آمار و احتمال کاربردی در هوش مصنوعی

Attila Asghari

2025

attilaasghari@gmail.com

https://attilaasghari.github.io/

سرفصل ها

بخش اول

1. آمار چیست و چگونه نمونه‌ی آماری بگیریم؟
2. متغیر آماری و کاربر آن در علم داده
3. روش‌های نمونه‌گیری از جمعیت
4. گشتاورهای آماری + مثال داده‌های اوبر
5. پنج عدد طلایی برای شناخت داده‌ها
6. توزیع‌های آماری
7. مقایسه‌ی توزیع‌های آماری و کاربردهای آن‌ها (KL و JSH)
8. همبستگی (Correlation) و کاربردهای آن
9. کار با نرم افزار JASP

بخش دوم

10. تست الف-ب (A-B Test) آماری و کاربردهای آن
11. تست فرضیه آماری (Hypothesis Test) - قسمت اول
12. تست فرضیه آماری - قسمت دوم، تست Z و تست T
13. تست فرضیه آماری - قسمت سوم، تست Z و تست T 1
14. تست فرضیه آماری - قسمت چهارم، تست Z و تست T 1

15. تست فرضیه آماری - قسمت پنجم، تست Z و تست T 1
16. تست فرضیه آماری - قسمت ششم، تست Z و T 1 و Anova
17. تست فرضیه آماری - قسمت هفتم، تست Z و T و Anova
18. تست فرضیه آماری - قسمت هشتم، تست U

فصل اول: آمار چیست و چگونه نمونه‌ی آماری بگیریم؟

تعریف آمار

آمار شاخه‌ای از ریاضیات است که به جمع‌آوری، تحلیل، تفسیر، و ارائه‌ی داده‌ها می‌پردازد. این علم به دانشمندان و تحلیل‌گران کمک می‌کند تا با استفاده از اطلاعات، رفتار و ویژگی‌های پدیده‌ها را به‌صورت کمی توصیف کرده و از این طریق بتوانند نتایج خود را به کل جمعیت تعمیم دهند. آمار به دو بخش اصلی تقسیم می‌شود:

1. **آمار توصیفی (Descriptive Statistics):** این شاخه از آمار به خلاصه‌سازی و توصیف داده‌ها به وسیله‌ی معیارهایی مانند میانگین، میانه، انحراف معیار و نمودارها می‌پردازد.
2. **آمار استنباطی (Inferential Statistics):** این شاخه با استفاده از نمونه‌های آماری تلاش می‌کند به استنتاج درباره‌ی کل جمعیت بپردازد. در این بخش، از مفاهیمی همچون تخمین و تست فرضیه استفاده می‌شود.

نمونه‌گیری در آمار

نمونه‌گیری به معنای انتخاب بخشی از اعضای یک جمعیت است به‌نحوی که این بخش، نماینده‌ای از ویژگی‌های جمعیت باشد. نمونه‌گیری به کاهش زمان، هزینه، و منابع موردنیاز برای جمع‌آوری و تحلیل داده‌ها کمک می‌کند. روش‌های نمونه‌گیری در آمار به انواع مختلفی تقسیم می‌شود و انتخاب روش مناسب بستگی به هدف تحقیق و نوع داده‌ها دارد.

انواع روش‌های نمونه‌گیری

1. **نمونه‌گیری تصادفی ساده (Simple Random Sampling)**

در این روش، هر عضو از جمعیت شانس مساوی برای انتخاب شدن در نمونه را دارد. به این منظور می‌توان از جدول اعداد تصادفی یا نرم‌افزارهای تولید اعداد تصادفی استفاده کرد.

مثال عددی:

فرض کنید جمعیت مورد بررسی شامل ۱۰۰۰ نفر است و می‌خواهیم نمونه‌ای ۵۰ نفری به‌طور تصادفی انتخاب کنیم. برای این کار می‌توانیم به هر نفر از ۱ تا ۱۰۰۰ یک عدد اختصاص دهیم و سپس با استفاده از تولید اعداد تصادفی، ۵۰ عدد را انتخاب کنیم.

2. **نمونه‌گیری سیستماتیک (Systematic Sampling)**

در این روش، ابتدا جمعیت به صورت تصادفی مرتب می‌شود و سپس هر nامین عضو انتخاب می‌شود.

مثال عددی:

فرض کنید جمعیت مورد بررسی شامل ۱۰۰۰ نفر است و می‌خواهیم نمونه‌ای ۱۰۰ نفری از آن انتخاب کنیم. در این روش، باید هر ۱۰امین نفر را انتخاب کنیم (۱۰۰/۱۰۰۰=۱۰). پس از انتخاب اولین فرد به‌طور تصادفی، هر دهمین نفر بعد از آن انتخاب خواهد شد

3. **نمونه‌گیری طبقه‌ای (Stratified Sampling)**

در این روش، جمعیت به گروه‌هایی به نام طبقات تقسیم می‌شود و سپس نمونه‌هایی از هر طبقه به صورت تصادفی انتخاب می‌شوند. این روش زمانی مفید است که ویژگی‌هایی مانند سن، جنسیت یا سطح تحصیلات به‌طور گسترده در جمعیت متفاوت باشد.

مثال عددی:

فرض کنید جمعیت مورد بررسی شامل ۶۰۰ نفر است که ۴۰۰ نفر مرد و ۲۰۰ نفر زن هستند. اگر بخواهیم نمونه‌ای شامل ۶۰ نفر از این جمعیت بگیریم، ۴۰ نفر از مردان و ۲۰ نفر از زنان را به صورت تصادفی انتخاب می‌کنیم.

4. **نمونه‌گیری خوشه‌ای (Cluster Sampling)**

در این روش، جمعیت به گروه‌های کوچک‌تر یا خوشه‌ها تقسیم می‌شود و سپس چند خوشه به‌صورت تصادفی انتخاب شده و تمامی اعضای آن خوشه‌ها به نمونه اضافه می‌شوند.

مثال عددی:

فرض کنید می‌خواهیم از مدارس یک شهر نمونه‌گیری کنیم. ابتدا مدارس را به عنوان خوشه در نظر می‌گیریم و سپس تعدادی از این مدارس را به‌طور تصادفی انتخاب کرده و تمامی دانش‌آموزان این مدارس را برای نمونه در نظر می‌گیریم.

جمع‌بندی روش‌های نمونه‌گیری و انتخاب بهترین روش

روش مناسب نمونه‌گیری به هدف تحقیق و نوع داده‌ها بستگی دارد. برای مثال، اگر داده‌ها به‌صورت همگن در سراسر جمعیت پخش شده‌اند، روش نمونه‌گیری تصادفی ساده مناسب است؛ اما اگر جمعیت به گروه‌های متمایز تقسیم شده باشد، روش نمونه‌گیری طبقه‌ای بهتر عمل خواهد کرد.

مثال جامع از فرآیند نمونه‌گیری

فرض کنید یک شرکت حمل‌ونقل می‌خواهد میزان رضایت مشتریان خود را ارزیابی کند. این شرکت دارای ۲۰۰۰ مشتری فعال در ماه گذشته است و به‌دنبال صرفه‌جویی در زمان و هزینه، تصمیم می‌گیرد از ۲۰۰ مشتری به‌عنوان نمونه استفاده کند.

مراحل نمونه‌گیری:

1. **تعیین جمعیت:** جمعیت کل شامل ۲۰۰۰ مشتری است.
2. **تعیین حجم نمونه:** شرکت می‌خواهد ۲۰۰ مشتری را انتخاب کند.
3. **انتخاب روش نمونه‌گیری:** برای تضمین نمایندگی مناسب، روش نمونه‌گیری سیستماتیک انتخاب می‌شود.
4. **انتخاب نمونه:** از میان ۲۰۰۰ مشتری، اولین نفر به‌طور تصادفی انتخاب شده و سپس هر دهمین نفر پس از او در نمونه قرار می‌گیرد.

به این ترتیب، شرکت می‌تواند از نتایج این ۲۰۰ نفر برای استنتاج نتایج کل ۲۰۰۰ مشتری استفاده کند.

جمع‌بندی فصل اول

در این فصل با مفاهیم اساسی آمار و اهمیت نمونه‌گیری آشنا شدیم. روش‌های مختلف نمونه‌گیری از جمله تصادفی ساده، سیستماتیک، طبقه‌ای و خوشه‌ای را بررسی کردیم و برای هر روش یک مثال عددی ارائه شد. نمونه‌گیری ابزار قدرتمندی برای کاهش هزینه‌ها و زمان موردنیاز برای جمع‌آوری داده‌ها است و به محققین اجازه می‌دهد نتایج خود را به کل جمعیت تعمیم دهند.

فصل دوم: متغیر آماری و کاربرد آن در علم داده

تعریف متغیر آماری

متغیر آماری به خصوصیتی اطلاق می‌شود که می‌تواند مقادیر مختلفی را بپذیرد. به عبارت دیگر، متغیرها ویژگی‌هایی هستند که برای توصیف و مقایسه داده‌ها استفاده می‌شوند. متغیرها در علم داده نقشی کلیدی دارند؛ زیرا داده‌های خام را به‌صورت کمی یا کیفی بیان می‌کنند و مبنای بسیاری از تحلیل‌ها و پیش‌بینی‌ها قرار می‌گیرند.

انواع متغیرها در آمار

متغیرهای آماری به دو دسته کلی تقسیم می‌شوند:

۱. متغیرهای کمی (Quantitative Variables)

این متغیرها می‌توانند به‌صورت عددی اندازه‌گیری شوند و به دو نوع زیر تقسیم می‌شوند:

- **پیوسته (Continuous)**: متغیرهایی که می‌توانند هر مقدار در یک بازه عددی را بپذیرند. مثلاً قد، وزن، دما، و زمان.
- **گسسته (Discrete)**: متغیرهایی که فقط مقادیر عددی خاصی را می‌پذیرند. مثلاً تعداد کتاب‌های خوانده شده، تعداد خودروهای موجود در پارکینگ.

۲. متغیرهای کیفی (Qualitative Variables)

این متغیرها بیانگر ویژگی‌هایی هستند که نمی‌توان آنها را به‌صورت عددی اندازه‌گیری کرد. به دو نوع زیر تقسیم می‌شوند:

- **اسمی (Nominal)**: این متغیرها هیچ ترتیبی بین مقادیر خود ندارند. مثلاً رنگ چشم، نوع خودرو، جنسیت.
- **رتبه‌ای (Ordinal)**: این متغیرها دارای ترتیبی بین مقادیر خود هستند، اما فاصله بین مقادیر آنها مشخص نیست. مثلاً سطح تحصیلات (ابتدایی، متوسطه، دانشگاهی)، رتبه رضایت مشتری (کم، متوسط، زیاد).

مقیاس‌های اندازه‌گیری متغیرها

متغیرها به چهار مقیاس اندازه‌گیری دسته‌بندی می‌شوند که در تحلیل و تفسیر داده‌ها تأثیرگذار هستند:

- **مقیاس اسمی (Nominal Scale)**: این مقیاس برای متغیرهای اسمی به کار می‌رود که نمی‌توان آنها را به‌صورت کمی مقایسه کرد. مثلاً نوع خودرو (سواری، شاسی‌بلند، کامیونت).
- **مقیاس ترتیبی (Ordinal Scale)**: این مقیاس برای متغیرهای رتبه‌ای استفاده می‌شود که دارای ترتیب هستند، اما فاصله‌ها مشخص نیستند. مثلاً رتبه‌های شغلی (کارمند، سرپرست، مدیر).
- **مقیاس فاصله‌ای (Interval Scale)**: این مقیاس برای متغیرهای کمی است که فاصله‌ها ثابت هستند، اما نقطه صفر واقعی وجود ندارد. مثلاً دما در مقیاس سلسیوس (۰ درجه به معنای نبود دما نیست).
- **مقیاس نسبی (Ratio Scale)**: این مقیاس برای متغیرهای کمی است که فاصله‌ها ثابت هستند و نقطه صفر واقعی دارند. مثلاً وزن و قد (وزن ۰ به معنای نبود جرم است).

کاربرد متغیرهای آماری در علم داده

در علم داده، متغیرها به منظور تحلیل و ساخت مدل‌های آماری و یادگیری ماشین بسیار مهم هستند. در این بخش به نحوه استفاده از متغیرهای آماری و کاربرد آنها در علم داده می‌پردازیم:

1. **تحلیل توصیفی**: با استفاده از متغیرهای کمی و کیفی می‌توان به خلاصه‌سازی داده‌ها پرداخت. به‌عنوان مثال، میانگین، واریانس، و سایر معیارهای آماری را می‌توان برای متغیرهای کمی محاسبه کرد تا رفتار کلی داده‌ها را بهتر درک کنیم.
2. **ساخت مدل‌های پیش‌بینی**: متغیرها به عنوان ویژگی‌های ورودی در مدل‌های یادگیری ماشین استفاده می‌شوند. مدل‌ها از این ویژگی‌ها استفاده می‌کنند تا خروجی‌های مورد نظر را پیش‌بینی کنند. به عنوان مثال، در مدل‌های پیش‌بینی قیمت خانه، متغیرهایی مانند مساحت، تعداد اتاق‌ها، و محل جغرافیایی خانه نقش کلیدی دارند.
3. **تحلیل همبستگی و روابط بین متغیرها**: برای درک بهتر داده‌ها، بررسی روابط بین متغیرها بسیار مهم است. به عنوان مثال، ممکن است در داده‌های مربوط به سلامتی مشاهده کنیم که بین قد و وزن همبستگی وجود دارد.

به عنوان مثال، در تحلیل داده‌های مشتریان یک فروشگاه آنلاین، می‌توانیم از متغیرهای کمی مانند تعداد خریدها و مبلغ خرید، و از متغیرهای کیفی مانند جنسیت و ترجیح برند استفاده کنیم. این متغیرها در نهایت به ایجاد مدل‌های پیش‌بینی رفتار خرید کمک می‌کنند.

مثال‌های عددی و کاربرد عملی

مثال ۱: تحلیل داده‌های فروش

فرض کنید در یک فروشگاه آنلاین، داده‌های زیر در مورد خریدها ثبت شده است:

شناسه مشتری	جنسیت	تعداد خرید (متغیر گسسته)	مبلغ خرید (متغیر پیوسته)
1	مرد	3	120000
2	زن	5	250000
3	مرد	2	80000
4	زن	6	310000

محاسبه میانگین مبلغ خرید:

میانگین مبلغ خرید = (120000 + 250000 + 80000 + 310000) / 4 = 760000 / 4 = 190000

مثال ۲: تحلیل داده‌های سلامت

فرض کنید داده‌هایی در مورد بیماران در یک بیمارستان دارید که شامل قد، وزن و وضعیت سلامتی آنهاست:

شناسه بیمار	قد (cm) (متغیر پیوسته)	وزن (kg) (متغیر پیوسته)	وضعیت سلامتی (متغیر رتبه‌ای)
1	175	70	متوسط
2	160	55	خوب
3	180	80	ضعیف
4	170	65	خوب

محاسبه میانگین قد:

میانگین قد = (175 + 160 + 180 + 170) / 4 = 685 / 4 = 171.25

جمع‌بندی فصل دوم

در این فصل با مفهوم متغیرهای آماری و انواع مختلف آن‌ها آشنا شدیم. متغیرها به دسته‌های کمی (پیوسته و گسسته) و کیفی (اسمی و رتبه‌ای) تقسیم می‌شوند و هرکدام از این دسته‌ها برای توصیف و تحلیل داده‌ها مورد استفاده قرار می‌گیرند. همچنین با مقیاس‌های اندازه‌گیری متغیرها آشنا شدیم که

برای انتخاب روش تحلیل و تفسیر داده‌ها ضروری است. در پایان، با ارائه‌ی مثال‌هایی از تحلیل داده‌های فروش و داده‌های سلامت، کاربرد متغیرها در علم داده را به صورت عملی مشاهده کردیم.

فصل سوم: روش‌های نمونه‌گیری از جمعیت

مقدمه

در آمار و علم داده، نمونه‌گیری یکی از مهم‌ترین مراحل جمع‌آوری داده‌هاست. نمونه‌گیری به ما این امکان را می‌دهد که با انتخاب تعداد کمی از اعضای یک جمعیت بزرگ، نتایج را به کل جمعیت تعمیم دهیم. هدف از نمونه‌گیری، دست‌یابی به داده‌های نماینده‌ای از جمعیت با کمترین هزینه و زمان است. در این فصل، به بررسی انواع روش‌های نمونه‌گیری و مزایا و معایب هرکدام خواهیم پرداخت.

انواع روش‌های نمونه‌گیری

روش‌های نمونه‌گیری به دو دسته اصلی تقسیم می‌شوند: روش‌های نمونه‌گیری احتمالی و روش‌های نمونه‌گیری غیراحتمالی.

۱. نمونه‌گیری احتمالی (Probability Sampling)

در روش‌های نمونه‌گیری احتمالی، هر عضو از جمعیت شانس مشخصی برای انتخاب شدن در نمونه دارد. این روش‌ها برای اطمینان از نمایندگی داده‌ها و کاهش خطاها بسیار مناسب هستند.

نمونه‌گیری تصادفی ساده (Simple Random Sampling)

در این روش، هر عضو از جمعیت به‌صورت تصادفی انتخاب می‌شود و شانس یکسانی برای حضور در نمونه دارد. از ابزارهایی مانند جداول اعداد تصادفی یا نرم‌افزارهای تولید اعداد تصادفی برای انتخاب اعضا استفاده می‌شود.

مثال عددی: فرض کنید جمعیتی شامل ۱۰۰۰ نفر داریم و می‌خواهیم نمونه‌ای شامل ۵۰ نفر انتخاب کنیم. ابتدا به هر نفر از ۱ تا ۱۰۰۰ یک عدد اختصاص می‌دهیم و سپس به‌طور تصادفی ۵۰ عدد انتخاب می‌کنیم.

نمونه‌گیری سیستماتیک (Systematic Sampling)

در این روش، ابتدا یک عضو به‌طور تصادفی انتخاب می‌شود و سپس هر nامین نفر از جمعیت انتخاب می‌شود.

مثال عددی: فرض کنید می‌خواهیم از میان ۲۰۰۰ نفر، نمونه‌ای ۱۰۰ نفری بگیریم. اگر جمعیت را به ترتیب خاصی مرتب کنیم و اولین نفر را به‌صورت تصادفی انتخاب کنیم، سپس هر ۲۰امین نفر بعد از او را برای نمونه انتخاب می‌کنیم (۲۰=۲۰۰۰/۱۰۰).

نمونه‌گیری طبقه‌ای (Stratified Sampling)

در این روش، جمعیت به گروه‌هایی به نام طبقات تقسیم می‌شود که هر طبقه شامل اعضایی با ویژگی‌های مشابه است. سپس از هر طبقه به‌صورت تصادفی نمونه‌گیری می‌شود. این روش زمانی مفید است که جمعیت شامل گروه‌های مختلف با ویژگی‌های متفاوت باشد.

مثال عددی: فرض کنید یک جامعه آماری شامل ۶۰۰ نفر داریم که ۳۰۰ نفر زن و ۳۰۰ نفر مرد هستند. اگر بخواهیم نمونه‌ای ۶۰ نفری انتخاب کنیم، ۳۰ نفر از مردان و ۳۰ نفر از زنان را به‌صورت تصادفی انتخاب می‌کنیم.

نمونه‌گیری خوشه‌ای (Cluster Sampling)

در این روش، جمعیت به خوشه‌های کوچک‌تر تقسیم می‌شود و سپس چند خوشه به‌صورت تصادفی انتخاب می‌شوند. تمامی اعضای خوشه‌های انتخاب شده در نمونه قرار می‌گیرند.

مثال عددی: فرض کنید قصد دارید از ۱۰ منطقه شهری نمونه‌گیری کنید. ابتدا ۱۰ منطقه را به خوشه‌های کوچک‌تر تقسیم می‌کنیم و سپس چند خوشه را به‌صورت تصادفی انتخاب کرده و تمامی افراد آن خوشه‌ها را مورد بررسی قرار می‌دهیم.

۲. نمونه‌گیری غیراحتمالی (Non-Probability Sampling)

در این روش‌ها، اعضای نمونه به‌طور تصادفی انتخاب نمی‌شوند و شانس انتخاب برای هر عضو مشخص نیست. این روش‌ها در زمانی استفاده می‌شوند که دسترسی به جمعیت کامل دشوار باشد.

نمونه‌گیری راحتی (Convenience Sampling)

در این روش، اعضای نمونه بر اساس دسترسی و راحتی انتخاب می‌شوند. این روش سریع و ارزان است، اما نتایج ممکن است نماینده کل جمعیت نباشند.

مثال عددی: فرض کنید می‌خواهیم نظرات افراد درباره یک محصول جدید را بررسی کنیم. با انتخاب افرادی که در نزدیک‌ترین فروشگاه به ما هستند و دسترسی به آنها آسان است، نمونه‌گیری می‌کنیم.

نمونه‌گیری هدفمند (Purposive Sampling)

در این روش، اعضای نمونه بر اساس قضاوت و دانش محقق انتخاب می‌شوند و معمولاً افرادی که دارای ویژگی‌های خاصی هستند به نمونه اضافه می‌شوند.

مثال عددی: فرض کنید در یک پژوهش پزشکی، می‌خواهید افرادی که بیماری خاصی دارند را بررسی کنید. به‌جای انتخاب تصادفی، افرادی که این بیماری را دارند را انتخاب می‌کنید.

نمونه‌گیری گلوله برفی (Snowball Sampling)

این روش زمانی استفاده می‌شود که جمعیت هدف دشوار برای شناسایی باشد. در این روش، از اعضای نمونه اولیه درخواست می‌شود که افراد مشابه خود را معرفی کنند.

مثال عددی: اگر می‌خواهید افرادی را که در یک زیرگروه خاص از جامعه فعالیت دارند شناسایی کنید، از یک عضو نمونه می‌خواهید که افراد دیگری با ویژگی مشابه را به شما معرفی کند.

مقایسه روش‌های نمونه‌گیری احتمالی و غیراحتمالی

روش نمونه‌گیری	نوع نمونه‌گیری	مزایا	معایب
تصادفی ساده	احتمالی	نمایندگی دقیق جمعیت، ساده و قابل اعتماد	ممکن است زمان‌بر و هزینه‌بر باشد
سیستماتیک	احتمالی	سریع و ساده	در صورت الگوی خاص در جمعیت ممکن است باعث اریبی شود
طبقه‌ای	احتمالی	مناسب برای جمعیت‌های ناهمگن	نیاز به اطلاعات قبلی از جمعیت دارد
خوشه‌ای	احتمالی	هزینه کمتر، مناسب برای جمعیت‌های بزرگ	دقت کمتر نسبت به روش تصادفی ساده
راحتی	غیراحتمالی	سریع و ارزان	نمایندگی کمتر و احتمال بالای اریبی
هدفمند	غیراحتمالی	مناسب برای جمعیت‌های خاص	نتایج ممکن است تعمیم‌پذیری کمی داشته باشد
گلوله برفی	غیراحتمالی	مناسب برای جمعیت‌های دشوار برای دسترسی	احتمال بالای اریبی و تعمیم‌پذیری محدود

مثال جامع: انتخاب روش نمونه‌گیری برای یک تحقیق علمی

فرض کنید یک محقق می‌خواهد تأثیر استفاده از فناوری‌های آموزشی جدید را در میان دانشجویان دانشگاه‌های یک کشور بررسی کند. جمعیت مورد نظر دانشجویان کل دانشگاه‌ها هستند و نمونه باید نماینده‌ی دقیق جمعیت باشد.

مراحل نمونه‌گیری:

1. تعیین جمعیت: تمامی دانشجویان در دانشگاه‌های کشور.
2. انتخاب روش نمونه‌گیری: روش طبقه‌ای (Stratified Sampling) به دلیل ناهمگنی جمعیت (تفاوت در مقاطع تحصیلی و دانشگاه‌ها).
3. تقسیم‌بندی به طبقات: طبقه‌بندی دانشجویان بر اساس دانشگاه و مقطع تحصیلی.
4. انتخاب نمونه: از هر طبقه (دانشگاه و مقطع) به‌طور تصادفی تعدادی دانشجو انتخاب می‌شود.

به این ترتیب، محقق می‌تواند به نتایج قابل تعمیم برای کل جمعیت دست یابد.

جمع‌بندی فصل سوم

در این فصل، با انواع روش‌های نمونه‌گیری آشنا شدیم و مزایا و معایب هرکدام را بررسی کردیم. روش‌های احتمالی شامل نمونه‌گیری تصادفی ساده، سیستماتیک، طبقه‌ای و خوشه‌ای هستند که هرکدام برای شرایط خاصی مناسب هستند و نتایج قابل اطمینانی ارائه می‌دهند. روش‌های غیراحتمالی شامل نمونه‌گیری راحتی، هدفمند و گلوله برفی است که به دلیل دسترسی آسان‌تر، در شرایط محدودیت منابع یا زمان استفاده می‌شوند.

فصل چهارم: گشتاورهای آماری

مقدمه

گشتاورهای آماری از ابزارهای مهم در توصیف خصوصیات توزیع داده‌ها هستند و در تحلیل‌های آماری و داده‌کاوی کاربرد فراوانی دارند. گشتاورها به ما کمک می‌کنند تا ویژگی‌های مختلف یک توزیع، مانند میانگین، پراکندگی، چولگی، و کشیدگی را بررسی کنیم. در این فصل، به معرفی انواع گشتاورها و محاسبه آنها با استفاده از داده‌های فرضی شرکت اوبر می‌پردازیم.

گشتاور چیست؟

گشتاور یک معیار آماری است که برای توصیف و خلاصه‌سازی اطلاعات مربوط به یک توزیع داده به‌کار می‌رود. گشتاورها به دسته‌های زیر تقسیم می‌شوند:

- **گشتاور اول (میانگین)**: معرف مرکز توزیع داده‌ها.
- **گشتاور دوم (واریانس)**: بیانگر پراکندگی داده‌ها حول میانگین.
- **گشتاور سوم (چولگی)**: نشان‌دهنده عدم تقارن توزیع.
- **گشتاور چهارم (کشیدگی)**: میزان پهن یا باریک بودن توزیع را توصیف می‌کند.

گشتاور اول: میانگین (Mean)

گشتاور اول در واقع همان میانگین است که به عنوان نقطه مرکزی داده‌ها استفاده می‌شود. میانگین با استفاده از فرمول زیر محاسبه می‌شود:

$$\mu = (1/n) \sum_{i=1}^{n} x_i$$

که در آن x_i داده‌های نمونه و n تعداد داده‌ها است.

مثال عددی: فرض کنید تعداد سفرهای روزانه ۵ راننده اوبر به‌صورت زیر ثبت شده است: ۱۰، ۱۵، ۱۲، ۱۸، و ۲۰.

میانگین سفرها برابر است با:

$$\mu = (10 + 15 + 12 + 18 + 20) / 5 = 75 / 5 = 15$$

گشتاور دوم: واریانس (Variance)

گشتاور دوم، واریانس، پراکندگی داده‌ها را حول میانگین نشان می‌دهد. واریانس از فرمول زیر محاسبه می‌شود:

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

مثال عددی: برای داده‌های قبلی با میانگین ۱۵، واریانس به صورت زیر محاسبه می‌شود:

$$\sigma^2 = (1/5) [(10-15)^2 + (15-15)^2 + (12-15)^2 + (18-15)^2 + (20-15)^2] = (1/5) [25 + 0 + 9 + 9 + 25] = 13.6$$

گشتاور سوم: چولگی (Skewness)

گشتاور سوم، چولگی، نشان می‌دهد که آیا توزیع داده‌ها نسبت به میانگین متقارن است یا خیر. چولگی مثبت نشان‌دهنده تمایل داده‌ها به سمت راست و چولگی منفی نشان‌دهنده تمایل به سمت چپ است. فرمول چولگی به‌صورت زیر است:

$$\text{Skewness} = (1/n) \sum_{i=1}^{n}((x_i - \mu) / \sigma)^3$$

مثال عددی: فرض کنید داده‌های ما چولگی ۰.۵ دارند که نشان‌دهنده تمایل کم داده‌ها به سمت راست است.

گشتاور چهارم: کشیدگی (Kurtosis)

گشتاور چهارم، کشیدگی، میزان پهن یا باریک بودن توزیع داده‌ها را نشان می‌دهد. کشیدگی بیشتر از صفر نشان‌دهنده توزیع با دم‌های سنگین‌تر و کشیدگی کمتر از صفر نشان‌دهنده توزیع با دم‌های سبک‌تر است. فرمول کشیدگی به‌صورت زیر است:

$$\text{Kurtosis} = (1/n) \sum_{i=1}^{n}((x_i - \mu) / \sigma)^4 - 3$$

مثال عددی: اگر داده‌ها دارای کشیدگی ۲ باشند، این نشان می‌دهد که توزیع داده‌ها دم‌های نسبتا سنگینی دارد.

مثال کاربردی از داده‌های اوبر

فرض کنید داده‌های مربوط به مدت زمان سفرهای رانندگان اوبر به‌صورت زیر است:

راننده	مدت زمان سفر (دقیقه)
A	12
B	18
C	15
D	20
E	10

محاسبه میانگین:

$\mu = (12 + 18 + 15 + 20 + 10) / 5 = 75 / 5 = 15$

محاسبه واریانس:

$\sigma^2 = (1/5) [(12-15)^2 + (18-15)^2 + (15-15)^2 + (20-15)^2 + (10-15)^2] = (1/5) [9 + 9 + 0 + 25 + 25] = 13.6$

محاسبه چولگی (Skewness):

فرمول چولگی به‌صورت زیر است:

$\text{Skewness} = (1/n) \sum_{i=1}^{n}((x_i - \mu) / \sigma)^3$

حال با جایگذاری مقادیر:

$\text{Skewness} = (1/5) [((12 - 15) / 3.69)^3 + ((18 - 15) / 3.69)^3 + ((15 - 15) / 3.69)^3 + ((20 - 15) / 3.69)^3 + ((10 - 15) / 3.69)^3]$

محاسبه کشیدگی (Kurtosis):

فرمول کشیدگی به‌صورت زیر است:

$\text{Kurtosis} = (1/n) \sum_{i=1}^{n}((x_i - \mu) / \sigma)^4 - 3$

حال با جایگذاری مقادیر:

$\text{Kurtosis} = (1/5) [((12 - 15) / 3.69)^4 + ((18 - 15) / 3.69)^4 + ((15 - 15) / 3.69)^4 + ((20 - 15) / 3.69)^4 + ((10 - 15) / 3.69)^4] - 3$

نتایج نهایی برای داده‌های اوبر:

- میانگین (Mean): 15
- واریانس (Variance): 13.6
- انحراف معیار (Standard Deviation): تقریباً 3.69
- چولگی (Skewness): 0.0
- کشیدگی (Kurtosis): -1.47

این مقادیر نشان می‌دهند که داده‌ها چولگی ندارند (چولگی برابر صفر)، و دارای توزیعی هستند که از لحاظ کشیدگی دم‌های سبک‌تری نسبت به یک توزیع نرمال دارند (کشیدگی منفی).

جمع‌بندی فصل چهارم

در این فصل، با مفهوم گشتاورهای آماری و کاربرد آن‌ها در تحلیل داده‌ها آشنا شدیم. گشتاور اول (میانگین) به‌عنوان مرکز داده، گشتاور دوم (واریانس) به‌عنوان پراکندگی، گشتاور سوم (چولگی) به‌عنوان عدم تقارن، و گشتاور چهارم (کشیدگی) به‌عنوان پهنی یا باریکی توزیع به ما کمک می‌کنند. با مثال داده‌های اوبر، روش‌های محاسبه این گشتاورها را در عمل مشاهده کردیم.

فصل پنجم: پنج عدد طلایی برای شناخت داده‌ها

معرفی پنج عدد طلایی

این پنج عدد به‌عنوان ابزارهایی اساسی در آمار توصیفی به ما کمک می‌کنند تا نگاه اولیه‌ای به توزیع و پراکندگی داده‌ها داشته باشیم.

1. **کمینه (Minimum)**

   کمینه یا حداقل داده، کوچک‌ترین مقداری است که در مجموعه داده‌ها یافت می‌شود. این عدد نشان‌دهنده‌ی پایین‌ترین حد از دامنه داده‌هاست و به تحلیل‌گر کمک می‌کند تا مقدار کمینه را به‌عنوان یک شاخص از حد پایینی مقادیر در نظر بگیرد.
2. **چارک اول (First Quartile - Q1)**

   چارک اول، نقطه‌ای است که ۲۵٪ از داده‌ها پایین‌تر از آن قرار می‌گیرند. این عدد به تحلیل‌گر اطلاعات مفیدی از توزیع داده‌ها در قسمت پایینی آن ارائه می‌دهد و گاهی به عنوان نقطه آستانه‌ای برای شناسایی نقاط دورافتاده نیز کاربرد دارد.
3. **میانه (Median - Q2)**

   میانه، عدد میانی داده‌ها است. برای داده‌هایی که به‌ترتیب صعودی مرتب شده‌اند، میانه نقطه‌ای است که نیمی از داده‌ها پایین‌تر و نیمی بالاتر از آن قرار می‌گیرند. اگر تعداد داده‌ها فرد باشد، میانه مقدار میانی خواهد بود و اگر تعداد داده‌ها زوج باشد، میانه میانگین دو مقدار میانی است.

   فرمول محاسبه میانه

   فرض کنید داده‌ها مرتب‌شده‌اند و n تعداد کل داده‌هاست. اگر n فرد باشد، میانه داده‌ی $(n+1)/2$-ام است. و اگر n زوج باشد، میانه میانگین داده‌های $n/2$-ام و $(n/2+1)$-ام است.
4. **چارک سوم (Third Quartile - Q3)**

   چارک سوم نقطه‌ای است که ۷۵٪ از داده‌ها کمتر از آن هستند. این عدد اطلاعاتی درباره توزیع داده‌ها در ناحیه بالایی ارائه می‌دهد و در ترکیب با چارک اول می‌تواند به تحلیل‌گر کمک کند تا نوسانات و دامنه پراکندگی داده‌ها را بهتر درک کند.

محاسبه‌ی چارک‌ها (Q1 و Q3)

چارک‌ها داده‌ها را به چهار بخش مساوی تقسیم می‌کنند:

چارک اول (Q1): مقدار وسط ۲۵٪ ابتدایی داده‌ها.

چارک سوم (Q3): مقدار وسط ۷۵٪ ابتدایی داده‌ها.

فرمول محاسبه چارک‌ها

برای محاسبه چارک‌ها روش‌های مختلفی وجود دارد. یکی از روش‌ها (روش تفسیری) به این صورت است:

داده‌ها را مرتب کنید.

چارک اول (Q1): اگر تعداد داده‌ها n باشد، Q1 تقریباً برابر است با مقدار $(n+1)/4$-ام. اگر این مقدار عدد صحیح نباشد، از میانگین دو مقدار نزدیک استفاده می‌شود.

چارک سوم (Q3): با محاسبه مقدار $3*(n+1)/4$ به دست می‌آید.

5. **بیشینه (Maximum)**

بیشینه یا حداکثر داده، بزرگ‌ترین مقداری است که در مجموعه داده‌ها یافت می‌شود و نمایان‌گر حد بالای داده‌هاست.

کاربرد پنج عدد طلایی و فاصله چارکی (IQR)

با استفاده از فاصله چارکی (IQR) می‌توانیم پراکندگی و نقاط دورافتاده را شناسایی کنیم. IQR به‌صورت زیر محاسبه می‌شود:

$$IQR = Q3 - Q1$$

تشخیص نقاط دورافتاده (Outliers):

نقاط دورافتاده، داده‌هایی هستند که به‌طور غیرعادی پایین‌تر یا بالاتر از بقیه داده‌ها قرار می‌گیرند و می‌توانند اطلاعات مهمی را درباره داده‌ها یا خطاهای احتمالی نشان دهند. برای شناسایی نقاط دورافتاده، از بازه زیر استفاده می‌کنیم:

$$\text{محدوده پایین‌تر} = Q1 - 1.5 \times IQR$$

$$\text{محدوده بالاتر} = Q3 + 1.5 \times IQR$$

داده‌هایی که بیرون از این محدوده‌ها قرار می‌گیرند، به‌عنوان نقاط دورافتاده در نظر گرفته می‌شوند.

مثال کاربردی از داده‌های اوبر (با جزئیات بیشتر)

فرض کنید مجموعه داده‌ای از مدت زمان سفرهای رانندگان اوبر به‌صورت زیر داریم:

10، 12، 15، 18، 20، 25، 30، 35، 40

محاسبه پنج عدد طلایی

- **کمینه:** کوچک‌ترین مقدار، ۱۰ است.
- **بیشینه:** بزرگ‌ترین مقدار، ۴۰ است.
- **میانه (Q2):**

داده‌ها ۹ عدد هستند، پس میانه برابر است با داده میانی (داده پنجم)، یعنی ۲۰.

- **چارک اول (Q1):**

برای محاسبه Q1، به داده‌های پایین‌تر از میانه نگاه می‌کنیم: 10، 12، 15، 18
میانه این داده‌ها (چارک اول) برابر با 12 است.

- **چارک سوم (Q3):**

برای محاسبه Q3، به داده‌های بالاتر از میانه نگاه می‌کنیم: 25، 30، 35، 40
میانه این داده‌ها (چارک سوم) برابر با 30 است.

فاصله چارکی (IQR):

$$IQR = Q3 - Q1 = 30 - 12 = 18$$

شناسایی نقاط دورافتاده

اکنون با استفاده از فاصله چارکی، محدوده شناسایی نقاط دورافتاده را محاسبه می‌کنیم:

$$\text{محدوده پایین‌تر: } Q1 - 1.5 \times IQR = 12 - 1.5 \times 18 = 12 - 27 = -15$$

$$\text{محدوده بالاتر: } Q3 + 1.5 \times IQR = 30 + 1.5 \times 18 = 30 + 27 = 57$$

با توجه به این محاسبات، هیچ داده‌ای خارج از محدوده $[-15, 57]$ قرار نمی‌گیرد، بنابراین در این مثال، نقطه دورافتاده‌ای نداریم.

کاربرد پنج عدد طلایی در تجسم داده‌ها: نمودار جعبه‌ای (Box Plot)

نمودار جعبه‌ای یکی از ابزارهای بصری است که برای نمایش توزیع داده‌ها با استفاده از پنج عدد طلایی طراحی شده است. این نمودار شامل جعبه‌ای است که چارک اول و سوم را نمایش می‌دهد و خطی در داخل جعبه که میانه را نشان می‌دهد. همچنین دو خط (دم‌ها) از جعبه به سمت کمینه و بیشینه گسترش پیدا می‌کنند و نقاط دورافتاده با علامت‌های خاص (مثل دایره یا ستاره) نمایش داده می‌شوند.

نمودار جعبه‌ای به تحلیل‌گر کمک می‌کند تا چولگی، پراکندگی و نقاط دورافتاده را به‌طور بصری مشاهده و تحلیل کند.

جمع‌بندی فصل پنجم

در این فصل با پنج عدد طلایی و کاربردهای آن‌ها در تحلیل داده‌ها آشنا شدیم. این پنج عدد شامل کمینه، چارک اول، میانه، چارک سوم و بیشینه است که نمایی کلی از توزیع و پراکندگی داده‌ها ارائه می‌دهند. فاصله چارکی (IQR) به ما کمک می‌کند تا نقاط دورافتاده را شناسایی کنیم و با استفاده از نمودار جعبه‌ای می‌توانیم به‌طور بصری توزیع داده‌ها را تحلیل کنیم.

فصل ششم: توزیع‌های آماری

مقدمه

توزیع‌های آماری در علم داده نقشی اساسی دارند و به تحلیل‌گر کمک می‌کنند تا با رفتار و ساختار داده‌ها آشنا شود و بتواند مدل‌سازی آماری و پیش‌بینی‌ها را با دقت بیشتری انجام دهد. توزیع‌های آماری، تابعی از داده‌ها هستند که نشان می‌دهند چگونه احتمال وقوع مقادیر مختلف یک متغیر تصادفی در مجموعه داده توزیع شده است.

در این فصل، چهار توزیع مهم و پرکاربرد شامل توزیع نرمال (گوسی)، برنولی، دوجمله‌ای و چندجمله‌ای را معرفی و کاربردهای آن‌ها را در علم داده توضیح می‌دهیم.

توزیع نرمال (گوسی)

تعریف و فرمول

توزیع نرمال که با نام توزیع گوسی نیز شناخته می‌شود، از پرکاربردترین توزیع‌های پیوسته در آمار است. این توزیع با میانگین (μ) و انحراف معیار (σ) مشخص می‌شود و تابع چگالی احتمال آن به صورت زیر است:

$$f(x) = 1 / (\sqrt{(2\pi\sigma^2)}) * e^{(-((x-\mu)^2 / (2\sigma^2)))}$$

این توزیع دارای شکل زنگوله‌ای است و بیشترین مقدار احتمال در نقطه‌ی میانگین قرار دارد. در این توزیع، هر دو طرف میانگین تقارن دارند.

کاربردها

- مدل‌سازی پدیده‌های طبیعی: بسیاری از داده‌های مربوط به پدیده‌های طبیعی مانند قد، وزن و نمرات امتحانات به توزیع نرمال نزدیک هستند.
- فرضیات آماری: در بسیاری از آزمون‌های آماری، مانند تست‌های فرضیه، از توزیع نرمال استفاده می‌شود.

مثال عددی

فرض کنید داده‌هایی از قد دانشجویان یک کلاس داریم که دارای میانگین (μ = 170) سانتی‌متر و انحراف معیار (σ = 10) سانتی‌متر هستند. احتمال اینکه قد یک دانشجو در محدوده 160 تا 180 سانتی‌متر باشد با استفاده از توزیع نرمال به‌سادگی قابل محاسبه است.

توزیع برنولی

تعریف و فرمول

توزیع برنولی یک توزیع گسسته است که تنها دو نتیجه ممکن دارد: موفقیت (۱) یا شکست (۰). این توزیع با احتمال موفقیت (p) تعریف می‌شود و احتمال شکست برابر (1-p) است. تابع احتمال توزیع برنولی به‌صورت زیر تعریف می‌شود:

$$P(X = x) = p^x * (1 - p)^(1 - x), x \in \{0, 1\}$$

کاربردها

- آزمایش‌های دوتایی: مانند پرتاب سکه که در آن نتیجه تنها شیر یا خط است.
- پیش‌بینی‌های دودویی: در مدل‌های یادگیری ماشین برای دسته‌بندی دودویی، از توزیع برنولی استفاده می‌شود.

مثال عددی

فرض کنید احتمال موفقیت یک دانش‌آموز در امتحان ریاضی ۰٫۷ باشد. توزیع برنولی می‌تواند پیش‌بینی کند که احتمال قبول شدن یا رد شدن دانش‌آموز در امتحان چقدر است.

توزیع دوجمله‌ای

تعریف و فرمول

توزیع دوجمله‌ای تعداد موفقیت‌ها در یک سری از آزمایش‌های مستقل با احتمال موفقیت ثابت (p) و تعداد آزمایش‌های (n) را مدل می‌کند. تابع احتمال آن به صورت زیر است:

$$P(X = k) = \binom{n}{k} * p^k * (1 - p)^{(n - k)}$$

که در آن (k) تعداد موفقیت‌ها و $\binom{n}{k}$ ترکیب‌های ممکن برای انتخاب (k) موفقیت از (n) آزمایش است.

کاربردها

- مدل‌سازی تعداد موفقیت‌ها: مانند تعداد افراد قبولی در یک آزمون از میان تعداد مشخصی داوطلب.
- تحلیل داده‌های گسسته: در بسیاری از مسائل، از توزیع دوجمله‌ای برای تحلیل داده‌های دسته‌بندی‌شده استفاده می‌شود.

مثال عددی

فرض کنید احتمال موفقیت در یک امتحان (p = 0.8) باشد و تعداد کل امتحانات (n = 10) است. توزیع دوجمله‌ای می‌تواند احتمال قبول شدن دقیقاً (k = 7) امتحان را محاسبه کند.

توزیع چندجمله‌ای

تعریف و فرمول

توزیع چندجمله‌ای تعمیمی از توزیع دوجمله‌ای است که در آن نتایج ممکن بیش از دو دسته دارند. به‌جای داشتن تنها دو نتیجه (موفقیت و شکست)، این توزیع می‌تواند دسته‌های مختلفی با احتمالات مختلف داشته باشد. احتمال در توزیع چندجمله‌ای به‌صورت زیر محاسبه می‌شود:

$$P(X1 = k1, X2 = k2, ..., Xr = kr) = (n!) / (k1! * k2! * ... * kr!) * (p1^{k1} * p2^{k2} * ... * pr^{kr})$$

کاربردها

- تحلیل دسته‌های چندگانه: مانند تحلیل نظرسنجی‌ها که در آن نظرات به بیش از دو دسته تقسیم می‌شوند.
- مدل‌سازی دسته‌های متعدد در علم داده: از توزیع چندجمله‌ای برای مدل‌سازی داده‌های دسته‌ای با چندین نتیجه ممکن استفاده می‌شود.

مثال عددی

فرض کنید در یک نظرسنجی، پاسخ‌ها به سه دسته "خوب"، "متوسط" و "بد" تقسیم شده‌اند. احتمال اینکه از بین ۱۰ پاسخ، ۵ نفر "خوب"، ۳ نفر "متوسط" و ۲ نفر "بد" را انتخاب کنند، با استفاده از توزیع چندجمله‌ای محاسبه می‌شود.

مقایسه توزیع‌ها و کاربرد آن‌ها

هر یک از این توزیع‌ها به‌طور خاص برای انواع مختلف داده‌ها مناسب هستند:

- توزیع نرمال: برای داده‌های پیوسته و زمانی که داده‌ها به‌صورت زنگوله‌ای توزیع شده‌اند.
- توزیع برنولی: برای داده‌های دودویی و آزمایش‌هایی که دو نتیجه ممکن دارند.
- توزیع دوجمله‌ای: برای تعداد موفقیت‌ها در یک سری آزمایش با تعداد مشخص.
- توزیع چندجمله‌ای: برای دسته‌بندی‌های چندگانه.

جمع‌بندی فصل ششم

در این فصل، با چندین توزیع آماری پایه و پرکاربرد آشنا شدیم. این توزیع‌ها ابزارهای اصلی برای تحلیل و مدل‌سازی داده‌ها در آمار و علم داده هستند. با شناخت این توزیع‌ها، تحلیل‌گران می‌توانند با دقت بیشتری به مدل‌سازی و پیش‌بینی داده‌ها بپردازند.

فصل هفتم: مقایسه‌ی توزیع‌های آماری و کاربردهای آن‌ها

مقدمه

در علم داده و یادگیری ماشین، گاهی نیاز داریم شباهت یا تفاوت بین دو توزیع آماری را اندازه‌گیری کنیم. این کار به ما کمک می‌کند تا بفهمیم داده‌ها تا چه حد از توزیع‌های نظری پیروی می‌کنند، یا اینکه دو

مجموعه داده چقدر به هم شبیه‌اند. روش‌های مختلفی برای مقایسه‌ی توزیع‌ها وجود دارد، اما دو معیار متداول KL Divergence و Jensen-Shannon Divergence (JSD) هستند. در این فصل، ابتدا این دو معیار را معرفی می‌کنیم و سپس کاربردهای آن‌ها را در مسائل مختلف بررسی خواهیم کرد.

KL-Divergence

تعریف و فرمول

KL-Divergence (Kullback-Leibler Divergence) معیاری برای سنجش تفاوت بین دو توزیع احتمال است. این معیار بیشتر در مواردی استفاده می‌شود که می‌خواهیم میزان تفاوت بین یک توزیع تجربی (مثلاً داده‌های نمونه‌گیری شده) و یک توزیع تئوریک (مثلاً توزیع نرمال) را اندازه‌گیری کنیم. KL-Divergence به‌صورت زیر تعریف می‌شود:

$$D_{KL}(P\|Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

در اینجا:

- $P(i)$ و $Q(i)$ به ترتیب مقادیر احتمال در توزیع‌های P و Q هستند.

اگر دو توزیع P و Q کاملاً یکسان باشند، مقدار KL برابر صفر خواهد بود. هر چه این مقدار بیشتر شود، نشان‌دهنده‌ی تفاوت بیشتر بین دو توزیع است.

کاربردها

- تخمین تفاوت بین داده‌ها و مدل‌های تئوریک: KL-Divergence برای مقایسه‌ی داده‌های مشاهده شده با مدل‌های آماری مناسب است. مثلاً اگر داده‌ها به‌طور ایده‌آل نرمال باشند، می‌توان میزان اختلاف با این توزیع را اندازه گرفت.
- مدل‌سازی در یادگیری ماشین: در الگوریتم‌هایی که به تطبیق توزیع داده‌ها با توزیع هدف نیاز دارند، KL-Divergence کاربردی است.

مثال عددی

فرض کنید توزیع احتمال یک نمونه‌ی آزمایشی P و توزیع تئوریک Q به‌صورت زیر باشند:

$$P = \{0.4, 0.6\} \quad \text{و} \quad Q = \{0.5, 0.5\}$$

KL-Divergence برای این دو توزیع به‌صورت زیر محاسبه می‌شود:

$$D_{KL}(P\|Q) = 0.4 \log\left(\frac{0.4}{0.5}\right) + 0.6 \log\left(\frac{0.6}{0.5}\right)$$

Jensen-Shannon Divergence (JSD)

تعریف و فرمول

Jensen-Shannon Divergence یک معیار متقارن برای مقایسه‌ی دو توزیع است که به کمک KL-Divergence محاسبه می‌شود. بر خلاف KL-Divergence، JSD همیشه مقداری محدود بین ۰ و ۱ دارد و به همین دلیل برای سنجش شباهت بین توزیع‌های مختلف مناسب‌تر است. فرمول JSD به صورت زیر است:

$$JSD(P\|Q) = \frac{1}{2}D_{KL}(P\|M) + \frac{1}{2}D_{KL}(Q\|M)$$

که در آن:

$$M = \frac{P+Q}{2}$$

این معیار به دلیل متقارن بودن ($JSD(P\|Q) = JSD(Q\|P)$) برای مواردی که توزیع‌ها نیاز به یکسان‌سازی دارند، مفید است.

کاربردها

- اندازه‌گیری شباهت توزیع‌ها: در تحلیل داده‌ها و یادگیری ماشین، JSD برای سنجش شباهت بین دو توزیع داده‌ای به‌کار می‌رود. این معیار به‌خصوص در پردازش زبان طبیعی و تحلیل متون کاربرد دارد.
- تحلیل‌های آماری در سنجش مدل‌ها: در مواقعی که دو مدل احتمال متفاوت داریم و می‌خواهیم میزان شباهت آن‌ها را مقایسه کنیم، از JSD استفاده می‌شود.

مثال عددی

فرض کنید توزیع‌های P و Q به ترتیب:

$$P = \{0.3, 0.7\} \quad \text{و} \quad Q = \{0.6, 0.4\}$$

ابتدا میانگین این دو توزیع $M = \{0.45, 0.55\}$ محاسبه می‌شود، سپس KL-Divergence بین P و M و بین Q و M برای محاسبه‌ی JSD به‌کار می‌رود.

انتخاب معیار مناسب: KL یا JSD؟

انتخاب بین KL-Divergence و JSD به نیاز مسئله بستگی دارد:

- KL-Divergence برای تخمین اختلاف بین یک توزیع اصلی و یک توزیع آزمایشی مناسب است، اما چون نامتقارن است، باید مراقب باشیم کدام توزیع را به عنوان اصلی و کدام را به عنوان آزمایشی انتخاب کنیم.
- JSD برای سنجش شباهت بین دو توزیع به‌کار می‌رود و به دلیل متقارن بودن، در مقایسه‌هایی که ترتیب توزیع‌ها مهم نیست، کاربرد بیشتری دارد.

کاربردهای عملی

- ارزیابی مدل‌های یادگیری ماشین: در یادگیری عمیق و مدل‌های احتمالاتی، تفاوت بین توزیع داده‌های ورودی و خروجی با KL-Divergence یا JSD ارزیابی می‌شود.
- تحلیل رفتار کاربران: در تحلیل‌های بازار و تحلیل رفتار کاربران، KL-Divergence به ارزیابی تفاوت رفتار کاربران از دوره‌ای به دوره‌ی دیگر کمک می‌کند.
- پردازش زبان طبیعی: در تشخیص موضوعات مختلف در متون، از JSD برای ارزیابی شباهت بین توزیع کلمات در دسته‌بندی‌ها استفاده می‌شود.

جمع‌بندی فصل هفتم

در این فصل با معیارهای KL-Divergence و Jensen-Shannon Divergence (JSD) برای مقایسه‌ی توزیع‌ها آشنا شدیم. این معیارها ابزارهای قدرتمندی برای اندازه‌گیری شباهت یا تفاوت بین توزیع‌های مختلف هستند و در کاربردهای مختلف آماری، یادگیری ماشین و تحلیل داده‌ها به‌کار می‌روند.

فصل هشتم: همبستگی (Correlation) و کاربردهای آن

مقدمه

همبستگی (Correlation) یکی از مفاهیم مهم در آمار و علم داده است که به ما امکان می‌دهد ارتباط بین دو متغیر کمی را بررسی کنیم. این شاخص نشان می‌دهد که تغییرات یک متغیر چگونه با تغییرات

متغیر دیگر مرتبط است. همبستگی به ما کمک می‌کند که رابطه‌های پنهان بین متغیرها را کشف کرده و از آن‌ها در مدل‌سازی‌های آماری و پیش‌بینی‌ها استفاده کنیم.

انواع همبستگی

دو نوع اصلی از همبستگی وجود دارد:

- همبستگی مثبت: زمانی که با افزایش یک متغیر، متغیر دیگر نیز افزایش یابد.
- همبستگی منفی: زمانی که با افزایش یک متغیر، متغیر دیگر کاهش یابد.

همچنین می‌توان همبستگی را به دسته‌های خطی و غیرخطی تقسیم‌بندی کرد.

ضریب همبستگی پیرسون (Pearson Correlation Coefficient)

ضریب همبستگی پیرسون یکی از رایج‌ترین معیارها برای اندازه‌گیری همبستگی خطی بین دو متغیر پیوسته است و به‌صورت زیر محاسبه می‌شود:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

در این فرمول:

- X و Y دو متغیر مورد بررسی هستند.
- $\bar{X}$ و $\bar{Y}$ میانگین‌های X و Y هستند.

ضریب r در بازه $[-1, 1]$ قرار می‌گیرد:

- اگر $r = 1$ باشد، رابطه‌ی خطی مثبت کامل بین دو متغیر وجود دارد.
- اگر $r = -1$ باشد، رابطه‌ی خطی منفی کامل بین دو متغیر وجود دارد.
- اگر $r = 0$ باشد، نشان‌دهنده‌ی عدم همبستگی خطی بین دو متغیر است.

کاربردها

- تحلیل بازار: برای تحلیل ارتباط بین عوامل مختلف مانند قیمت و تقاضا.
- تحلیل داده‌های پزشکی: مثلاً بررسی ارتباط بین فشار خون و وزن افراد.

مثال عددی

فرض کنید داده‌های زیر برای قد و وزن تعدادی فرد داریم:

قد (سانتی‌متر)	وزن (کیلوگرم)
160	55
170	65
180	75
190	85

برای محاسبه‌ی ضریب همبستگی پیرسون، ابتدا میانگین قد و وزن را محاسبه کرده و سپس مقدار r را به دست می‌آوریم.

ضریب همبستگی اسپیرمن (Spearman Correlation Coefficient)

ضریب همبستگی اسپیرمن، برای بررسی همبستگی رتبه‌ای استفاده می‌شود و زمانی مناسب است که رابطه بین متغیرها غیرخطی باشد. ضریب اسپیرمن بر اساس رتبه‌های متغیرها محاسبه می‌شود و فرمول آن به‌صورت زیر است:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

در اینجا:

- d_i اختلاف رتبه بین دو متغیر برای مشاهده‌ی iام است.
- n تعداد مشاهدات است.

این ضریب نیز مانند ضریب پیرسون در بازه $[-1, 1]$ قرار دارد.

کاربردها

- تحلیل داده‌های دسته‌بندی‌شده: برای داده‌هایی که از نظر کمی اندازه‌گیری نشده‌اند، اما می‌توان به آن‌ها رتبه داد.
- پیش‌بینی در بازاریابی: بررسی رتبه‌بندی‌ها و ارتباط بین داده‌ها در سطح رتبه‌ای، مانند امتیاز مشتریان به محصولات.

مثال عددی

فرض کنید امتیازات زیر از دو مجموعه داده داریم:

مجموعه A	مجموعه B
5	7
3	4
8	6
1	2

ابتدا داده‌ها را رتبه‌بندی کرده و سپس اختلاف رتبه‌ها و مقدار ضریب اسپیرمن را محاسبه می‌کنیم.

مقایسه‌ی همبستگی پیرسون و اسپیرمن

انتخاب بین ضریب پیرسون و اسپیرمن به نوع رابطه بین متغیرها بستگی دارد:

- پیرسون: برای داده‌های کمی و زمانی که رابطه خطی بین متغیرها وجود دارد، مناسب است.
- اسپیرمن: برای داده‌های غیرخطی یا رتبه‌ای مناسب‌تر است و در مواقعی که داده‌ها دارای توزیع نرمال نیستند یا دارای جهش هستند، عملکرد بهتری دارد.

کاربردهای همبستگی در علم داده

- تشخیص ویژگی‌های مهم: در یادگیری ماشین، از همبستگی برای شناسایی ویژگی‌های مرتبط با خروجی استفاده می‌شود. ویژگی‌هایی که همبستگی بالاتری با خروجی دارند، اغلب در مدل‌سازی انتخاب می‌شوند.
- تحلیل رابطه بین متغیرها: در تحلیل داده‌ها، همبستگی به ما کمک می‌کند تا روابط پنهان بین متغیرها را پیدا کنیم. مثلاً در تحلیل بازار، می‌توانیم ارتباط بین تبلیغات و فروش را بررسی کنیم.
- پیش‌بینی‌های مالی: در بازارهای مالی، از همبستگی برای بررسی ارتباط بین دارایی‌ها و مدیریت ریسک استفاده می‌شود. مثلاً سهام‌هایی که همبستگی منفی دارند، می‌توانند به تنوع‌بخشی به سبد سرمایه‌گذاری کمک کنند.

جمع‌بندی فصل هشتم

در این فصل با مفهوم همبستگی و کاربردهای آن آشنا شدیم. همبستگی به‌عنوان یک ابزار مهم در تحلیل داده‌ها و پیش‌بینی‌ها نقش کلیدی دارد و به ما کمک می‌کند تا روابط بین متغیرها را درک کرده و از آن‌ها برای مدل‌سازی استفاده کنیم.

فصل نهم: کار با نرم‌افزار JASP (آمار کاربردی بدون برنامه‌نویسی)

مقدمه

نرم‌افزار JASP یک ابزار منبع‌باز و رایگان برای تحلیل‌های آماری است که به دلیل رابط کاربری ساده و کارپسند خود، جایگاه ویژه‌ای در میان دانشجویان، پژوهشگران و افرادی که علاقه‌مند به تحلیل آماری بدون نیاز به برنامه‌نویسی هستند، یافته است. برخلاف نرم‌افزارهای آماری پیچیده مانند SPSS و R که نیاز به دانش برنامه‌نویسی و مهارت‌های پیشرفته دارند، JASP امکان انجام طیف گسترده‌ای از تحلیل‌های آماری را تنها با چند کلیک فراهم می‌کند. این نرم‌افزار نه تنها به راحتی تحلیل‌های توصیفی و آزمون‌های آماری مختلفی را اجرا می‌کند، بلکه امکانات بصری و نمودارهای متنوعی نیز برای درک بهتر داده‌ها در اختیار کاربران قرار می‌دهد.

۱. قابلیت‌های JASP

۱.۱. تحلیل‌های توصیفی

تحلیل توصیفی شامل محاسبه و نمایش شاخص‌های مرکزی (مانند میانگین و میانه) و پراکندگی (مانند واریانس و انحراف استاندارد) است که به کاربران کمک می‌کند تا یک دید کلی از داده‌ها داشته باشند و ساختار و ویژگی‌های آن‌ها را درک کنند. در JASP، می‌توان به سادگی این شاخص‌ها را انتخاب کرده و در قالب جدول و نمودار مشاهده کرد.

۱.۲. آزمون‌های فرضیه

JASP از آزمون‌های آماری مختلفی برای بررسی فرضیات آماری پشتیبانی می‌کند. برخی از این آزمون‌ها شامل موارد زیر هستند:

- آزمون T: برای مقایسه میانگین دو گروه مستقل یا وابسته استفاده می‌شود.
- آزمون Z: آزمونی مناسب برای مقایسه میانگین‌ها در نمونه‌های بزرگتر از ۳۰ نفر.

- آزمون ANOVA: برای مقایسه میانگین‌ها در بیش از دو گروه مورد استفاده قرار می‌گیرد.

این آزمون‌ها در JASP با تنظیماتی ساده قابل اجرا هستند و خروجی به صورت جدول و نمودار برای تفسیر سریع و دقیق ارائه می‌شود.

۱.۳. تحلیل همبستگی

تحلیل همبستگی به کاربران این امکان را می‌دهد تا ارتباط و شدت رابطه بین دو متغیر را بررسی کنند. JASP امکان محاسبه ضرایب همبستگی پیرسون و اسپیرمن را فراهم می‌کند که برای داده‌های نرمال و غیرنرمال به ترتیب مناسب هستند. خروجی این تحلیل شامل ضریب همبستگی r است که نشان‌دهنده شدت و نوع رابطه بین متغیرهاست.

۱.۴. تحلیل رگرسیون

تحلیل رگرسیون یکی از مهم‌ترین روش‌ها برای مدل‌سازی رابطه بین متغیرها و پیش‌بینی مقادیر است. JASP قابلیت اجرای رگرسیون خطی و غیرخطی را دارد که به پژوهشگران و تحلیل‌گران امکان می‌دهد تا اثر متغیرهای مستقل بر متغیر وابسته را بررسی کنند. خروجی این تحلیل شامل ضرایب رگرسیونی، مقدار p، و خطای استاندارد می‌باشد.

۱.۵. آزمون‌های ناپارامتریک

در JASP، آزمون‌های ناپارامتریک نیز برای تحلیل داده‌های رتبه‌ای یا غیرنرمال فراهم شده‌اند. برخی از این آزمون‌ها عبارتند از:

- آزمون U من-ویتنی: برای مقایسه دو گروه مستقل استفاده می‌شود.
- آزمون کای-دو: برای تحلیل ارتباط بین دو متغیر کیفی مورد استفاده قرار می‌گیرد.

۲. نصب و راه‌اندازی JASP

1. به سایت رسمی JASP به آدرس jasp-stats.org مراجعه کنید.
2. نسخه مناسب سیستم عامل خود را (ویندوز، مک، یا لینوکس) دانلود کنید.
3. پس از دانلود، نرم‌افزار را نصب کنید. نصب JASP ساده است و به سرعت شما را به محیط کاربرپسند نرم‌افزار هدایت می‌کند.
4. پس از باز کردن نرم‌افزار، محیطی را مشاهده خواهید کرد که به شما امکان می‌دهد فایل‌های داده خود را وارد کرده و تحلیل‌های آماری را آغاز کنید.

۳. تحلیل‌های توصیفی در JASP

بارگذاری داده‌ها

برای شروع تحلیل توصیفی در JASP، ابتدا داده‌های خود را در قالب فایل CSV یا Excel وارد نرم‌افزار کنید. این کار از طریق منوی اصلی و گزینه "Open" انجام می‌شود.

انتخاب تحلیل توصیفی

در منوی Descriptive Statistics می‌توانید شاخص‌های آماری توصیفی مانند میانگین، واریانس، میانه، چولگی و کشیدگی را انتخاب کنید. JASP به صورت خودکار جداولی حاوی این شاخص‌ها برای متغیرهای انتخابی ایجاد می‌کند.

تفسیر خروجی

نرم‌افزار به طور خودکار جدول‌ها و نمودارهایی برای شاخص‌های توصیفی ایجاد می‌کند. این خروجی‌ها کمک می‌کنند تا با دیدگاه کلی‌تری نسبت به داده‌ها، درک بهتری از توزیع و ساختار آن‌ها داشته باشید.

مثال عددی

فرض کنید مجموعه داده‌ای شامل قد و وزن افراد را در JASP بارگذاری کرده‌اید. با انتخاب تحلیل توصیفی، نرم‌افزار به شما میانگین، میانه و انحراف معیار هر متغیر را نمایش می‌دهد. این اطلاعات به شما کمک می‌کند تا با نمای کلی از داده‌ها آشنا شوید و تحلیل‌های بعدی را بر اساس این اطلاعات پایه‌گذاری کنید.

۴. انجام آزمون‌های فرضیه در JASP

انتخاب آزمون T

پس از وارد کردن داده‌ها، به منوی T-Tests بروید و نوع آزمون را انتخاب کنید:

- تک‌نمونه‌ای: مقایسه میانگین یک نمونه با یک مقدار مشخص.
- مستقل: مقایسه میانگین دو گروه مستقل.
- وابسته: مقایسه میانگین‌ها در گروه‌های جفت‌شده (مانند قبل و بعد از یک مداخله).

تنظیم فرضیات

برای اجرای آزمون، فرض صفر و فرض مقابل خود را مشخص کرده و سطح معناداری (مثلاً 0.05) را تعیین کنید.

تحلیل نتایج

JASP خروجی آزمون را به صورت جداول آماری ارائه می‌دهد و مقدار p-value را نمایش می‌دهد. بر اساس مقدار p، می‌توانید فرضیه صفر را رد یا قبول کنید.

مثال عددی

فرض کنید قصد دارید میانگین قد یک گروه از افراد را با یک مقدار مشخص مقایسه کنید. با استفاده از آزمون T تک‌نمونه‌ای در JASP، می‌توانید نتیجه این آزمون و مقدار p-value آن را دریافت کنید و تصمیم‌گیری کنید که آیا تفاوت معناداری بین میانگین قد گروه و مقدار مورد نظر وجود دارد یا خیر.

۵. تحلیل همبستگی در JASP

برای تحلیل همبستگی بین دو متغیر، به منوی Correlation بروید و متغیرهای مورد نظر را انتخاب کنید.

انتخاب نوع همبستگی

بسته به نوع داده‌ها و فرضیات شما، می‌توانید ضریب همبستگی پیرسون یا اسپیرمن را انتخاب کنید.

تفسیر نتایج

JASP جدول همبستگی و مقدار ϑ (r) را نمایش می‌دهد که شدت و نوع رابطه بین متغیرها را نشان می‌دهد. مقدار ϑ (r) بین -1 و 1 قرار دارد؛ مقادیر نزدیک به 1 یا -1 نشان‌دهنده همبستگی قوی و مقادیر نزدیک به 0 نشان‌دهنده عدم همبستگی می‌باشند.

۶. تحلیل رگرسیون در JASP

انتخاب تحلیل رگرسیون

برای تحلیل رگرسیون، پس از بارگذاری داده‌ها، به منوی Regression بروید.

انتخاب متغیر وابسته و مستقل

متغیر وابسته و متغیرهای مستقل خود را مشخص کنید. برای مثال، می‌توانید وزن را به عنوان متغیر وابسته و قد را به عنوان متغیر مستقل انتخاب کنید.

تفسیر خروجی

JASP خروجی رگرسیون شامل ضریب‌های رگرسیونی، مقدار p و خطای استاندارد را ارائه می‌دهد. این اطلاعات به شما کمک می‌کند تا رابطه بین متغیرها را مدل‌سازی کرده و به پیش‌بینی بپردازید.

مثال عددی

فرض کنید می‌خواهید وزن افراد را بر اساس قد آن‌ها پیش‌بینی کنید. با استفاده از تحلیل رگرسیون در JASP، می‌توانید مدل رگرسیونی و ضرایب مربوطه را به دست آورید و ارتباط بین این دو متغیر را بسنجید.

۷. رسم نمودارها در JASP

JASP انواع مختلفی از نمودارها را برای تجسم داده‌ها فراهم می‌کند که شامل نمودارهای پراکندگی، جعبه‌ای، هیستوگرام و غیره است. این نمودارها به تحلیل بهتر داده‌ها و تفسیر نتایج کمک می‌کنند و در درک الگوها و توزیع داده‌ها مفید هستند.

جمع‌بندی فصل نهم

در این فصل با نرم‌افزار JASP و قابلیت‌های آن آشنا شدیم. JASP ابزاری منبع‌باز و رایگان برای تحلیل‌های آماری است که استفاده از آن نیاز به برنامه‌نویسی ندارد و گزینه‌ای مناسب برای دانشجویان، پژوهشگران و تحلیل‌گران داده است. این نرم‌افزار با ارائه تحلیل‌های آماری گسترده، آزمون‌های فرضیه، ابزارهای همبستگی و رگرسیون، و همچنین رسم نمودارهای متنوع، به کاربران کمک می‌کند تا داده‌ها را به راحتی تحلیل و تفسیر کنند.

فصل دهم: تست الف-ب (A-B Test) آماری و کاربردهای آن

مقدمه

تست الف-ب (A-B Test) یکی از روش‌های مهم در آمار و علم داده است که برای مقایسه و ارزیابی دو گروه یا دو نسخه از یک محصول به کار می‌رود. این آزمون به‌ویژه در بازاریابی دیجیتال، طراحی وب‌سایت و اپلیکیشن‌ها و نیز بهبود تجربه کاربری بسیار رایج است.

مفهوم تست الف-ب

در تست الف-ب، دو نسخه از یک متغیر (مثلاً دو طراحی مختلف از یک وب‌سایت) را با هم مقایسه می‌کنیم تا بررسی کنیم کدام‌یک عملکرد بهتری دارد. برای انجام این آزمون، نمونه‌ای از کاربران به‌طور

تصادفی به دو گروه تقسیم می‌شوند: گروه A و گروه B. هر گروه به‌صورت مستقل در معرض یکی از نسخه‌ها قرار می‌گیرد و سپس نتایج ارزیابی و با استفاده از آزمون‌های آماری تحلیل می‌شود.

مراحل اجرای تست الف-ب

1. **تعریف فرضیات:** ابتدا باید فرضیه صفر (H_0) و فرضیه مقابل (H_1) را تعریف کنیم. به عنوان مثال، فرضیه صفر می‌تواند این باشد که "تفاوتی بین دو گروه A و B وجود ندارد"، و فرضیه مقابل نشان‌دهنده‌ی این است که "تفاوت معناداری بین دو گروه وجود دارد."
2. **تعیین معیارهای ارزیابی:** معیارهای مورد نظر برای سنجش موفقیت نسخه‌ها را مشخص می‌کنیم. این معیارها می‌تواند نرخ کلیک (CTR)، نرخ تبدیل (Conversion Rate) یا زمان سپری شده در صفحه باشد.
3. **تقسیم تصادفی نمونه:** نمونه‌ی کاربران به دو گروه A و B تقسیم می‌شود تا تعصبات احتمالی در نتایج کاهش یابد.
4. **جمع‌آوری داده‌ها:** هر گروه با نسخه‌ی مختص خود تعامل دارد و داده‌ها جمع‌آوری می‌شود.
5. **تحلیل آماری نتایج:** با استفاده از آزمون‌های آماری، مانند آزمون T، تفاوت میانگین‌ها یا نرخ تبدیل بین دو گروه ارزیابی می‌شود.
6. **تفسیر نتایج و تصمیم‌گیری:** در نهایت، نتایج به‌دست‌آمده تحلیل و تصمیم‌گیری می‌شود که کدام نسخه عملکرد بهتری دارد.

فرمول‌های آماری مرتبط با تست الف-ب

در اکثر موارد، برای تحلیل داده‌های تست الف-ب از آزمون T استفاده می‌شود. فرمول کلی آزمون T برای مقایسه‌ی میانگین‌های دو گروه به‌صورت زیر است:

$$t = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

در اینجا:

- $\overline{X}_A$ و $\overline{X}_B$: میانگین نتایج گروه‌های A و B هستند.
- s_A و s_B: انحراف معیارهای گروه‌های A و B هستند.
- n_A و n_B: تعداد نمونه‌های گروه‌های A و B هستند.

کاربردهای تست الف-ب

- **بهبود تجربه کاربری (UX):** با استفاده از تست الف-ب می‌توان طراحی‌های مختلفی از یک صفحه‌ی وب را با هم مقایسه کرد و تأثیر آن‌ها را بر تجربه‌ی کاربران سنجید.

- **بازاریابی دیجیتال:** در کمپین‌های تبلیغاتی، نسخه‌های مختلفی از یک تبلیغ یا خبرنامه را با هم مقایسه کرده تا میزان تأثیرگذاری آن‌ها را بر نرخ کلیک و نرخ تبدیل بررسی کنند.
- **بهینه‌سازی اپلیکیشن‌ها:** در طراحی اپلیکیشن‌های موبایل، تست الف-ب می‌تواند به بهبود ویژگی‌های مختلف اپلیکیشن کمک کند.

مثال عددی

فرض کنید یک وب‌سایت دو نسخه مختلف از صفحه‌ی ثبت‌نام خود را آزمایش می‌کند. نسخه A دارای طراحی ساده است و نسخه B دارای یک طرح جذاب‌تر و جدید است. هدف، مقایسه‌ی نرخ تبدیل (Conversion Rate) در این دو نسخه است.

جمع‌آوری داده‌ها:

- تعداد کاربران در نسخه A: 1000 نفر
- تعداد کاربران در نسخه B: 1000 نفر
- تعداد کاربرانی که در نسخه A ثبت‌نام کرده‌اند: 120 نفر
- تعداد کاربرانی که در نسخه B ثبت‌نام کرده‌اند: 150 نفر

محاسبه نرخ تبدیل:

نرخ تبدیل در نسخه A: $\frac{120}{1000} = 0.12$ یا ۱۲٪

نرخ تبدیل در نسخه B: $\frac{150}{1000} = 0.15$ یا ۱۵٪

اجرای آزمون T:

با استفاده از آزمون T برای مقایسه‌ی این دو نرخ تبدیل می‌توان بررسی کرد که آیا تفاوت معناداری بین نرخ تبدیل دو نسخه وجود دارد یا خیر.

تحلیل نتایج

با اجرای آزمون T و بررسی مقدار *p-value* می‌توان نتیجه‌گیری کرد که اگر *p-value* کمتر از سطح معناداری (معمولاً ۰.۰۵) باشد، فرض صفر رد شده و نتیجه می‌گیریم که تفاوت معناداری بین دو نسخه وجود دارد. در غیر این صورت، فرض صفر رد نمی‌شود و تفاوتی معنادار بین دو نسخه مشاهده نمی‌شود.

جمع‌بندی فصل دهم

در این فصل با تست الف-ب و کاربردهای آن آشنا شدیم. این آزمون ابزار قدرتمندی برای ارزیابی و بهبود طراحی‌ها و کمپین‌های بازاریابی است که به تصمیم‌گیری بهتر و علمی‌تر کمک می‌کند.

فصل یازدهم: تست فرضیه آماری (Hypothesis Test) - قسمت اول

مقدمه

تست فرضیه آماری یکی از ابزارهای اصلی در آمار است که به ما کمک می‌کند با استفاده از نمونه‌های آماری، نتایج و استنباط‌هایی درباره جمعیت اصلی به‌دست آوریم. هدف اصلی از این آزمون‌ها بررسی یک ادعا یا فرضیه درباره یک متغیر یا پارامتر جمعیت است. تست فرضیه در علوم مختلف، از جمله در علوم اجتماعی، بهداشت، بازاریابی و هوش مصنوعی کاربرد گسترده‌ای دارد.

مراحل تست فرضیه آماری

تست فرضیه شامل چندین مرحله است که باید به ترتیب و با دقت دنبال شوند:

1. **تعریف فرضیه صفر (H_0) و فرضیه مقابل (H_1):**
   - **فرضیه صفر (H_0):** فرضی است که معمولاً بیانگر عدم وجود تفاوت یا رابطه‌ای خاص در داده‌ها است و به‌عنوان فرضیه‌ای که باید آزمون شود، مطرح می‌شود.
   - **فرضیه مقابل (H_1):** فرضی است که خلاف فرضیه صفر را بیان می‌کند و هدف ما تعیین این است که آیا شواهد کافی برای پذیرش آن وجود دارد یا خیر.
2. **انتخاب سطح معناداری (α):** سطح معناداری یا احتمال خطای نوع اول (α) میزان تحمل ما برای پذیرش احتمال خطای رد کردن فرضیه صفر در صورت درست بودن آن را تعیین می‌کند. معمولاً از سطح‌های 0.05 (5%) یا 0.01 (1%) استفاده می‌شود.
3. **انتخاب آزمون آماری مناسب:** بسته به نوع داده‌ها و فرضیات، آزمون‌های مختلفی از جمله آزمون‌های T، آزمون‌های Z، آزمون کای-دو و آزمون‌های ناپارامتریک برای تست فرضیه استفاده می‌شوند.
4. **محاسبه آماره‌ی آزمون و مقدار p-value:** پس از انتخاب آزمون، آماره‌ی مناسب محاسبه می‌شود. سپس مقدار p-value که نشان‌دهنده احتمال وقوع آماره آزمون تحت فرضیه صفر است، به‌دست می‌آید.
5. **تفسیر نتایج و تصمیم‌گیری:** اگر مقدار p کمتر از سطح معناداری (α) باشد، فرضیه صفر رد می‌شود و نتیجه‌گیری می‌کنیم که شواهد کافی برای پذیرش فرضیه مقابل وجود دارد. در غیر این

صورت، فرضیه صفر رد نمی‌شود.

مثال ساده‌ای از تست فرضیه

فرض کنید یک فروشگاه آنلاین ادعا می‌کند که میانگین زمان تحویل سفارش‌ها کمتر از 24 ساعت است. برای آزمون این ادعا، نمونه‌ای از سفارش‌ها را انتخاب کرده و تست فرضیه را انجام می‌دهیم.

فرضیات:

- فرضیه صفر (H_0): میانگین زمان تحویل سفارش‌ها برابر یا بیشتر از 24 ساعت است.
- فرضیه مقابل (H_1): میانگین زمان تحویل سفارش‌ها کمتر از 24 ساعت است.

انتخاب سطح معناداری:

سطح معناداری $\alpha = 0.05$ انتخاب می‌شود.

انتخاب آزمون و محاسبه آماره‌ی آزمون:

فرض کنید میانگین زمان تحویل سفارش‌ها در نمونه 22 ساعت و انحراف معیار نمونه 4 ساعت باشد و تعداد سفارش‌ها 30 باشد.

آماره آزمون با استفاده از فرمول آزمون T تک‌طرفه محاسبه می‌شود:

$$t = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

در اینجا:

- $\overline{X} = 22$: میانگین نمونه
- $\mu = 24$: میانگین فرض‌شده
- $s = 4$: انحراف معیار نمونه
- $n = 30$: تعداد نمونه‌ها

تفسیر نتایج:

اگر مقدار p کمتر از 0.05 باشد، فرضیه صفر رد و نتیجه می‌گیریم که میانگین زمان تحویل سفارش‌ها کمتر از 24 ساعت است.

اهمیت سطح معناداری و احتمال خطای نوع اول و دوم

- **خطای نوع اول (Alpha Error):** رد کردن فرضیه صفر در حالی که درست است.
- **خطای نوع دوم (Beta Error):** پذیرش فرضیه صفر در حالی که نادرست است.

کاهش سطح معناداری به کاهش احتمال خطای نوع اول کمک می‌کند، اما معمولاً باعث افزایش احتمال خطای نوع دوم می‌شود. بنابراین، تعیین سطح معناداری مناسب بسته به هدف و حساسیت آزمون ضروری است.

جمع‌بندی فصل یازدهم

در این فصل با مفهوم تست فرضیه و مراحل اجرای آن آشنا شدیم. در بخش‌های بعدی، به آزمون‌های خاصی مانند آزمون‌های T و Z می‌پردازیم و مثال‌های عددی بیشتری را بررسی می‌کنیم.

فصل دوازدهم: تست فرضیه آماری (Hypothesis Test) - قسمت دوم، تست Z و تست T

مقدمه

آزمون‌های Z و T از مهم‌ترین روش‌های آماری برای تست فرضیه هستند. این آزمون‌ها به ما کمک می‌کنند تا بر اساس نمونه‌های آماری، در مورد جمعیت‌ها فرضیه‌هایی را بررسی کنیم. در این فصل، به بررسی شرایط استفاده از این آزمون‌ها، فرمول‌ها و مثال‌های عددی خواهیم پرداخت.

آزمون Z

شرایط استفاده از آزمون Z

آزمون Z معمولاً در شرایط زیر به کار می‌رود:

- **حجم نمونه بزرگ:** وقتی تعداد نمونه‌ها (n) بیشتر از 30 باشد، می‌توان از آزمون Z استفاده کرد.
- **توزیع نرمال:** اگر جمعیت دارای توزیع نرمال باشد یا وقتی حجم نمونه بزرگ باشد، می‌توان از آزمون Z بهره برد.
- **معرفی انحراف معیار جمعیت:** در آزمون Z، انحراف معیار جمعیت (σ) باید شناخته شده باشد.

فرمول آزمون Z

آماره آزمون Z با استفاده از فرمول زیر محاسبه می‌شود:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

در اینجا:

- $\overline{X}$: میانگین نمونه
- μ: میانگین جمعیت (فرضیه صفر)
- σ: انحراف معیار جمعیت
- n: تعداد نمونه

آزمون T

شرایط استفاده از آزمون T

آزمون T معمولاً در شرایط زیر به کار می‌رود:

- **حجم نمونه کوچک:** وقتی تعداد نمونه‌ها (n) کمتر از 30 باشد، معمولاً از آزمون T استفاده می‌شود.
- **توزیع نرمال:** اگر جمعیت دارای توزیع نرمال باشد، می‌توان از آزمون T بهره برد، حتی با حجم نمونه کوچک.
- **معرفی انحراف معیار جمعیت:** در آزمون T، انحراف معیار جمعیت (σ) معمولاً ناشناخته است و به جای آن از انحراف معیار نمونه (s) استفاده می‌شود.

فرمول آزمون T

آماره آزمون T با استفاده از فرمول زیر محاسبه می‌شود:

$$T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$$

در اینجا:

- s: انحراف معیار نمونه

مقایسه آزمون Z و T

ویژگی	آزمون Z	آزمون T

حجم نمونه	بزرگتر از 30	کوچکتر از 30
انحراف معیار	شناخته شده	ناشناخته
توزیع	نرمال	نرمال یا تقریباً نرمال
توزیع آماره	توزیع نرمال	توزیع T با $n-1$ درجه آزادی

مثال عددی برای آزمون Z

فرض کنید یک شرکت تولیدی ادعا می‌کند که میانگین زمان تولید یک محصول 100 ساعت است. برای بررسی این ادعا، نمونه‌ای از 36 محصول تولید شده (با میانگین 98 ساعت و انحراف معیار 10 ساعت) انتخاب می‌شود. آیا شواهد کافی برای رد فرضیه صفر وجود دارد؟

فرضیات:

- $H_0 : \mu = 100$
- $H_1 : \mu < 100$

محاسبه آماره آزمون Z:

برای محاسبه آماره Z، از فرمول زیر استفاده می‌کنیم:

$$Z = \frac{98-100}{\frac{10}{\sqrt{36}}} = \frac{-2}{\frac{10}{6}} = -1.2$$

محاسبه p-value:

با استفاده از جداول توزیع Z، مقدار p را برای $Z = -1.2$ محاسبه می‌کنیم.

تفسیر نتایج:

اگر مقدار p کمتر از 0.05 باشد، فرضیه صفر رد می‌شود.

مثال عددی برای آزمون T

فرض کنید یک مدرسه ادعا می‌کند که میانگین نمرات دانش‌آموزان در یک امتحان 75 است. نمونه‌ای از 10 دانش‌آموز (میانگین 78 و انحراف معیار 5) انتخاب شده است. آیا شواهد کافی برای رد فرضیه صفر وجود دارد؟

فرضیات:

- $H_0 : \mu = 75$
- $H_1 : \mu \neq 75$

محاسبه آماره آزمون T:

برای محاسبه آماره T، از فرمول زیر استفاده می‌کنیم:

$$T = \frac{78-75}{\frac{5}{\sqrt{10}}} = \frac{3}{1.58} \approx 1.90$$

محاسبه p-value:

با استفاده از جداول توزیع T و 9 درجه آزادی، مقدار p را محاسبه می‌کنیم.

تفسیر نتایج:

اگر مقدار p کمتر از 0.05 باشد، فرضیه صفر رد می‌شود.

جمع‌بندی فصل دوازدهم

در این فصل با آزمون‌های Z و T و شرایط و روش‌های استفاده از آن‌ها آشنا شدیم. در فصل بعدی، به آزمون‌های بیشتری در تست فرضیه می‌پردازیم و جزئیات بیشتری را بررسی خواهیم کرد.

فصل سیزدهم: تست فرضیه آماری (Hypothesis Test) - قسمت سوم، تست Z و تست T

مقدمه

آزمون‌های Z و T ابزارهای قدرتمندی برای آزمون فرضیات در آمار هستند. در این فصل، به بررسی شرایط و موارد خاص استفاده از این آزمون‌ها خواهیم پرداخت و مثال‌هایی برای درک بهتر مفهوم آن‌ها ارائه خواهیم کرد.

آزمون Z

تست Z برای میانگین جمعیت

وقتی اطلاعاتی درباره انحراف معیار جمعیت (σ) داریم، می‌توانیم از آزمون Z استفاده کنیم. به عنوان مثال، فرض کنید یک شرکت تولیدی می‌گوید که میانگین عمر مفید یک محصول 50 ساعت است. برای بررسی این ادعا، نمونه‌ای از 40 محصول (با میانگین 48 ساعت و انحراف معیار 5 ساعت) بررسی می‌شود.

فرضیات:

- $H_0 : \mu = 50$
- $H_1 : \mu < 50$

محاسبه آماره آزمون Z:

برای محاسبه آماره Z، از فرمول زیر استفاده می‌کنیم:

$$Z = \frac{48-50}{\frac{5}{\sqrt{40}}} = \frac{-2}{\frac{5}{6.32}} \approx -2.53$$

محاسبه p-value:

با استفاده از جداول توزیع Z، مقدار p برای $Z = -2.53$ محاسبه می‌شود.

تفسیر نتایج:

اگر مقدار p کمتر از 0.05 باشد، فرضیه صفر رد می‌شود.

آزمون T

تست T برای میانگین جمعیت

آزمون T زمانی استفاده می‌شود که اطلاعات دقیقی از انحراف معیار جمعیت نداریم و تنها می‌توانیم از انحراف معیار نمونه استفاده کنیم. فرض کنید یک محقق می‌خواهد بررسی کند که آیا یک دارو اثر مثبتی بر کاهش فشار خون دارد یا خیر. نمونه‌ای از 15 بیمار (با میانگین کاهش فشار 8 میلی‌متر جیوه و انحراف معیار 2 میلی‌متر جیوه) جمع‌آوری شده است.

فرضیات:

- $H_0 : \mu = 0$ (هیچ کاهش معنی‌داری در فشار خون وجود ندارد)
- $H_1 : \mu > 0$ (کاهش معنی‌دار در فشار خون وجود دارد)

محاسبه آماره آزمون T:

برای محاسبه آماره T، از فرمول زیر استفاده می‌کنیم:

$$T = \frac{8-0}{\frac{2}{\sqrt{15}}} = \frac{8}{3.87} \approx 2.07$$

محاسبه p-value:

با استفاده از جداول توزیع T و 14 درجه آزادی، مقدار p را محاسبه می‌کنیم.

تفسیر نتایج:

اگر مقدار p کمتر از 0.05 باشد، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که دارو اثر مثبتی بر کاهش فشار خون دارد.

نتیجه‌گیری در آزمون Z و T

آزمون Z: زمانی که حجم نمونه بزرگ باشد و انحراف معیار جمعیت شناخته شده باشد، از آزمون Z استفاده می‌شود. این آزمون بیشتر برای داده‌های نرمال و بزرگ مناسب است.

آزمون T: زمانی که حجم نمونه کوچک است و اطلاعاتی درباره انحراف معیار جمعیت نداریم، از آزمون T استفاده می‌شود. این آزمون به دلیل توزیع نرمالی که برای حجم‌های کوچک دارد، کاربرد دارد.

جمع‌بندی فصل سیزدهم

در این فصل، با جزئیات بیشتری به بررسی آزمون‌های Z و T پرداختیم و مثال‌های عددی را بررسی کردیم. در فصل بعدی، به آزمون‌های بیشتری خواهیم پرداخت و نحوه کاربرد آن‌ها را بررسی خواهیم کرد.

فصل چهاردهم: تست فرضیه آماری (Hypothesis Test) - قسمت چهارم، تست Z و تست T

مقدمه

در فصل‌های قبل، به معرفی آزمون‌های Z و T پرداختیم. در این فصل، بر روی جزئیات بیشتر و مثال‌های کاربردی متمرکز خواهیم شد. همچنین، به بررسی شرایط خاصی که ممکن است در هنگام استفاده از این آزمون‌ها به وجود آید، خواهیم پرداخت.

تست فرضیه با آزمون Z

مثال عملی

فرض کنید یک کارخانه‌ی تولیدی ادعا می‌کند که میانگین وزن محصول تولیدی آن 100 گرم است. برای بررسی این ادعا، یک نمونه‌ی تصادفی از 50 محصول انتخاب شده است که میانگین وزن آن‌ها 98 گرم و انحراف معیار جمعیت 4 گرم است. آیا شواهد کافی برای رد ادعای کارخانه وجود دارد؟

فرضیات:

- $H_0 : \mu = 100$
- $H_1 : \mu < 100$

محاسبه آماره آزمون Z:

برای محاسبه آماره Z، از فرمول زیر استفاده می‌کنیم:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{98 - 100}{\frac{4}{\sqrt{50}}} = \frac{-2}{0.5657} \approx -3.54$$

محاسبه p-value:

با توجه به جداول توزیع Z، مقدار p برای $Z = -3.54$ محاسبه می‌شود.

تفسیر نتایج:

اگر $p < 0.05$، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که میانگین وزن محصولات کمتر از 100 گرم است.

تست فرضیه با آزمون T

مثال عملی

فرض کنید یک گروه از محققان می‌خواهند بررسی کنند که آیا میانگین دما در یک منطقه خاص در تابستان بیشتر از 30 درجه سانتی‌گراد است یا خیر. آن‌ها نمونه‌ای از 12 روز تابستانی را جمع‌آوری کرده‌اند

که میانگین دما 32 درجه و انحراف معیار 3 درجه است.

فرضیات:

- $H_0 : \mu = 30$
- $H_1 : \mu > 30$

محاسبه آماره آزمون T:

برای محاسبه آماره T، از فرمول زیر استفاده می‌کنیم:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{32 - 30}{\frac{3}{\sqrt{12}}} = \frac{2}{0.866} \approx 2.31$$

محاسبه p-value:

با استفاده از جداول توزیع T و 11 درجه آزادی، مقدار p برای $T = 2.31$ محاسبه می‌شود.

تفسیر نتایج:

اگر $p < 0.05$، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که میانگین دما در این منطقه بیشتر از 30 درجه است.

نکات مهم در استفاده از آزمون‌های Z و T

- **بررسی توزیع داده‌ها:** قبل از انتخاب آزمون، بررسی توزیع داده‌ها بسیار مهم است. برای داده‌های غیر نرمال، ممکن است نیاز به آزمون‌های ناپارامتریک باشد.
- **حجم نمونه:** در آزمون Z، حجم نمونه بزرگتر از 30 و در آزمون T معمولاً کمتر از 30 است.
- **انحراف معیار:** در آزمون T، انحراف معیار جمعیت معمولاً ناشناخته است و از انحراف معیار نمونه استفاده می‌شود.
- **آزمون‌های دوطرفه و یکطرفه:** توجه به نوع آزمون (دوطرفه یا یکطرفه) اهمیت دارد. در آزمون‌های یکطرفه، فقط یک سمت توزیع بررسی می‌شود، در حالی که در آزمون‌های دوطرفه، هر دو سمت توزیع مورد بررسی قرار می‌گیرد.

جمع‌بندی فصل چهاردهم

در این فصل، با جزئیات بیشتری به بررسی آزمون‌های Z و T پرداختیم و مثال‌های عملی بیشتری را بررسی کردیم. در فصل بعدی، به آزمون‌های بیشتری خواهیم پرداخت و چگونگی کاربرد آن‌ها را بررسی خواهیم کرد.

فصل پانزدهم: تست فرضیه آماری (Hypothesis Test) - قسمت پنجم، تست Z و تست T

مقدمه

در این فصل، به تفصیل بیشتری در مورد آزمون‌های Z و T خواهیم پرداخت. علاوه بر این، شرایط خاصی که در هنگام استفاده از این آزمون‌ها باید در نظر گرفته شود را بررسی خواهیم کرد و مثال‌های عددی بیشتری ارائه خواهیم داد.

تست فرضیه با آزمون Z

مثال عملی

فرض کنید یک رستوران ادعا می‌کند که میانگین زمان انتظار مشتریان برای سرو غذا 20 دقیقه است. یک محقق 36 مشتری را انتخاب کرده و میانگین زمان انتظار آن‌ها 22 دقیقه و انحراف معیار جمعیت 5 دقیقه است. آیا شواهد کافی برای رد ادعای رستوران وجود دارد؟

فرضیات:

- $H_0 : \mu = 20$
- $H_1 : \mu > 20$

محاسبه آماره آزمون Z:

برای محاسبه آماره Z، از فرمول زیر استفاده می‌کنیم:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{22 - 20}{\frac{5}{\sqrt{36}}} = \frac{2}{0.8333} \approx 2.4$$

محاسبه p-value:

با توجه به جداول توزیع Z، مقدار p برای $Z = 2.4$ محاسبه می‌شود.

تفسیر نتایج:

اگر $p < 0.05$، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که زمان انتظار مشتریان برای سرو غذا بیشتر از 20 دقیقه است.

تست فرضیه با آزمون T

مثال عملی

فرض کنید یک گروه از محققان می‌خواهند بررسی کنند که آیا یک برنامه آموزشی جدید موجب بهبود نمرات دانش‌آموزان می‌شود یا خیر. 10 دانش‌آموز به صورت تصادفی انتخاب شده‌اند و میانگین نمرات آن‌ها قبل از آموزش 75 و بعد از آموزش 82 بوده است. انحراف معیار نمرات بعد از آموزش 4 است. آیا شواهد کافی برای رد فرضیه وجود دارد؟

فرضیات:

- $H_0 : \mu = 75$
- $H_1 : \mu > 75$

محاسبه آماره آزمون T:

برای محاسبه آماره T، از فرمول زیر استفاده می‌کنیم:

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{82 - 75}{\frac{4}{\sqrt{10}}} = \frac{7}{1.2649} \approx 5.53$$

محاسبه p-value:

با استفاده از جداول توزیع T و 9 درجه آزادی، مقدار p برای $T = 5.53$ محاسبه می‌شود.

تفسیر نتایج:

اگر $p < 0.05$، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که برنامه آموزشی موجب بهبود نمرات دانش‌آموزان شده است.

نکات مهم در استفاده از آزمون‌های Z و T

- **تعیین نوع آزمون:** بسته به نوع فرضیه (یکطرفه یا دوطرفه) و نوع توزیع داده‌ها (نرمال یا غیر نرمال)، نوع آزمون باید مشخص شود.
- **توزیع داده‌ها:** اگر داده‌ها دارای توزیع نرمال نیستند و حجم نمونه کوچک است، باید از آزمون‌های ناپارامتریک استفاده کرد.

- **تحلیل اثر اندازه:** در کنار آزمون فرضیه، تحلیل اثر اندازه می‌تواند به ما کمک کند تا درک بهتری از اهمیت نتیجه داشته باشیم.
- **استفاده از نرم‌افزارها:** استفاده از نرم‌افزارهای آماری مانند JASP یا SPSS می‌تواند فرایند محاسبه آماره‌های آزمون و p-value را تسهیل کند.

جمع‌بندی فصل پانزدهم

در این فصل، به بررسی عمیق‌تری از آزمون‌های Z و T پرداختیم و مثال‌های عملی بیشتری را مورد بررسی قرار دادیم. در فصل بعدی، به آزمون‌های بیشتری خواهیم پرداخت و نحوه کاربرد آن‌ها را بررسی خواهیم کرد.

فصل شانزدهم: تست فرضیه آماری (Hypothesis Test) - قسمت ششم، تست Z و تست T و ANOVA

مقدمه

آزمون ANOVA (تحلیل واریانس) ابزاری قوی برای مقایسه میانگین‌های چند گروه است. این آزمون به ما امکان می‌دهد تا بررسی کنیم که آیا تفاوت معنی‌داری بین میانگین‌های گروه‌های مختلف وجود دارد یا خیر. در این فصل، نحوه انجام آزمون ANOVA و مثال‌های عملی را بررسی خواهیم کرد.

تحلیل واریانس (ANOVA)

آزمون یک‌طرفه ANOVA

آزمون یک‌طرفه ANOVA زمانی استفاده می‌شود که بخواهیم میانگین‌های سه یا چند گروه را با هم مقایسه کنیم. به عنوان مثال، فرض کنید یک محقق می‌خواهد بررسی کند که آیا سه نوع مختلف کود تأثیر متفاوتی بر رشد گیاهان دارند یا خیر. محقق سه گروه از گیاهان را با سه نوع کود مختلف پرورش می‌دهد و ارتفاع گیاهان را اندازه‌گیری می‌کند.

فرضیات:

- $H_0 : \mu_1 = \mu_2 = \mu_3$ (هیچ تفاوت معنی‌داری بین میانگین‌ها وجود ندارد)
- H_1 : حداقل یکی از میانگین‌ها متفاوت است.

جمع‌آوری داده‌ها:

فرض کنید ارتفاع گیاهان در سه گروه به صورت زیر باشد:

- گروه 1: 20، 22، 19
- گروه 2: 25، 23، 26
- گروه 3: 30، 31، 29

محاسبه میانگین و واریانس:

میانگین گروه 1: $\bar{X}_1 = \frac{20+22+19}{3} = 20.33$

میانگین گروه 2: $\bar{X}_2 = \frac{25+23+26}{3} = 24.67$

میانگین گروه 3: $\bar{X}_3 = \frac{30+31+29}{3} = 30$

واریانس گروه 1: $s_1^2 = \frac{(20-20.33)^2+(22-20.33)^2+(19-20.33)^2}{3-1} = 3.33$

واریانس گروه 2: $s_2^2 = \frac{(25-24.67)^2+(23-24.67)^2+(26-24.67)^2}{3-1} = 1.33$

واریانس گروه 3: $s_3^2 = \frac{(30-30)^2+(31-30)^2+(29-30)^2}{3-1} = 0.67$

محاسبه ANOVA:

محاسبه میانگین کل: $\bar{X} = \frac{\bar{X}_1+\bar{X}_2+\bar{X}_3}{3} = \frac{20.33+24.67+30}{3} = 25.67$

محاسبه مجموع مربعات بین گروه‌ها (SSB):

$$SS_B = n((\bar{X}_1 - \bar{X})^2 + (\bar{X}_2 - \bar{X})^2 + (\bar{X}_3 - \bar{X})^2)$$

$$SS_B = 3((20.33 - 25.67)^2 + (24.67 - 25.67)^2 + (30 - 25.67)^2) = 51.33$$

محاسبه مجموع مربعات درون گروه‌ها (SSW):

$$SS_W = (n-1)(s_1^2 + s_2^2 + s_3^2)$$

$$SS_W = 2(3.33 + 1.33 + 0.67) = 10$$

محاسبه آماره F:

با توجه به درجات آزادی $df_B = k - 1 = 3 - 1 = 2$ و $df_W = N - k = 9 - 3 = 6$، آماره F به صورت زیر محاسبه می‌شود:

$$F = \frac{SS_B/df_B}{SS_W/df_W} = \frac{51.33/2}{10/6} = 15.4$$

محاسبه p-value:

با توجه به جداول توزیع F و $df_B = 2$ و $df_W = 6$، مقدار p محاسبه می‌شود.

تفسیر نتایج:

اگر $p < 0.05$، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که حداقل یکی از میانگین‌ها متفاوت است.

جمع‌بندی فصل شانزدهم

در این فصل، با آزمون ANOVA و نحوه استفاده از آن برای مقایسه میانگین‌های چند گروه آشنا شدیم. این آزمون به ما کمک می‌کند تا بررسی کنیم که آیا تفاوت معنی‌داری بین میانگین‌های گروه‌های مختلف وجود دارد یا خیر. در فصل بعدی، به آزمون‌های بیشتری خواهیم پرداخت و نحوه کاربرد آن‌ها را بررسی خواهیم کرد.

فصل هفدهم: تست فرضیه آماری (Hypothesis Test) - قسمت هفتم، تست Z و تست T و ANOVA

مقدمه

در این فصل، بر روی جزئیات بیشتر آزمون ANOVA و همچنین آزمون‌های چندگانه متمرکز خواهیم شد. این آزمون‌ها به ما این امکان را می‌دهند که اگر اختلاف معناداری بین گروه‌ها وجود داشت، بتوانیم مشخص کنیم که کدام گروه‌ها با هم متفاوت‌اند.

آزمون ANOVA دوطرفه

آزمون ANOVA دوطرفه به ما این امکان را می‌دهد که تأثیر دو عامل مختلف را بر یک متغیر وابسته بررسی کنیم. به عنوان مثال، فرض کنید یک محقق می‌خواهد بررسی کند که آیا نوع کود و نوع خاک تأثیر

معناداری بر رشد گیاهان دارد.

فرضیات:

- فرضیه صفر: $H_0 : \mu_A = \mu_B = \mu_C$ (تأثیر نوع کود و نوع خاک بر رشد گیاهان وجود ندارد)
- فرضیه مقابل: H_1: حداقل یکی از میانگین‌ها متفاوت است.

جمع‌آوری داده‌ها: فرض کنید داده‌های زیر برای رشد گیاهان در دو نوع کود (کود A و کود B) و دو نوع خاک (خاک X و خاک Y) جمع‌آوری شده است:

کود	خاک	رشد گیاه (سانتی‌متر)
A	X	20
A	Y	22
B	X	25
B	Y	27

محاسبه میانگین‌ها:

- میانگین رشد گیاه در کود A و خاک X: $\overline{X}_{AX} = 20$
- میانگین رشد گیاه در کود A و خاک Y: $\overline{X}_{AY} = 22$
- میانگین رشد گیاه در کود B و خاک X: $\overline{X}_{BX} = 25$
- میانگین رشد گیاه در کود B و خاک Y: $\overline{X}_{BY} = 27$

محاسبه میانگین کل:

$$\overline{X}_{Total} = \frac{20+22+25+27}{4} = 23.5$$

محاسبه مجموع مربعات بین گروه‌ها (SSB):

$$\overline{X}_{AX} - \overline{X}_{Total})^2 + n_A(\overline{X}_{AY} - \overline{X}_{Total})^2 + n_B(\overline{X}_{BX} - \overline{X}_{Total})^2 + n_B(\overline{X}_{BY} - \overline{X}_{Total})^2$$

با $n_A = n_B = 2$ (دو مشاهدات در هر گروه):

$$SSB = 2(20 - 23.5)^2 + 2(22 - 23.5)^2 + 2(25 - 23.5)^2 + 2(27 - 23.5)^2$$

$$SSB = 2(12.25) + 2(2.25) + 2(2.25) + 2(12.25) = 60$$

محاسبه مجموع مربعات درون گروه‌ها (SSW):

SSW برای هر گروه به این صورت محاسبه می‌شود:

$$SSW = (n-1)(s_1^2 + s_2^2 + s_3^2 + s_4^2)$$

فرض کنید واریانس‌ها به صورت زیر باشد:

- $s_1^2 = 0, s_2^2 = 0, s_3^2 = 0, s_4^2 = 0$

در این صورت:

$$SSW = (2-1)(0+0+0+0) = 0$$

محاسبه آماره F:

$$F = \frac{SSB/df_B}{SSW/df_W}$$

با $df_B = k - 1 = 4 - 1 = 3$ و $df_W = N - k = 8 - 4 = 4$:

$$F = \frac{60/3}{0/4}$$

(توجه کنید که در اینجا به دلیل وجود سطوح با واریانس صفر نمی‌توانیم F را محاسبه کنیم).

تفسیر نتایج:

اگر F محاسبه‌شده از F جدول بزرگ‌تر باشد، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که حداقل یکی از میانگین‌ها متفاوت است.

آزمون‌های چندگانه

هنگامی که نتایج ANOVA نشان می‌دهد که تفاوت معناداری وجود دارد، می‌توانیم از آزمون‌های چندگانه مانند آزمون Tukey یا Bonferroni برای شناسایی اینکه کدام گروه‌ها با هم متفاوت هستند، استفاده کنیم.

- **آزمون Tukey:** این آزمون برای مقایسه همه گروه‌ها به‌صورت زوجی استفاده می‌شود و تفاوت‌های معنادار بین میانگین‌ها را شناسایی می‌کند.
- **آزمون Bonferroni:** این آزمون برای کنترل خطای نوع اول در مقایسه‌های چندگانه استفاده می‌شود و سطح معناداری را تقسیم بر تعداد مقایسه‌ها می‌کند.

جمع‌بندی فصل هفدهم

در این فصل، به بررسی عمیق‌تری از آزمون ANOVA دوطرفه و آزمون‌های چندگانه پرداختیم. این ابزارها به ما کمک می‌کنند تا بتوانیم تفاوت‌های معنادار بین گروه‌های مختلف را شناسایی کنیم و تحلیل‌های بهتری ارائه دهیم. در فصل بعدی، به آزمون‌های بیشتری خواهیم پرداخت و نحوه کاربرد آن‌ها را بررسی خواهیم کرد.

فصل هجدهم: تست فرضیه آماری (Hypothesis Test) - قسمت هشتم، تست U

مقدمه

آزمون U من-ویتنی (Mann-Whitney U Test) یک آزمون غیرپارامتری است که برای مقایسه دو گروه مستقل و تعیین اینکه آیا یکی از گروه‌ها به طور معناداری بزرگ‌تر یا کوچک‌تر از دیگری است، استفاده می‌شود. این آزمون معمولاً زمانی به کار می‌رود که شرایط لازم برای آزمون t مستقل (نرمال بودن داده‌ها و واریانس‌های برابر) رعایت نشده باشد.

فرضیات

- فرضیه صفر (H_0): دو گروه دارای توزیع یکسان هستند.
- فرضیه مقابل (H_1): توزیع یکی از گروه‌ها بزرگ‌تر از دیگری است.

جمع‌آوری داده‌ها

فرض کنید دو گروه از داده‌ها به صورت زیر داریم:

- گروه A: 12، 15، 14، 10، 13
- گروه B: 20، 22، 21، 19، 25

مراحل آزمون

مرحله 1: رتبه‌بندی داده‌ها

تمام داده‌ها را در یک مجموعه قرار داده و آن‌ها را رتبه‌بندی می‌کنیم.

داده	گروه	رتبه
10	A	1
12	A	2
13	A	3
14	A	4
15	A	5
19	B	6
20	B	7
21	B	8
22	B	9
25	B	10

مرحله 2: محاسبه مجموع رتبه‌ها

مجموع رتبه‌های هر گروه را محاسبه می‌کنیم:

مجموع رتبه‌های گروه A:

$$R_A = 1 + 2 + 3 + 4 + 5 = 15$$

مجموع رتبه‌های گروه B:

$$R_B = 6 + 7 + 8 + 9 + 10 = 40$$

مرحله 3: محاسبه U

فرمول محاسبه U به صورت زیر است:

$$U_A = n_A n_B + \frac{n_A(n_A + 1)}{2} - R_A$$

$$U_B = n_A n_B + \frac{n_B(n_B+1)}{2} - R_B$$

که در آن:

- n_A و n_B تعداد مشاهدات در گروه A و B هستند.

با توجه به داده‌ها:

$$n_A = 5$$
$$n_B = 5$$

محاسبه U برای گروه A:

$$U_A = 5 \cdot 5 + \frac{5 \cdot (5+1)}{2} - 15$$
$$U_A = 25 + 15 - 15 = 25$$

محاسبه U برای گروه B:

$$U_B = 5 \cdot 5 + \frac{5 \cdot (5+1)}{2} - 40$$
$$U_B = 25 + 15 - 40 = 0$$

مرحله 4: محاسبه مقدار U نهایی

مقدار U نهایی برابر با کمینه U بین دو گروه است:

$$U = \min(U_A, U_B) = \min(25, 0) = 0$$

مرحله 5: تعیین سطح معناداری

برای تعیین اینکه آیا U معنادار است یا خیر، می‌توانیم از جداول توزیع U من-ویتنی استفاده کنیم یا از نرم‌افزارهای آماری استفاده کنیم.

مرحله 6: تفسیر نتایج

اگر مقدار U محاسبه شده از مقدار U بحرانی (طبق سطح معناداری 0.05) کوچک‌تر باشد، فرضیه صفر رد می‌شود و نتیجه می‌گیریم که بین دو گروه تفاوت معناداری وجود دارد.

جمع‌بندی فصل هجدهم

در این فصل، با آزمون U من-ویتنی آشنا شدیم و یاد گرفتیم که چگونه این آزمون را برای مقایسه دو گروه مستقل استفاده کنیم. این آزمون به ما کمک می‌کند تا در شرایطی که داده‌ها نرمال نیستند، مقایسه‌های معناداری انجام دهیم.