرده، بلکه چون اون فصل استثنایی از اول هم یه اتفاق دور از میانگین بوده. این پدیده باعث سوءتفاهم‌های زیادی می‌شه، مثل این باور غلط که «تنبیه کردن بهتر از تشویق کردن جواب می‌ده» (چون بعد از یه عملکرد خیلی بد، طبیعتاً بهبود پیش میاد، حتی بدون تنبیه—فقط بازگشت به میانگینه).

۲۲ .اصل پارتو (قانون ۲۰-۸۰)

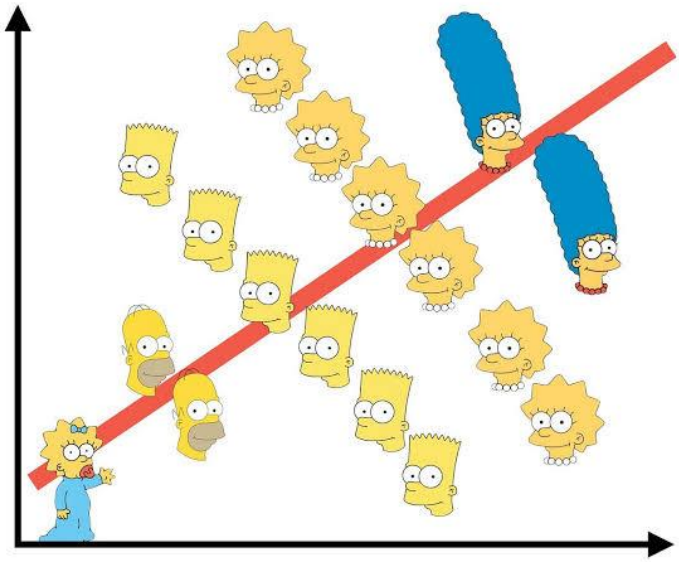
این اصل می‌گه در خیلی از پدیده‌ها، تقریباً ۸۰% نتایج از ۲۰% عوامل ناشی می‌شه.

مثال‌های عددی:

- ۸۰% فروش یه شرکت از ۲۰% مشتریانش میاد. مثلاً اگه یه شرکت ۱۰۰۰ مشتری داره و کل فروشش ۱۰ میلیارد تومانه، احتمالاً همون ۲۰۰ مشتری برتر، حدود ۸ میلیارد تومان از اون فروش رو تشکیل می‌دن.

- ۸۰٪ باگ‌های یه نرم‌افزار توی ۲۰٪ بخش‌های کد پیدا می‌شن.
- ۲۰٪ کارهای روزانه‌مون معمولاً ۸۰٪ نتیجه رو می‌سازن.

این اصل دقیقاً ۲۰-۸۰ نیست همیشه (گاهی ۱۰-۹۰ یا ۳۰-۷۰ هم هست)، ولی ایده‌ی کلیش (نامتوازن بودن توزیع تأثیرگذاری عوامل) توی مدیریت زمان و کسب‌وکار خیلی کاربرد داره.



۲۳. پارادوکس سیمپسون (Simpson's Paradox)

یکی از عجیب‌ترین و کمتر شناخته‌شده‌ترین پدیده‌های آماری. گاهی یه روند توی چند گروه مجزا وجود داره، ولی وقتی گروه‌ها رو با هم ترکیب می‌کنیم، روند برعکس می‌شه!

مثال عددی واقعی (شبیه به مورد معروف پزشکی):

فرض کن دو روش درمانی A و B رو روی بیماران با دو سطح شدت بیماری (خفیف و شدید) امتحان می‌کنیم:

روش B	روش A	وضعیت بیماری
۹۰۰ از ۱۰۰۰ بهبود (**۹۰٪**)	۹۵ از ۱۰۰ بهبود (**۹۵٪**)	**بیماران خفیف**
۴۰ از ۱۰۰ بهبود (**۴۰٪**)	۴۵۰ از ۱۰۰۰ بهبود (**۴۵٪**)	**بیماران شدید**
۹۴۰ از ۱۱۰۰ بهبود (**۸۵,۴٪**)	۵۴۵ از ۱۱۰۰ بهبود (**۴۹,۵٪**)	**جمع کل**

- در بیماران **خفیف**، روش A بهتر از B است.
- در بیماران **شدید**، باز هم روش A بهتر از B است.
- اما وقتی همه را با هم جمع می‌زنیم، **روش B** بهتر از A به نظر می‌رسد!

نسخه‌ی معروف‌تر این پارادوکس: یه دانشگاه متهم به تبعیض جنسیتی توی پذیرش شده چون کلاً ۴۵% مردان و فقط ۳۰% زنان پذیرفته شدن. ولی وقتی جدا جدا بر اساس دانشکده نگاه می‌کنیم:

- دانشکده‌ی مهندسی: ۸۰% مردان پذیرفته شدن، ۸۲% زنان پذیرفته شدن (زنان کمی بهتر!)
- دانشکده‌ی هنر: ۲۰% مردان پذیرفته شدن، ۲۵% زنان پذیرفته شدن (زنان کمی بهتر!)

توی هر دو دانشکده، زنان نرخ پذیرش بهتری داشتن! پس چرا در کل، مردان نرخ پذیرش بالاتری داشتن؟ چون **بیشتر زنان برای دانشکده‌ی هنر (که نرخ پذیرش پایینی داره) درخواست داده بودن**، و بیشتر مردان برای دانشکده‌ی مهندسی (که نرخ پذیرش بالایی داره) درخواست داده بودن. این توزیع نامتوازن درخواست‌ها، نتیجه‌ی کلی رو برعکس نشون داد. این دقیقاً نشون می‌ده چرا نگاه کردن فقط به آمار کلی، بدون تفکیک زیرگروه‌ها، می‌تونه کاملاً گمراه‌کننده باشه.

دانشکده	متقاضیان مرد	پذیرفته‌شدگان مرد	نرخ پذیرش مرد	متقاضیان زن	پذیرفته‌شدگان زن	نرخ پذیرش زن
مهندسی	نفر ۴۰۰	نفر ۳۲۰	**۸۰%**	نفر ۱۰۰	نفر ۸۲	**۸۲%**(!بهتر)
هنر	نفر ۶۰۰	نفر ۱۲۰	**۲۰%**	نفر ۹۰۰	نفر ۲۲۵	**۲۵%**(!بهتر)
جمع کل	**نفر ۱۰۰۰**	**نفر ۴۴۰**	**۴۴% (حدود ۴۵%)**	**نفر ۱۰۰۰**	**نفر ۳۰۷**	**۳۰,۷% (حدود ۳۰%)**

بخش پنجم: فهرست کامل اشتباهات رایج آماری

بیایم حالا مهم‌ترین اشتباهاتی که آدم‌ها (حتی متخصص‌ها) توی تفسیر آمار مرتکب می‌شن رو مرور کنیم:

۲۴ .سوگیری نمونه‌گیری(Sampling Bias)

وقتی نمونه‌ای که انتخاب کردیم نماینده‌ی واقعی جامعه نیست.

مثال تاریخی معروف :توی انتخابات ریاست‌جمهوری آمریکا در سال ۱۹۳۶، یه مجله به اسم Literary Digest با نظرسنجی از ۲.۴ میلیون نفر (از طریق شماره تلفن و لیست خودروهای ثبت‌شده) پیش‌بینی کرد که «لندن» برنده می‌شه. ولی «روزولت» با اختلاف زیاد برنده شد! چرا؟ چون در اون سال‌ها، فقط افراد ثروتمند تلفن یا ماشین داشتن، و نمونه به شدت به سمت طبقه‌ی مرفه (که بیشتر به لندن رأی می‌دادن) سوگیری داشت. این یکی از معروف‌ترین شکست‌های تاریخ آمار و نظرسنجیه.

نظرسنجی لیتراری دایجست یکی از بزرگ‌ترین و پر هزینه‌ترین نظرسنجی‌هایی بود که تا آن زمان انجام شده بود؛ با حجم نمونه‌ای نزدیک به ۴/۲ میلیون نفر. جالب این که در همان زمان، جورج گالوپ توانست با حجم نمونه‌ی خیلی کمتر یعنی ۵۰ هزار نفر به خوبی و با خطایی اندک پیروزی روزولت را پیش بینی کند. پیش‌بینی غلط لیتراری دایجست نشان داد اگر روش نمونه‌گیری اشتباه باشد، افزایش حجم نمونه مشکل‌گشا نیست. در حقیقت مسئله اساسی در نمونه گیری، حجم نمونه نیست، این است که نمونه واقعا نماینده جامعه هدف باشد.

۲۵ .سوگیری بازمانده(Survivorship Bias)

وقتی فقط «موفق‌ها» رو بررسی می‌کنیم و شکست‌خورده‌ها رو نادیده می‌گیریم.

مثال معروف جنگ جهانی دوم :ارتش آمریکا هواپیماهای برگشته از مأموریت رو بررسی می‌کرد و می‌دید بیشتر گلوله‌ها به بدنه و بال‌ها خورده، پس تصمیم داشتن اونجاها رو زره‌پوش‌تر کنن. ولی آماردان آبراهام والد گفت: نه! باید جاهایی که **کمتر** گلوله خورده (مثل موتور) رو زره‌پوش کنیم، چون هواپیماهایی که از اونجا اصابت خورده بودن، اصلاً برنگشتن که بررسی بشن! ما فقط «بازمانده‌ها» رو می‌بینیم، نه کل جامعه.

مثال روزمره :گفتن «همه‌ی میلیاردرهای موفق دانشگاه رو نصفه رها کردن، پس دانشگاه لازم نیست» بدون در نظر گرفتن هزاران نفری که دانشگاه رو رها کردن و شکست خوردن و اسمشون هیچ‌جا نیومده.

۲۶ .مغالطه‌ی هم‌رخدادی یا اتاق(Conjunction Fallacy)

آدم‌ها گاهی فکر می‌کنن ترکیب دو شرط، محتمل‌تر از یه شرط تنهاست—که ریاضیاتاً همیشه غلطه چون P(A , B) هیچ‌وقت نمی‌تونه بزرگ‌تر از P(A) تنها باشه.

مثال معروف (مسئله‌ی لیندا) :به آدم‌ها توضیح می‌دن: «لیندا ۳۱ ساله، مجرد، صریح و خیلی باهوشه. فوق‌لیسانس فلسفه گرفته و دغدغه‌ی شدیدی نسبت به عدالت اجتماعی و تبعیض داشته.» بعد می‌پرسن کدوم محتمل‌تره: «لیندا یه صندوق‌دار بانکه» یا «لیندا یه صندوق‌دار بانکه و فعال جنبش فمینیستی هم هست»؟ اکثر مردم گزینه‌ی دوم رو انتخاب می‌کنن، چون با شخصیت لیندا هم‌خوانی بیشتری داره. ولی این ریاضیاتاً غلطه! چون «صندوق‌دار و فمینیست» یه زیرمجموعه از «صندوق‌دار» است، پس هیچ‌وقت نمی‌تونه محتمل‌تر از خود «صندوق‌دار» باشه.

۲۷ .تفسیر غلط ازp-value

توی پژوهش‌های علمی، p-valueکمتر از ۰.۰۵ معمولاً به عنوان «معنی‌دار آماری» در نظر گرفته می‌شه. ولی خیلی‌ها اشتباه فکر می‌کنن p-value یعنی «احتمال درست بودن فرضیه»—در حالی که این‌طور نیست p-value !یعنی «اگه فرض کنیم هیچ اثری وجود نداره (فرض صفر درست باشه)، چقدر احتمال داره داده‌هایی به این حدت یا حادتر مشاهده کنیم». این یه تفاوت ظریف ولی بسیار مهمه که حتی خیلی از محققان هم توش اشتباه می‌کنن.

مشکل :p-hacking بعضی محققان مدام آزمایش‌های مختلف رو امتحان می‌کنن تا بالاخره یه نتیجه با p کمتر از ۰.۰۵ پیدا کنن، بدون در نظر گرفتن اینکه اگه به اندازه‌ی کافی تست کنی، صرفاً به شانس هم به نتایج «معنی‌دار» کاذب می‌رسی.

۲۸ .تعمیم بیش از حد از نمونه‌ی کوچیک

هرچقدر نمونه کوچیک‌تر باشه، احتمال نوسان تصادفی (و در نتیجه نتیجه‌گیری غلط) بیشتره.

مثال عددی :فرض کن یه داروی جدید رو روی ۱۰ نفر امتحان می‌کنن و ۸ نفرشون بهبود پیدا می‌کنن (۸۰٪). آیا می‌شه گفت دارو ۸۰٪ مؤثره؟ نه لزوماً! با نمونه‌ی ۱۰ نفری، حاشیه‌ی خطا خیلی بزرگه (شاید بین ۴۵٪ تا ۹۵٪). ولی اگه همین دارو روی ۱۰,۰۰۰ نفر امتحان بشه و باز ۸۰٪ بهبود پیدا کنن، اون‌وقت اطمینان خیلی بیشتریه—چون طبق قانون اعداد بزرگ، نمونه‌ی بزرگ‌تر به واقعیت جامعه نزدیک‌تره.

۲۹ .چیدن گلچین(Cherry Picking)

انتخاب فقط داده‌هایی که ادعای ما رو تأیید می‌کنن و نادیده گرفتن بقیه.

مثال :یه شرکت می‌گه «فروش ما نسبت به ماه پیش ۵۰٪ رشد کرده!»، ولی نمی‌گه که ماه پیش به خاطر یه بحران خاص، فروش خیلی پایین بوده و در واقع نسبت به میانگین سالانه، هنوز افت داره. انتخاب نقطه‌ی مقایسه (baseline)می‌تونه کاملاً روایت رو تغییر بده.

۳۰ .سوگیری تأیید(Confirmation Bias)

تمایل به دیدن، جستجو و به‌خاطر سپردن اطلاعاتی که باورهای قبلی‌مون رو تأیید می‌کنه، و نادیده گرفتن شواهد مخالف. این یه سوگیری شناختیه که مستقیماً روی نحوه‌ی تفسیر آمار هم تأثیر می‌ذاره؛ آدم‌ها معمولاً آماری که با باورشون هم‌خونی داره رو راحت‌تر می‌پذیرن و در آماری که خلاف باورشونه، به دنبال ایراد می‌گردن.

۳۱ .گمراهی نمودارها(Misleading Graphs)

خیلی وقت‌ها با تغییر مقیاس محور Y یه نمودار (مثلاً شروع از عدد ۹۰ به جای صفر)، یه تغییر کوچیک می‌تونه خیلی بزرگ به نظر برسه.

مثال عددی :فرض کن فروش از ۱۰۰ به ۱۰۵ واحد رسیده (فقط ۵٪ رشد). اگه محور Y نمودار از ۹۵ تا ۱۱۰ رسم بشه (به جای از صفر تا ۱۱۰)، این ۵٪ رشد کوچیک، بصری شبیه یه جهش عظیم به نظر می‌رسه. این ترفند توی تبلیغات و حتی گزارش‌های خبری زیاد استفاده می‌شه.

جمع‌بندی

آمار فقط عدد و فرمول نیست؛ یه چارچوب فکریه برای اینکه گول ظاهر داده‌ها رو نخوریم. توی این مقاله از مفاهیم پایه (جامعه، نمونه، متغیرها) شروع کردیم، با محاسبات دقیق وارد دنیای میانگین، میانه، واریانس و انحراف معیار شدیم، بعد قدم به قدم احتمال رو با مثال‌های عددی کامل (از مغالطه‌ی قمارباز تا قضیه‌ی بیز) بررسی کردیم، به قوانین بزرگ آماری مثل قانون اعداد بزرگ، قضیه‌ی حد مرکزی و پارادوکس سیمپسون رسیدیم، و در آخر یه فهرست کامل از اشتباهات رایجی که حتی متخصص‌ها هم توشون می‌افتن رو مرور کردیم.

وقتی این مفاهیم رو بشناسیم، دیگه راحت‌تر می‌تونیم اخبار، آمارهای شبکه‌های اجتماعی، نتایج تست‌های پزشکی، تبلیغات و حتی تصمیم‌های شخصی زندگیمون رو با نگاه نقادانه‌تری بررسی کنیم.

قدم بعدی؟ دفعه‌ی بعد که یه آمار یا عدد جایی دیدی، از خودت این سؤال‌ها رو بپرس:

- این آمار از کجا اومده و نمونه‌اش چقدر بزرگ و تصادفی بوده؟
- آیا این یه میانگینه یا میانه؟ آیا داده‌ی پرتی وجود داره که میانگین رو گمراه‌کننده کرده؟
- آیا این یه رابطه‌ی همبستگیه یا واقعاً علّی؟ عامل سوم احتمالی چیه؟
- آیا این نتیجه، فقط بازگشت به میانگین یا نوسان تصادفیه، یا واقعاً یه روند معنی‌داره؟
- آیا نمودار یا ارائه‌ی این آمار، به شکلی طراحی شده که گمراه‌کننده باشه؟

همین سؤال‌های ساده، شروع تفکر آماریه و همین تفکره که تو رو از یه مصرف‌کننده‌ی منفعل اعداد، به یه تحلیل‌گر هوشمند تبدیل می‌کنه.