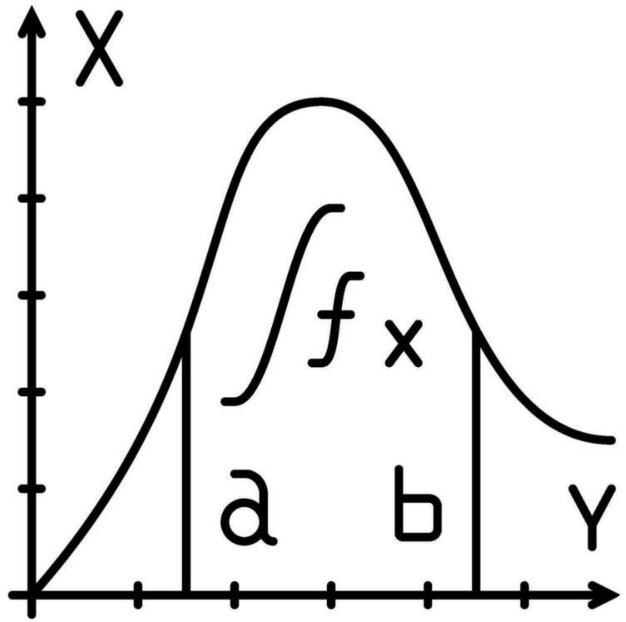
مبانی آمار و احتمال

هادی عاشری



میانگین متغیر تصادفی

- میانگین (مقدار مورد انتظار) نمایانگر مرکز مشاهدات (داده‌ها) است و به ما می‌گوید که «به طور متوسط چه انتظاری از مقدار متغیر تصادفی داریم
- میانگین برای متغیر گسسته

$$E[X] = \sum xP(x)$$

Iran Annual AI Award

واریانس متغیر تصادفی

- واریانس نشان می‌دهد که داده‌های واقعی چقدر نسبت به میانگین پراکنـده شده‌اند. اگر واریانس کم باشد، داده‌ها همگـن و نسـبتاً یکنواخـت هسـتند؛ ولی اگر واریانس زیاد باشد، داده‌ها نوسان زیادی دارند
- واریانس برای متغیر گسسته:

$$Var(X) = E[(X - E[X])^2]$$

$$= \sum_x (x - E[X])^2 P(x) = E[X^2] - (E[X])^2$$

- انحراف معیار جذر واریانس است و به طور مستقیم نشان می‌دهد داده‌ها به چه میزان از میانگین فاصله دارند.

متغیر تصادفی پیوسته

- میانگین در حالت پیوسته به صورت یک انتگرال تعریف می‌شود که نقش جمع در حالت گسسته را بازی می‌کند.

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

- واریانس برای متغیر پیوسته

$$Var[X] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x) dx$$

- یادآوری: برای متغیرهای گسسته از مجموع روی احتمالات و برای متغیرهای پیوسته از انتگرال روی تابع چگالی برای محاسبه میانگین و واریانس استفاده می‌شود.

Iran Annual AI Award

میانگین و واریانس توزیع‌های رایج

نام توزیع	تابع (چگالی/جرم) احتمال	میانگین $(E[X])$	واریانس $(Var[X])$
نرمال	$\mathcal{N}(\mu, \sigma^2)$	μ	σ^2
برنولی	$Bern(p)$	p	$p(1-p)$
دوجمله‌ای	$Bin(n.p)$	np	$np(1-p)$
پواسون	$Poisson(\lambda)$	λ	λ
نمایی	$Exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda}$

Iran Annual AI Award

توزیع نرمال توأمان

- توزیع نرمال توأمان/چند متغیره/برداری تعمیم توزیع نرمال یک‌متغیره است.
- برای بردار تصادفی $\boldsymbol{X} = [X_1 \quad X_2 \quad \cdots \quad X_n]^T$ تابع چگالی احتمال به‌صورت زیر تعریف می‌شود:

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- بردار میانگین: $\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_n]^T$
- ماتریس کوواریانس: $\Sigma_{n\times n} = [\sigma_{ij}]$

کواریانس بین دو متغیر

- در ماتریس کوواریانس هر عنصر قطری نشان‌دهنده واریانس متغیر نظیر آن است.
- هر عنصر غیرقطری نشان‌دهنده کوواریانس (همبستگی) بین دو متغیر نظیر آن است:

$$Cov(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

- کوواریانس مثبت نشان دهنده رفتار هم‌راستا بین دو متغیر رفتار است یعنی هنگامی که X_i از میانگین خود بالاتر است، X_j نیز از میانگین خود بالاتر است.
- کوواریانس منفی نشان دهنده رفتار غیرهم‌راستا است.

Iran Annual AI Award

ویژگی‌های توزیع نرمال

- هر ترکیب خطی از اجزای بردار X با توزیع نرمال، نیز دارای توزیع نرمال است.
- اگر ماتریس کوواریانس قطری باشد، اجزا مستقل از یکدیگر محسوب می‌شوند.
- قانون ۶۸-۹۵-۹۹.۷: تقریباً ۶۸ درصد از داده‌ها در فاصله یک انحراف معیار از میانگین، ۹۵ درصد در فاصله دو انحراف معیار و ۹۹.۷ درصد در فاصله سه انحراف معیار قرار می‌گیرند.

آمار توصیفی

- آمار توصیفی شاخه‌ای از آمار است که تمرکز آن بر جمع‌آوری، سازماندهی، خلاصه‌سازی و نمایش داده‌ها به صورتی است که فهم آن‌ها تسهیل شود.
- کاربردها:
  - خلاصه‌سازی و توصیف داده‌ها
  - تشخیص الگوها و روندهای موجود
  - پایه‌گذاری برای تحلیل‌های پیشرفته‌تر
  - کمک به تصمیم‌گیری در حوزه‌های مختلف
  - ابزار کمکی در تهیه گزارش و ارائه نتایج

Iran Annual AI Award

آمار استنباطی

- آمار استنباطی به استفاده از داده‌های نمونه برای تعمیم و استنباط ویژگی‌های کلی جامعه پرداخته و شامل نتیجه‌گیری‌های احتمالی و آزمون فرضیه است.
  - تخمین پارامتر: تخمین میانگین، واریانس و سایر شاخص‌ها برای جامعه از روی نمونه
  - آزمون فرضیه: تعیین صحت یک ادعا یا فرضیه در رابطه با جامعه آماری
  - فواصل اطمینان: ارائه دامنه‌ای که با یک سطح اطمینان مشخص، پارامتر واقعی داخل آن قرار می‌گیرد
- مزایا و چالش‌ها:
  - امکان تعمیم نتایج به کل جامعه
  - وابستگی به نمونه انتخابی و ریسک خطاهای نگراشی

شاخص‌های مرکزی – میانگین

- میانگین حاصل جمع تمامی داده‌ها تقسیم بر تعداد آن‌هاست.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- ویژگی‌ها:
  - حساس به داده‌های پرت
  - مناسب برای داده‌های دارای توزیع متقارن
  - نمایانگر «مرکز ثقل» داده‌ها

شاخص‌های مرکزی – میانه

- میانه نقطه‌ی وسط داده‌ها پس از مرتب‌سازی است؛ به طوری که نیمی از داده‌ها کمتر و نیمی بیشتر از آن مقدار هستند.
- ویژگی‌ها:
  - نسبت به داده‌های پرت مقاوم‌تر
  - نمایانگر موقعیت مرکزی در داده‌های نامتقارن
  - کاربرد در توزیع‌هایی که میانگین تحت تأثیر داده‌های بزرگ/کوچک قرار می‌گیرد
- روش محاسبه: داده‌ها را به ترتیب صعودی مرتب کرده و مقدار میانی را تعیین می‌کنیم. در صورتی که تعداد داده‌ها زوج باشد، میانگین دو میانه انتخاب می‌شود.

Iran Annual AI Award

شاخص‌های مرکزی – مد

- مد مقداری است که بیشترین فراوانی را در داده‌ها دارد.
- ویژگی‌ها:
  - مناسب برای داده‌های اسمی و رتبه‌بندی شده
  - امکان وجود چند مد (در صورت تکرار بیش از یک مقدار به یک اندازه)
  - تفسیر مستقیم „تند بودن“ یا الگوی تکرار در داده‌ها

شاخص‌های پراکندگی – واریانس

- واریانس میانگین مربعات فاصله تمام داده‌ها از میانگین است.
- واریانس جامعه آماری:

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

- واریانس نمونه آماری:

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- انحراف معیار جذر واریانس است و بیانگر فاصله متوسط داده‌ها از میـانگین می‌باشد.

نمودار هیستوگرام

- هیستوگرام نموداری میله‌ای است که فراوانی وقوع داده‌ها در بازه‌های مشخص (Bins) را نمایش می‌دهد.
- ویژگی‌ها:
  - نمایش تصویری از توزیع داده‌ها
  - انتخاب تعداد و ابعاد بین‌ها (Bins) بر نتایج تأثیرگذار است
  - ابزار مهم در شناسایی الگوهایی مانند تقارن، چولگی (Skewness) یا چند قله‌ای بودن (Multimodality)

Iran Annual AI Award

نمودار جعبه‌ای

- نمودار جعبه‌ای (Box Plot) ابزاری گرافیکی برای نمایش توزیع داده‌ها به کمک پنج عدد خلاصه (کمینه، چارک اول، میانه، چارک سوم، بیشینه) به همراه شناسایی نقاط پرت است.
- مولفه‌های نمودار جعبه‌ای:
  - جعبه (Box): نشان‌دهنده بازه بین چارک اول (Q1) و چارک سوم (Q3)
  - خط میانی: نشان‌دهنده میانه داده‌ها
  - "شوارز" (Whiskers): نمایش محدوده داده‌های موجود خارج از جعبه
  - نقاط پرت: داده‌هایی که به طور قابل توجهی خارج از بازه شوارز قرار دارند

مزایای نمودار جعبه‌ای

- ارائه سریع تصویر از پراکندگی داده
- مقایسه بین چند گروه داده
- شناسایی ناهنجاری‌ها و داده‌های خارج از روند

مقایسه نمودارهای هیستوگرام و جعبه‌ای

نمودار	مزایا	معایب
هیستوگرام	نمایش توزیع داده‌ها به‌صورت بصری نمایش شکل کلی داده‌ها مانند چولگی و چنداوجی مناسب برای داده‌های پیوسته	وابسته به تعداد و اندازه‌ی بازه‌ها سختی مقایسه‌ی مجموعه‌های داده‌ای مختلف
جعبه‌ای	خلاصه‌سازی آماری مؤثر نمایش مقدار میانه، چارک‌ها و دامنه بین‌چارکی تشخیص نقاط پرت به‌صورت مستقیم	نمایش توزیع کلی بدون نمایش جزئیات داخلی عدم نمایش شکل واقعی داده‌ها مانند چولگی

Iran Annual AI Award

اهمیت آمار استباطی در تحلیل داده

- این شاخه از آمار به ما کمک می‌کند تا از نمونه‌های داده‌ای، نتیجه‌گیری‌هایی درباره‌ی کل جامعه آماری انجام دهیم:
- بسیاری از الگوریتم‌های یادگیری ماشین مانند رگرسیون خطی یا مدل‌های بیز، از روش‌های استنباط آماری برای تخمین مقادیر بهینه‌ی پارامترها استفاده می‌کنند
- برای ارزیابی مدل‌ها و بررسی ارتباط بین متغیرها، آزمون‌های آماری استفاده می‌شود.
- مدل‌های داده‌محور همیشه با مقداری عدم قطعیت مواجه‌اند. آمار استنباطی به ما امکان بررسی میزان اعتبار و دقت نتایج را می‌دهد.
- با استفاده از روش‌های آماری، می‌توان تعیین کرد که کدام متغیرها تأثیر بیشتری بر پیش‌بینی‌های مدل دارند و ویژگی‌های غیرضروری را حذف کرد.
- تحلیل همبستگی و رگرسیون آماری، بینش‌های ارزشمندی درباره‌ی روابط میان متغیرها ارائه می‌دهد

Iran Annual AI Award

تخمین پارامتر

- تخمین پارامتر به معنای استفاده از داده‌های نمونه به‌دست آمده برای برآورد پارامترهای ناشناخته جامعه (مثل میانگین) است.
- تفاوت میان پارامتر و آمار
  - پارامتر عددی ثابت اما ناشناخته که ویژگی جامعه است در حالی که آمار عددی است که از روی داده‌های نمونه محاسبه می‌شود.
- انواع تخمین پارامتر
  - تخمین نقطه‌ای (Point Estimation) برای ارائه یک مقدار تخمینی از پارامتر (مثال: میانگین نمونه به عنوان تخمین میانگین جامعه).
  - تخمین فاصله‌ای (Interval Estimation) برای ارائه یک بازه (فاصله اطمینان) که پارامتر واقعی با احتمال مشخصی در آن قرار دارد.

Iran Annual AI Award

خواص مطلوب تخمین‌گرها

- **عدم سوگیری (Unbiasedness):**
  - میانگین تخمین‌ها برابر با مقدار واقعی پارامتر باشد.
- **سازگاری (Consistency):**
  - با افزایش اندازه نمونه، تخمین به مقدار واقعی نزدیک می‌شود.
- **کارایی (Efficiency):**
  - کمترین واریانس در میان تخمین‌گرهای بی‌سوگای ممکن.
- **کفایت (Sufficiency):**
  - استفاده کامل از اطلاعات موجود در داده‌های نمونه جهت تخمین پارامتر.

روش‌های اصلی تخمین پارامتر

- عدم سوگیری (Unbiasedness):
  - میانگین تخمین‌ها برابر با مقدار واقعی پارامتر باشد.
- سازگاری (Consistency):
  - با افزایش اندازه نمونه، تخمین به مقدار واقعی نزدیک می‌شود.
- کارایی (Efficiency):
  - کمترین واریانس در میان تخمین‌گرهای بی‌سوگای ممکن.
- کفایت (Sufficiency):
  - استفاده کامل از اطلاعات موجود در داده‌های نمونه جهت تخمین پارامتر.

Iran Annual AI Award

تابع درست‌نمایی

- اگر مجموعه نمونه‌ی $X = \{x_1, x_2, \cdots, x_N\}$ یک توزیع احتمال مشاهده شده باشد تابع درست‌نمایی به‌صورت زیر تعریف می‌شود:

$$L(\theta|\boldsymbol{X}) = P(\boldsymbol{X}|\theta)$$

- $L(\theta|X)$ مقدار احتمال شرطی داده‌های مشاهده‌شده با توجه به مقادیر پارامتر θ است.
- این تابع نشان می‌دهد که یک مجموعه داده چقدر تحت پارامترهای مفروض مدل، محتمل است.

کاربردهای تابع درست‌نمایی

- تخمین پارامتر:
  - تعیین مقداری از پارامتر که تابع درست‌نمایی داده‌های مشاهده‌شده را بیشینه کند (Maximum Likelihood Estimation - MLE).
- آزمون فرض آماری:
  - برای بررسی میزان تطابق داده‌ها با یک مدل آماری خاص.
- بهینه‌سازی مدل‌های یادگیری ماشین:
  - بسیاری از مدل‌های آماری و یادگیری ماشین بر اساس تخمین بیشینه‌ی درست‌نمایی آموزش داده می‌شوند.
- در حقیقت، تابع درست‌نمایی به ما امکان می‌دهد تا با استفاده از داده‌های نمونه، برآوردی بهینه از پارامترهای مدل داشته باشیم و احتمال وقوع مشاهدات را بهتر تحلیل کنیم.

Iran Annual AI Award

تخمین درست‌نمایی بیشینه

- در نظر بگیرید مجموعه‌ی داده مستقل $X = \{x_1, x_2, \cdots, x_N\}$ از یک توزیع نرمال ناشناخته را داریم.

$$\theta = \{\mu, \sigma^2\}$$

- ابتدا تابع درست‌نمایی را می‌نویسیم:

$$L(\mu, \sigma^2) = f(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

- بدست آوردن لگاریتم تابع درست‌نمایی (Log-Likelihood):

$$LL(\mu, \sigma^2) = \sum_{i=1}^{N} \log(f(x_i)) = \sum_{i=1}^{N} -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(x_i - \mu)^2$$

تخمین درست‌نمایی بیشینه

- برای تخمین مقادیر بهینه پارامترها، کافی است مشتق تابع لگاریتم را نسبت به هر پارامتر بدست‌آورده و هر یک را مساوی صفر قرار دهیم.

$$\frac{\partial}{\partial \mu} L(\mu, \sigma^2) = 0 \;\Rightarrow\; \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\frac{\partial}{\partial \mu} L(\mu, \sigma^2) = 0 \;\Rightarrow\; \widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2$$

استنباط آماری:رویکرد فراوانی‌گرا

- در رویکرد frequentist، احتمال به عنوان نسبت وقوع یک رویداد در میان تعداد بسیار زیاد تکرارهای یک آزمایش تعریف می‌شود.
- به عنوان مثال، وقتی می‌گوییم احتمال پرتاب سکه برابر ۰.۵ است، منظور این است که اگر سکه را بی‌نهایت بار پرت کنیم، نسبت دفعات وقوع شیر به تعداد کل پرتاب‌ها نزدیک به ۰.۵ خواهد شد.
- پارامترهای جامعه (مانند میانگین یا واریانس) ثابت هستند. مشکل ما در استنباط این پارامترها به دلیل ناشناخته بودن آن‌هاست، نه به دلیل تغییر آن‌ها.
- استدلال‌ها بر پایه نمونه‌های تصادفی و فرضیات مربوط به تکرارپذیری تجربیات استوارند. به همین دلیل، آزمون‌های فرض و فاصله‌های اطمینان بر اساس فراوانی نسبی داده‌های تکراری تعریف می‌شوند.

استنباط آماری:رویکرد باورگرا

- در رویکرد Bayesian، احتمال نشان‌دهندهٔ میزان باور یا اطمینان ما به وقوع یک رویداد یا صحت یک فرضیه است.
- برای مثال، اگر پیش از انجام یک آزمایش به این باور داشته باشیم که یک داروی جدید ۷۰٪ مؤثر است، این احتمال نشان‌دهنده‌ی اعتقاد اولیه ماست.
- برخلاف رویکردهای فراوانی‌گرا که پارامترها را ثابت فرض می‌کنند، در رویکرد Bayesian پارامترها نیز دارای توزیع احتمالی (توزیع پیشین یا prior distribution) هستند.
- با مشاهده داده‌های جدید، با استفاده از قضیه بیز، این توزیع به‌روز شده و به توزیع پسین (posterior distribution) تبدیل می‌شود که نمایانگر باور به‌روز شده ما نسبت به پارامتر است.

استنباط آماری:رویکرد باورگرا

- ویژگی جذاب روش Bayesian این است که شما می‌توانید اطلاعات قبلی یا دانش پیشین خود را وارد اعمال کنید و با مشاهده داده‌های جدید، باورهایتان را به‌صورت سیستماتیک به‌روز کنید. به عبارت دیگر، نتیجه‌گیری شما پویا بوده و با دریافت اطلاعات تازه تغییر می‌کند.

Iran Annual AI Award

خلاصه شهودی

- فراوانی‌گرا : تصور کنید یک ماشین شمارنده دارید. شما به تعداد زیادی آزمایش می‌پردازید و در نهایت نسبت وقوع رویدادها (مثلاً تعداد شیرها بر تعداد کل پرتاب‌ها) را محاسبه می‌کنید. در اینجا پارامترها ثابت هستند و هدف شما فقط مشاهده نتایج تکراری است.
- نگاه باورگرا: حال فرض کنید که قبل از شروع آزمایش‌ها، یک حدس یا باور اولیه دارید (مثلاً ۷۰٪ احتمال موفقیت برای یک دارو). با دریافت داده‌های جدید، این باور اولیه تغییر کرده و به‌صورت یک توزیع جدید (پسین) ارائه می‌شود. در این رویکرد، شما همزمان از دانش قبلی و اطلاعات تازه استفاده می‌کنید تا تصویر دقیق‌تری از وضعیت به‌دست آورید.

Iran Annual AI Award

قضیه بیز

- قانون بیز به‌صورت ریاضی:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \left(\frac{P(D|\theta)}{P(D)}\right)P(\theta)$$

توزیع احتمال	نام
$P(\theta\|D)$	توزیع پسین
$P(D\|\theta)$	تابع درست‌نمایی
$P(\theta)$	توزیع پیشین
$P(D)$	عامل نرمالسازی/شواهد

توزیع‌های پیشین

- توزیع پیشین (Prior) بازتاب‌دهنده باور اولیه درباره پارامترها پیش از مشاهده داده‌ها است.
- مثال‌ها:
  - توزیع یکنواخت
  - توزیع نرمال
- تعیین گرایش اولیه مدل و تأثیر مستقیم بر استنتاج‌های بعدی؛ در شرایطی که داده‌های کمی داریم، انتخاب پیشین مناسب اهمیت بیشتری پیدا می‌کند.

Iran Annual AI Award

توزیع‌های پسین

- توزیع پسین نتیجه به‌روزرسانی باورها پس از مشاهدهٔ داده‌ها است. به عبارت دیگر، پس از ترکیب توزیع پیشین و تابع درست‌نمایی از قضیه بیز به‌دست می‌آید.

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

- توزیع پسین نشان‌دهنده باور به یک پارامتر پس از اعمال شواهد جدید و کاهش عدم قطعیت نسبت به حالت پیشین است.

به‌روزرسانی باورها بر اساس شواهد جدید

مفهوم به‌روزرسانی:

- هر بار که داده یا شواهد جدید وارد می‌شود، توزیع پیشین قبلی به عنوان توزیع اولیه در نظر گرفته شده و پس از اعمال داده‌ها، به توزیع پسین جدید تبدیل می‌شود.

فرآیند تکراری:

- به‌روزرسانی مداوم باورها در هنگام دریافت شواهد جدید
- امکان پیش‌بینی بهتر و کاهش عدم قطعیت با گذشت زمان (مانند الگوریتم‌های یادگیری تدریجی).

تشخیص بیماری

- موضوع: بکارگیری استدلال بیزی در تشخیص بیماری با توجه به نتایج آزمایش‌های پزشکی.
- رویکرد بیزین:
  - پیشین: احتمال اولیه ابتلا به بیماری (مثلاً براساس شیوع در جامعه).
  - تابع درست‌نمایی: احتمال گرفتن نتیجه مثبت یا منفی آزمایش در افراد بیمار یا سالم.
  - تابع پسین: احتمال نهایی ابتلا به بیماری پس از دریافت نتایج آزمایش.
- اگر پیشین ابتلا به بیماری ۵٪ باشد و آزمون تشخیص با حساسیت و ویژگی بالا ارائه شود، نتیجه مثبت آزمون می‌تواند احتمال ابتلا را به طور قابل توجهی افزایش دهد.

iAAA
Iran Annual AI Award

Iran Annual AI Award

www.iaaa.ai
support@iaaa.ai
021-91096992
iaaa.event
iaaa_ai

باتشــــکر
از توجه شما