Statistics

مقدمه :

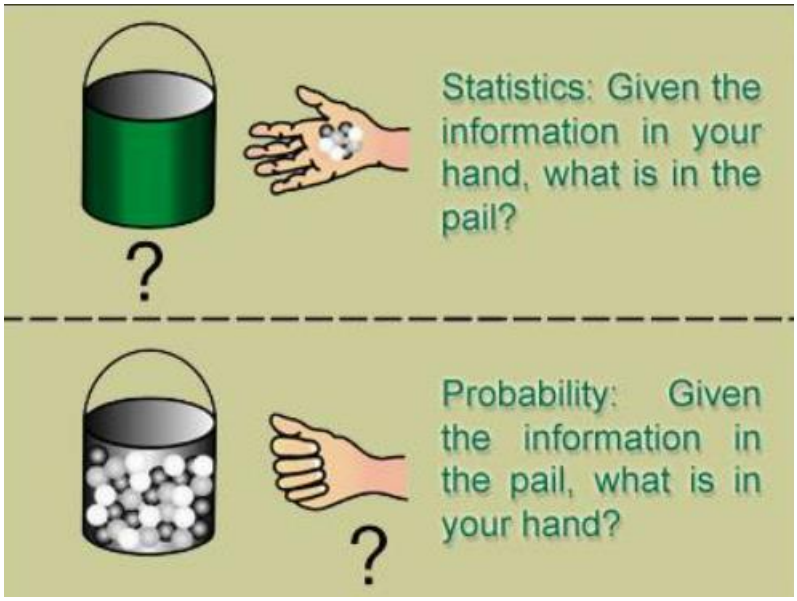
اکثر مسائلی که تا اینجای کار بررسی کردیم به این شکل بودند که ما ویژگی‌های یک جامعه بزرگ (به عنوان مثال توزیع احتمال یک متغیر تصادفی در آن جامعه) را می‌دانستیم و می‌خواستیم بر اساس آن یک رویداد را پیش‌بینی کنیم. اما اگر بخواهیم ویژگی‌های جامعه را به دست آوریم باید چه کرد؟ برای مثال اگر بخواهیم پیش‌بینی کنیم در انتخابات چه نامزدی رأی خواهد آورد باید نظر تمام افراد جامعه را پرسید تا دانست اکثریت جامعه به کدام نامزد رأی خواهد داد اما اینکار بسیار هزینه بر است. آیا راه حل دیگری وجود دارد؟

استدلال استنتاجی در مقابل استدلال استقرایی :

در منطق دو شیوه برای استدلال وجود دارد.

۱- شیوه اول که به آن استدلال استنتاجی (Deductive Reasoning) می‌گویند یک شیوه رسیدن از کل به جزء است که در آن با استفاده از حقایق و کلیاتی که از قبل می‌دانیم به نتایج جدید می‌رسیم.

۲- شیوه دوم استدلال استقرایی (Inductive Reasoning) است که یک شیوه رسیدن از جزء به کل است و با مشاهده نتایج جزئی تلاش می‌کنیم آن‌ها را به کل تعمیم دهیم.



حال که با دو شیوه استدلال آشنا شدیم می‌توان گفت که علم احتمال نمونه‌ای از استدلال استنتاجی و آمار نمونه‌ای از استدلال استقرایی است. در مبحث احتمال دیدیم که می‌توان ویژگی‌های یک نمونه کوچک را از روی ویژگی‌های جامعه تعیین کرد. علم آمار به ما می‌گوید عکس این هم امکان‌پذیر است. یعنی می‌توان از روی ویژگی‌های یک نمونه کوچک از جامعه به ویژگی‌های کل جامعه پی‌برد. برای مثال اگر بخواهیم نتیجه انتخابات را پیش‌بینی کنیم می‌توانیم صرفاً از یک نمونه ۱۰۰ یا ۱۰۰۰ نفری استفاده کنیم و اکثریت آن را به جامعه نیز تعمیم دهیم. البته این کار به شرطی امکان پذیر است که نمونه ما به اندازه کافی تصادفی باشد.

در نهایت می‌توان گفت که آمار و احتمال مکمل همدیگر هستند. در یک شیوه علمی ما ابتدا با جمع‌آوری نمونه‌ها برای کلیت جامعه مدل‌های آماری می‌سازیم و سپس با کمک احتمال و از روی مدل‌های ساخته شده به سؤال‌های خود درباره جامعه پاسخ می‌دهیم.

❖ کاربردهای آمار :

برای انجام تحلیل‌های آماری در مرحله اول به طراحی آزمایش و جمع‌آوری داده نیاز داریم. اما سؤال این است که پس از جمع‌آوری این اطلاعات چه کارهایی می‌توان با آن‌ها انجام داد؟

۱- فرض کنید طبق معمول یک سکه داریم که این بار احتمال شیر آمدنش را نداریم. برای اینکه بتوانیم این احتمال را حدس بزنیم سکه را چندین بار پرتاب می‌کنیم و تعداد شیرها را ثبت می‌کنیم. پس از انجام نمونه‌گیری به این شیوه ، باید یک روش مناسب برای حدس زدن مجهولمان داشته باشیم که به آن تخمین می‌گویند. تخمین می‌تواند به صورت نقطه‌ای یا بازه‌ای باشد و در آن مجهول ما یک متغیر تصادفی فرض شود یا صرفاً یک پارامتر نامعلوم. در ادامه درس با انواع این تخمین‌ها آشنا خواهیم شد.

۲- گاهی اوقات پس از جمع‌آوری داده‌ها می‌خواهیم درستی یا نادرستی یک فرض درباره جامعه را بررسی کنیم. مثلاً پس از جمع‌آوری وزن افراد یک نمونه می‌خواهیم بررسی کنیم آیا وزن افراد جامعه از توزیع نرمال پیروی می‌کند یا خیر؟ برای این کار نیاز به ابزاری داریم تا به ما در تصمیم‌گیری اینکه فرضمان درست است یا خیر کمک کند. نام این ابزار آزمون فرض است و در ادامه با انواع آن و آزمون‌های مختلف آشنا می‌شویم.

۳- بخشی از آمار به ما کمک می‌کند تا بتوانیم به کمک آزمایش های متعدد روابط بین متغیرها را کشف کنیم که به این بخش از آمار تحلیل رگرسیون می‌گویند.

این موارد چند نمونه محدود از تحلیل‌های متنوعی بودند که می‌توانیم روی داده‌های یک آزمایش انجام دهیم و نتایج جالبی از آن بدست بیاوریم.

آماره (Statistic) :

بعد از نمونه گیری باید اطلاعاتی درباره نمونه ها استخراج کرد. به هر تابعی که روی نمونه تصادفی اعمال شود آماره می گویند.

دو تا از پر کاربردترین آماره ها میانگین و واریانس هستند.

آزمون فرض (Hypothesis Test) یا Z-Test :

❖ مقدمه :

ما در زندگی با فرضیات مختلفی روبه‌رو می‌شویم و می‌خواهیم درست یا غلط بودن این فرضیات را بررسی کنیم. اغلب اوقات این فرضیه ها بیانگر وجود یا عدم وجود رابطه بین دو متغیراند.

به‌طور مثال ممکن است شما در زندگی روزمره خود با این جمله که سرمایه‌گذاری در بازار ارز پرسودتر از سرمایه‌گذاری در بازار سهام است رو به رو شده باشید و به این فکر بیفتید که این فرضیه را بررسی کنید و وجود تفاوت بین سرمایه‌گذاری در این دو بازار را نشان دهید.

یا در مثالی دیگر ممکن است بخواهید نشان دهید دانش‌آموزان ساکن پایتخت درصد قبولی بالاتری در کنکور دارند. در این جا شما می خواهید وجود رابطه بین محل سکونت و قبولی کنکور را نشان دهید یا به عبارت دیگر ادعای عدم وجود رابطه میان این دو متغیر را رد کنید.

یا فرض کنید که یک شرکت دارویی دارید و به تازگی داروی جدیدی تولید کرده‌اید. می‌خواهید ببینید که آیا این دارو برای بیماری سرطان موثر است یا خیر. عدم تاثیر دارو فرضی است که تغییر چندانی در شرایط فعلی ایجاد نمی کند و آن چه شما به دنبال آن هستید رد کردن این فرض و نشان دادن تاثیرگذار بودن این دارو است.

در تمام این مثال ها فرضی وجود دارد که اثر یا تفاوت بین متغیرها را انکار می کند به طور مثال یکسان بودن سرمایه‌گذاری در بازار ارز و سهام ، عدم تاثیر محل سکونت در قبولی کنکور و بی تاثیر بودن دارو بر سرطان. چنین فرض‌هایی فرض پوچ یا فرض صفر (H_0) نام دارند و ما در بررسی خود معمولا به دنبال رد این فرضیه و نشان دادن یک فرض جایگزین (H_1) هستیم ، که معمولا درستی آن برای ما سودمندتر است. در مثال‌های بالا پرسود تر بودن سرمایه‌گذاری در بازار ارز نسبت به

بازار سهام ، تاثیر محل سکونت در قبولی کنکور و موثر بودن دارو در درمان سرطان ، فرض‌های جایگزین هستند که با رد کردن فرض صفر می توانیم به آن‌ها برسیم.
پس از مشخص کردن هرچه دقیق‌تر فرض صفر و فرض جایگزین در بررسی خود ، لازم است که با استفاده از مشاهدات انجام شده به آزمون فرض بپردازیم.

❖ **فرض صفر (H_0 or Null Hypothesis) :**

فرضی است که در مورد پذیرش یا عدم پذیرش آن بحث می کنیم. اغلب به دنبال رد این فرض هستیم.

❖ **فرض جایگزین (H_1 or Alternative Hypothesis) :**

فرضی است که در صورت رد کردن فرض صفر آن را می پذیریم.

نکته ! هدف از آزمون آماری این است که آیا شواهد موجود برای رد فرض صفر کافی است یا خیر.

نکته ! آماره در آزمون های آماری متغیری است که از داده های مشاهده شده بدست می آوریم. با استفاده از آن طی یک آزمون آماری می توانیم شواهد کافی برای پذیرش یا عدم پذیرش فرض صفر را بدست بیاوریم.

❖ **ناحیه رد (Rejection Region) :**

ناحیه ای است که اگر آماره در آن قرار گیرد ، فرض صفر را رد می کنیم.

❖ **ناحیه قبول (Acceptance Region) :**

ناحیه ای است که اگر آماره در آن قرار گیرد ، نمی توان فرض صفر را رد کرد.

❖ **مقدار بحرانی (Critical Point) :**

نقطه ای از توزیع مورد آزمون است که با آماره مقایسه می شود تا تصمیم نهایی درباره فرض صفر گرفته شود.

نکته ! از Z-Test در مواقعی استفاده می شود که انحراف معیار و میانگین جمعیت مورد آزمایش معلوم باشد.

نکته ! از Z-Test زمانی استفاده می شود که توزیع داده ها به صورت نرمال باشد.

✓ **مثال :** میانگین طول عمر یک مدل لامپ ۶ ماه با انحراف معیار ۰.۵ است. کمترین طول عمر این مدل لامپ ۵ ماه است. شرکت ادعا می کند که طول عمر لامپ ها کمتر از ۵ ماه نمی شود. با استفاده از آزمون فرض بررسی کنید که آیا ادعا شرکت درست می باشد یا خیر.

داریم :

$\mu = 6$
$\sigma = 0.5$
$H_0: \mu \geq 5$
$H_1: \mu < 5$

فردی برای رد ادعای شرکت ۴۰ نمونه از این لامپ ها را تست کرده است و میانگین طول عمر این ۴۰ لامپ برابر ۴.۵ ماه بوده است.

پس داریم :

$n = 40$
$\bar{x} = 4.5$

بر اساس فرمول داریم :

$$Z - score = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{4.5 - 6}{\frac{0.5}{\sqrt{40}}} = -18.97$$

ضریب اطمینان را ۹۵ درصد در نظر می گیریم که بر این اساس α برابر ۰.۰۵ می شود.

با توجه به ضریب اطمینان و مقادیر دیگر ، با استفاده از آدرس زیر می توان تایید یا رد فرض صفر را نتیجه گرفت.

https://www.statskingdom.com/110MeanNormal1.html

پس فرض صفر رد می شود.

با استفاده از پایتون می توان Z-Test را پیاده سازی کرد :

```python
import numpy as np
from statsmodels.stats.weightstats import ztest

np.random.seed(7)
a = np.random.normal(loc = 4.5, scale = 0.5, size = 40)
pop_mean = 6
std = 0.5
n = 40
sample_mean = 4.5

ztest(a, value = pop_mean, alternative = "smaller")
```

T-Test :

نکته ! از T-Test در مواقعی استفاده می شود که فقط میانگین جمعیت مورد آزمایش معلوم باشد.

نکته ! از T-Test زمانی استفاده می شود که توزیع داده ها به صورت نرمال باشد.

- ✓ **مثال :** تصور کنید شما تبلیغ یک رستوران را مشاهده کرید که در آن نوشته شده است غذای شما کمتر از ۳۰ دقیقه آماده می شود. از این رستوران یک بار غذا سفارش می دهید و مشاهده می کنید که غذای شما ۴۰ دقیقه ای آماده می شود. می خواهید ببینید که آیا این ادعای رستوران درست است و به صورت شانسی غذای شما دیر آماده شده است یا خیر. برای اثبات ادعای رستوران ۴۰ بار در مجموع طی روز های مختلف غذا سفارش می دهید. میانگین آماده شدن غذا در این ۴۰ سفارش شما برابر با ۳۹ می شود. با استفاده از T-Test درستی یا عدم درستی ادعای رستوران را نشان دهید.

داریم :

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

همچنین داریم :

$n = 40$
$\bar{x} = 39$

بر اساس فرمول داریم :

$$T - score = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{39 - 30}{\frac{5}{\sqrt{40}}} = 11.38$$

s انحراف معیار نمونه ها است که در این جا مقدار آن ۵ در نظر گرفته شده است.

ضریب اطمینان را نیز برای این مثال ۹۵ درصد در نظر می گیریم.

با توجه به ضریب اطمینان و مقادیر دیگر ، با استفاده از آدرس زیر می توان تایید یا رد فرض صفر را نتیجه گرفت.

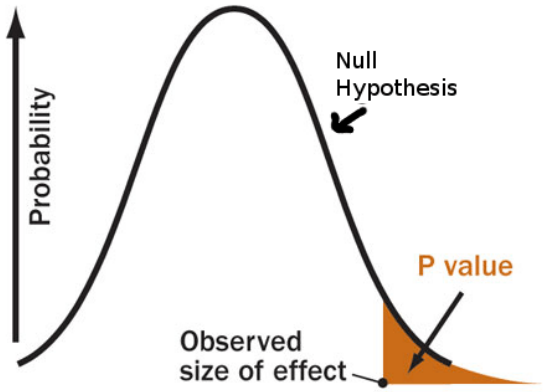
https://www.statskingdom.com/130MeanT1.html

پس فرض صفر رد می شود.

P-Value :

مثال سفارش غذا را به خاطر بیاورید. در آن مثال فرض صفر آن بود که حداکثر تحویل غذا ۳۰ دقیقه طول می کشد. چندین بار غذا سفارش داده شد و مشاهده شد که میانگین تحویل غذا ۳۹ دقیقه بود. با فرض آن که فرض صفر درست است ، P-Value احتمال آن است که در نمونه برداری های دیگر یا طی سفارشات دیگر میانگین تحویل غذا حداقل ۳۹ دقیقه باشد.

یا به عبارت دیگر P-Value را احتمال مشاهده نتایج به دست آمده و نتایج شدیدتر به شرط برقراری فرض صفر تعریف می کنند.



توزیع بالا ، توزیعی نرمال است که مربوط به فرض صفر است.

نکته ! P-Value نمی تواند احتمال درست یا غلط بودن فرض صفر را بیان کند.
نکته ! از P-Value بسیار کم می توان استنباط کرد ، احتمال وقوع نمونه آماری جمع آوری شده در صورت برقراری فرض صفر ، بسیار کم است. اما نمی توان بر اساس آن فرض صفر را رد یا تایید کرد.

با استفاده از آدرس زیر می توان مقدار P-Value را بدست آورد.
https://www.statskingdom.com/p_value.html

نکته ! منظور از DF ، Degrees of Freedom یا درجه آزادی است که مقدار آن یک عدد کمتر از تعداد نمونه ها است. این مقدار در نسخه های 1.10.0. اضافه شده است.

با استفاده از پایتون می توان T-Test پیاده سازی کرد و P-Value را بدست آورد :

```python
import numpy as np
from scipy import stats

np.random.seed(142)
a = np.random.normal(loc = 39, scale = 5, size = 40)
stats.ttest_1samp(a, popmean = 30, alternative = "greater")
```