

ماژول اول آمار و احتمال

بخش پنجم - آمار استنباطی

المپیاد هوش مصنوعی

۱ نمونه‌گیری و استنباط آماری

۱ نمونه‌گیری و استنباط آماری

جمعیت (Population): مجموعه‌ای از تمام اشیاء یا افراد مورد نظر برای مطالعه است.

نمونه (Sample): زیرمجموعه‌ای از جمعیت که به صورت تصادفی یا هدفمند انتخاب می‌شود. در صورتی که نمونه شامل کل جمعیت باشد، به آن سرشماری (Census) گفته می‌شود.

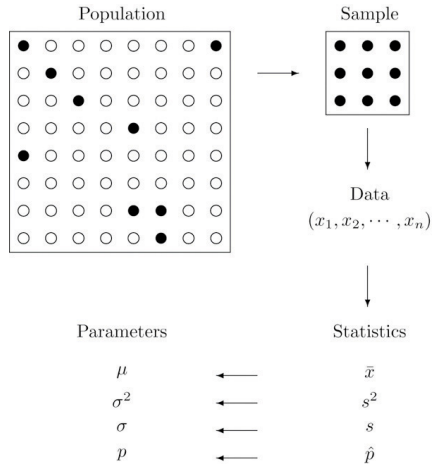
داده‌های نمونه

داده نمونه: هر اندازه‌گیری یا مشاهده‌ای است که از اعضای نمونه به دست می‌آید. این داده‌ها شامل اعداد یا ویژگی‌هایی هستند که برای تحلیل آماری جمع‌آوری می‌شوند.

پارامترها و آماره‌ها

پارامتر (Parameter): عدد یا ویژگی کلی که جمعیت را توصیف می‌کند (مانند میانگین یا واریانس).
آماره (Statistic): عددی که از داده‌های نمونه به دست آمده و برای برآورد پارامترهای جمعیت استفاده می‌شود.

تصویر نمونه



شکل ۱: نمونه‌ای از نمونه‌گیری

چالش‌های داده‌های تصادفی

در زندگی واقعی، داده‌ها به دلایل مختلف مانند خطاهای اندازه‌گیری، تغییرات زمانی و فرآیندهای نمونه‌گیری دارای ویژگی تصادفی هستند.

مثال: برای پیش‌بینی نتیجه انتخابات، از یک نمونه تصادفی از رأی‌دهندگان استفاده می‌کنیم. علاوه بر تصادفی بودن در انتخاب نمونه، نظرات افراد ممکن است در طول زمان تغییر کند. این شرایط نیازمند استفاده از روش‌های استنباط آماری برای نتیجه‌گیری درباره کل جمعیت است.

استنباط آماری (Statistical Inference) فرآیندی است که از داده‌های نمونه برای نتیجه‌گیری درباره ویژگی‌های جمعیت استفاده می‌کند. این روش‌ها به ما کمک می‌کنند تا با در نظر گرفتن عدم قطعیت ناشی از تصادفی بودن داده‌ها، تخمین‌های قابل اعتمادی ارائه دهیم.

چارچوب کلی استنباط آماری

مراحل اصلی استنباط آماری به شرح زیر است:

- ۱ تعریف پارامتر ناشناخته: مانند میانگین یا نسبت در جمعیت.
- ۲ جمع‌آوری داده‌های نمونه: انتخاب داده‌ها به صورت تصادفی از جمعیت.
- ۳ محاسبه آماره نمونه: مانند میانگین نمونه به عنوان برآورد پارامتر.
- ۴ ارائه نتیجه‌گیری: همراه با بیان عدم قطعیت (مثلاً با استفاده از فاصله‌های اطمینان).

دیدگاه frequentist در استنباط آماری

در رویکرد frequentist، فرض می‌شود که پارامترهای جمعیت (مانند θ) مقدارهای ثابتی هستند، در حالی که داده‌های نمونه به صورت تصادفی از جمعیت استخراج می‌شوند.

مثال: اگر θ درصد رأی‌دهندگان به یک کاندیدا باشد، برآورد $\hat{\theta} = \frac{Y}{n}$ (که Y تعداد رأی‌های مثبت است و n اندازه نمونه) یک متغیر تصادفی محسوب می‌شود که وابسته به نمونه انتخاب شده است.

در فرآیند جمع‌آوری داده‌ها، معمولاً چندین مشاهده از یک متغیر تصادفی ثبت می‌شود.
مثال: برای بررسی توزیع قد افراد، n نفر به صورت تصادفی انتخاب شده و قد آن‌ها به ترتیب X_1, X_2, \dots, X_n ثبت می‌شود.

اگرچه در عمل معمولاً از نمونه‌گیری بدون جایگزینی استفاده می‌شود، فرض نمونه‌گیری با جایگزینی به دلیل استقلال مشاهدات و سادگی تحلیل، به عنوان یک تقریب مناسب در نظر گرفته می‌شود.

برآوردگر نقطه‌ای

برآوردگر نقطه‌ای (Point Estimator): تابعی از داده‌های نمونه است که برای برآورد یک پارامتر ناشناخته به کار می‌رود.
مثال: برآورد میانگین قد جمعیت با استفاده از:

$$\hat{\Theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

این برآوردگر، بسته به نمونه انتخاب‌شده، ممکن است مقداری متفاوت داشته باشد؛ اما با افزایش اندازه نمونه، به مقدار واقعی نزدیک‌تر می‌شود.