ریاضیات یادگیری ماشین
(آمار و احتمال)

سجاد ثقفی

محمدرضا احمدی

سرفصل دوره

- آمار و احتمالات
  - آمار توصیفی
  - آمار استنباطی
  - مبانی احتمال
- مقدمه ای بر جبر خطی
  - مفاهیم اولیه
  - انواع ماتریس
  - عملیات های روی ماتریس
- مشتق و گرادیان

کاربرد ریاضیات در یادگیری ماشین

ریاضیات ستون فقرات یادگیری ماشین است. بدون مفاهیم پایه‌ای مثل جبر خطی، آمار و احتمال، حسابان و نظریه اطلاعات، طراحی و درک درست الگوریتم‌های یادگیری ماشین ممکن نیست.

نقش کلیدی ریاضیات در یادگیری ماشین

۱. جبر خطی (Linear Algebra):

- نمایش داده‌ها به صورت بردار و ماتریس
- عملیات روی تصاویر (پیکسل‌ها)، متن‌ها و داده‌های چندبعدی

۲. آمار و احتمال (Statistics & Probability):

- تحلیل داده‌ها و مدیریت عدم قطعیت
- مدل‌سازی توزیع داده‌ها و پیش‌بینی نتایج

۳. حسابان (Calculus):

- محاسبه گرادیان‌ها برای بهینه‌سازی
- آموزش مدل‌ها با روش‌هایی مثل گرادیان نزولی (Gradient Descent)

۱. آمار توصیفی

- انواع داده ها
- مقیاس های اندازه گیری
- شاخص های مرکزی
- شاخص های پراکندگی
- توزیع داده ها و نمودار ها

آمار توصیفی شاخه‌ای از علم آمار است که با جمع‌آوری، سازمان‌دهی، خلاصه‌سازی و نمایش داده‌ها سروکار دارد. هدف آن ارائه‌ی تصویری روشن از ویژگی‌های داده‌های موجود است، بدون آن‌که نتایج به جامعه‌ی بزرگ‌تر تعمیم داده شوند.

آمار توصیفی مجموعه‌ای از روش‌هاست که داده‌های خام را به شکل قابل فهم و منظم تبدیل می‌کند و به ما کمک می‌کند دید کلی و سریع از داده‌ها داشته باشیم.

۱.۱ انواع داده ها

- داده های کمی (Quantitative Data):
  - گسسته (Discrete): مقادیر شمارشی، مثل تعداد دانش‌آموزان یک کلاس.
  - پیوسته (Continuos): مقادیر اندازه‌گیری‌شده، مثل قد، وزن یا دما.

- داده های کیفی (Categorical Data):
  - اسمی (Nominal): دسته‌بندی بدون ترتیب، مثل جنسیت یا رنگ.
  - ترتیبی (Ordinal): دسته‌بندی با ترتیب مشخص، مثل سطح تحصیلات (دیپلم، کارشناسی، کارشناسی ارشد).

۱.۲ مقیاس های اندازه گیری داده ها

نوع مقیاس	ویژگی	مثال
اسمی (Nominal)	دسته بندی بدون ترتیب	رنگ چشم، جنسیت
ترتیبی (Ordinal)	دسته‌بندی با ترتیب	رتبه‌ی مسابقه، سطح تحصیلات
فاصله ای (Interval)	داده‌های عددی با فاصله‌ی مساوی، بدون صفر مطلق	دما بر حسب سانتی‌گراد
نسبی (Ratio)	داده‌های عددی با صفر مطلق	وزن، قد، درآمد

۱.۳ شاخص های مرکزی (Central Tendency)

شاخص های مرکزی معیاری آماری است که نشان میدهد داده‌ها بطور کلی در چه مقدار یا نقطه‌ای متمرکز شده‌اند و یک مقدار نماینده برای کل داده‌ها ارائه می‌دهد:

- میانگین (Mean):
  میانگین برابر مجموع داده‌ها تقسیم بر تعداد آن‌هاست.
  کاربرد: در یادگیری ماشین برای محاسبه مقدار متوسط ویژگی‌ها، نرمال‌سازی داده‌ها و الگوریتم‌هایی مثل رگرسیون و K-Means استفاده می‌شود.
- میانه (Median):
  میانه مقداری است که داده‌های مرتب‌شده را به دو بخش مساوی تقسیم می‌کند.
  کاربرد: وقتی داده‌ها نویز یا داده‌های پرت (Outlier) دارند، در تحلیل داده و پیش‌پردازش ML گزینه‌ی مقاوم‌تری نسبت به میانگین است.
- مد (Mode):
  مد پرتکرارترین مقدار در یک مجموعه داده است.
  کاربرد: در داده‌های دسته‌ای (Categorical)مثل کلاس‌ها، برچسب‌ها و تحلیل توزیع خروجی مدل‌ها استفاده می‌شود.

۱.۴ میانگین (Mean)

$$\frac{1}{n}\sum xi$$

۱.۵ میانه (Median)

- حالت اول: در این حالت تعداد داده ها فرد است:

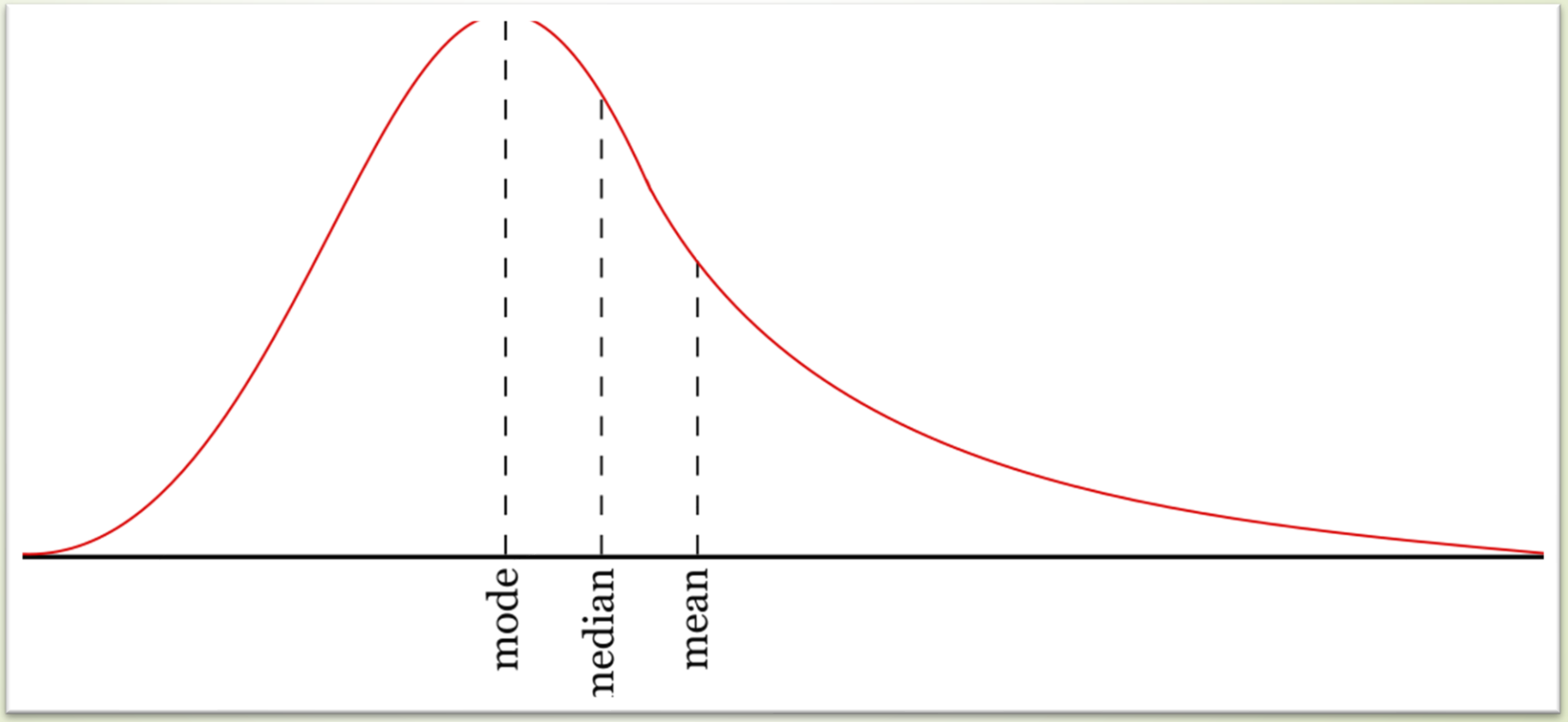
7	47	96	6	3	21	58

- حالت دوم: در این حالت تعداد داده ها زوج است:

1	12	9	4	3	56

۱.۶ مد (Mode)

7	3	5	5	5	47	96	6	3	21	58



۱.۷ شاخص های پراکندگی (Dispersion Measures)

شاخص‌های پراکندگی (Dispersion Measures) میزان پخش‌شدگی و تغییرات داده‌ها حول شاخص مرکزی را نشان می‌دهند:

- واریانس (Variance):
  واریانس میانگین مربعات فاصله داده‌ها از میانگین است.
  کاربرد: در یادگیری ماشین برای سنجش میزان نوسان ویژگی‌ها، تحلیل ریسک و الگوریتم هایی مثل PCA استفاده می‌شود.
- انحراف معیار (Standard Deviation):
  انحراف معیار ریشه دوم واریانس است و میزان پراکندگی داده ها را در همان واحد داده نشان می‌دهد.
  کاربرد: در نرمال سازی داده ها، تشخیص داده پرت و تحلیل پایداری مدل ها کاربرد دارد.
- کوواریانس (Covariance):
  کوواریانس میزان تغییر هم زمان دو متغیر نسبت به میانگین‌هایشان را نشان می‌دهد.
  کاربرد: برای بررسی رابطه بین ویژگی‌ها و پایه‌ی محاسبات ماتریس کوواریانس در الگوریتم‌هایی مثل PCA و تحلیل همبستگی استفاده می‌شود.

۱.۸ واریانس (var) و انحراف معیار

$$S^2 = \frac{1}{n}\sum_{i=0}^{n}(\mathcal{X}i - \bar{x})^2$$

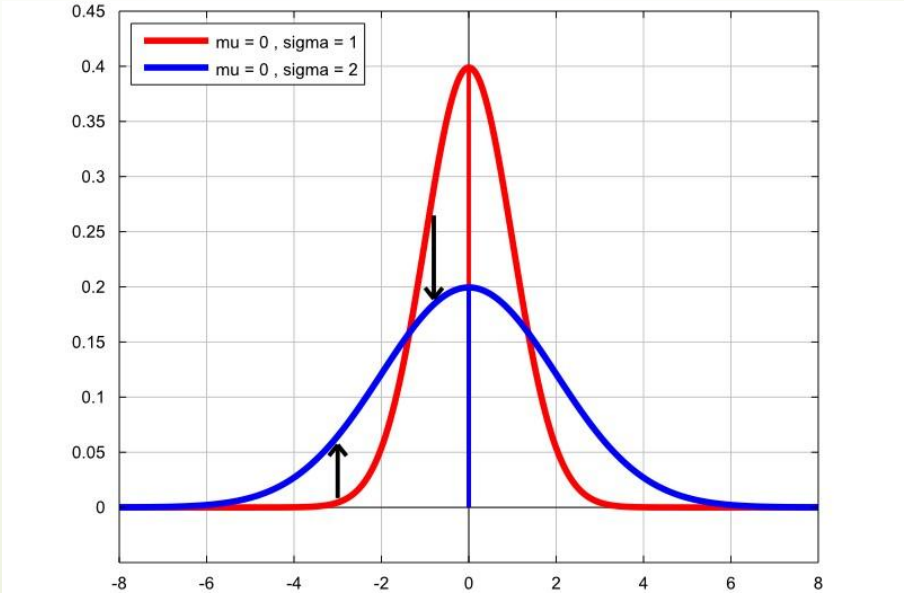
۱.۹ کواریانس(cov)

$$cov(x,y) = \frac{\sum(xi - \bar{x})(yi - \bar{y})}{n}$$

۱.۱۰ توزیع نرمال

توزیع نرمال، یک توزیع پیوسته است که در آن مقادیر متغیر تصادفی به‌صورت متقارن حول یک مقدار میانگین توزیع می‌شوند.

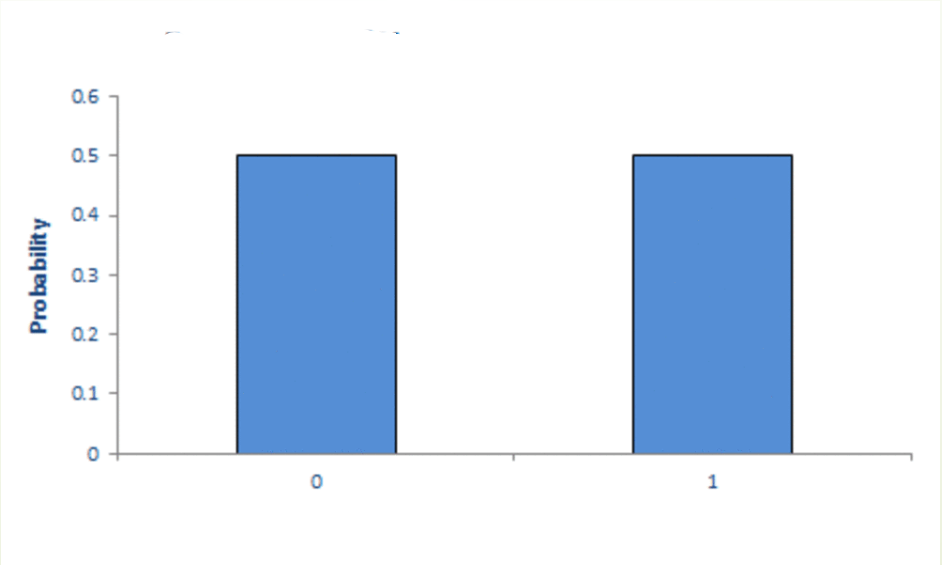
در توزیع نرمال، مقادیر حول یک مقدار مرکزی به نام میانگین متمرکز می‌شوند. این منحنی به‌صورت متقارن شکل‌گرفته و نشان‌دهنده این است که نیمی از مقادیر بالاتر و نیمی دیگر پایین‌تر از میانگین قرار دارند. از دیگر ویژگی‌های این توزیع، می‌توان به انحراف معیار اشاره نمود که نشان‌دهنده میزان پراکندگی داده‌ها نسبت به میانگین است. انحراف معیار تعیین می‌کند که منحنی چقدر پهن یا باریک باشد.



۱.۱۱ توزیع برنولی

توزیع برنولی برای رویدادهایی اعمال می شود که یک آزمایش و دو نتیجه ممکن دارند. اینها به عنوان آزمایشات برنولی شناخته می شوند. به هر نوع آزمایشی فکر کنید که یک سوال بله یا خیر بپرسد.
برای مثال، آیا وقتی این سکه را برگردانم روی شیر فرود میاید؟ آیا با این تاس یک شش رو می کنم؟ آیا دانش آموز Y در آزمون ریاضی خود موفق خواهد شد؟

به عنوان مثال، با استفاده از این ابزار می توان احتمال عوارض جانبی ناشی از یک داروی جدید را اندازه گیری کرد. می تواند احتمال موفقیت یا شکست یک آزمایش پزشکی را تعیین کند. برای سنجش احتمال اسپم بودن ایمیل استفاده می شود. در بازاریابی، این قضیه احتمال خرید یا عدم خرید یک محصول خاص توسط مشتری را پیش بینی می کند.

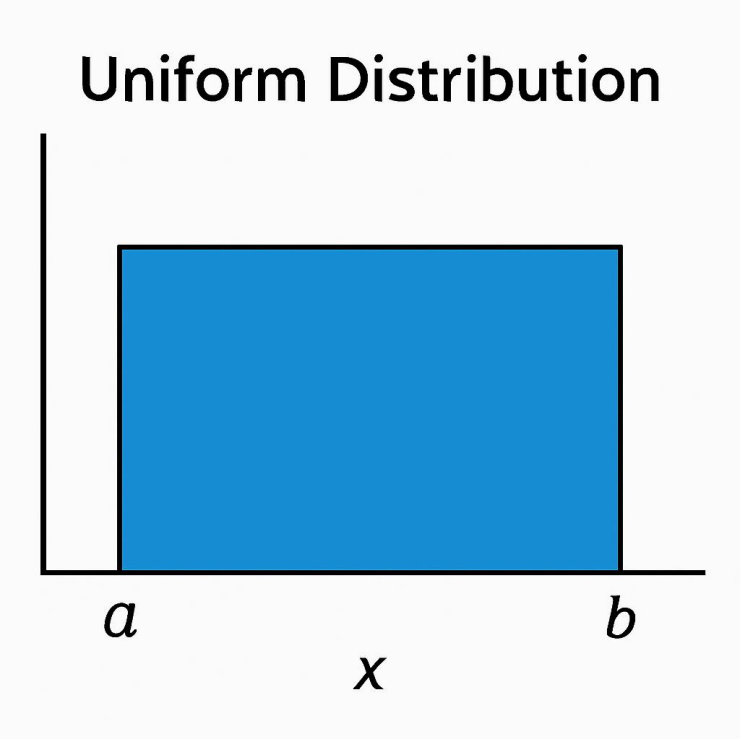


۱۲.۱توزیع یکنواخت

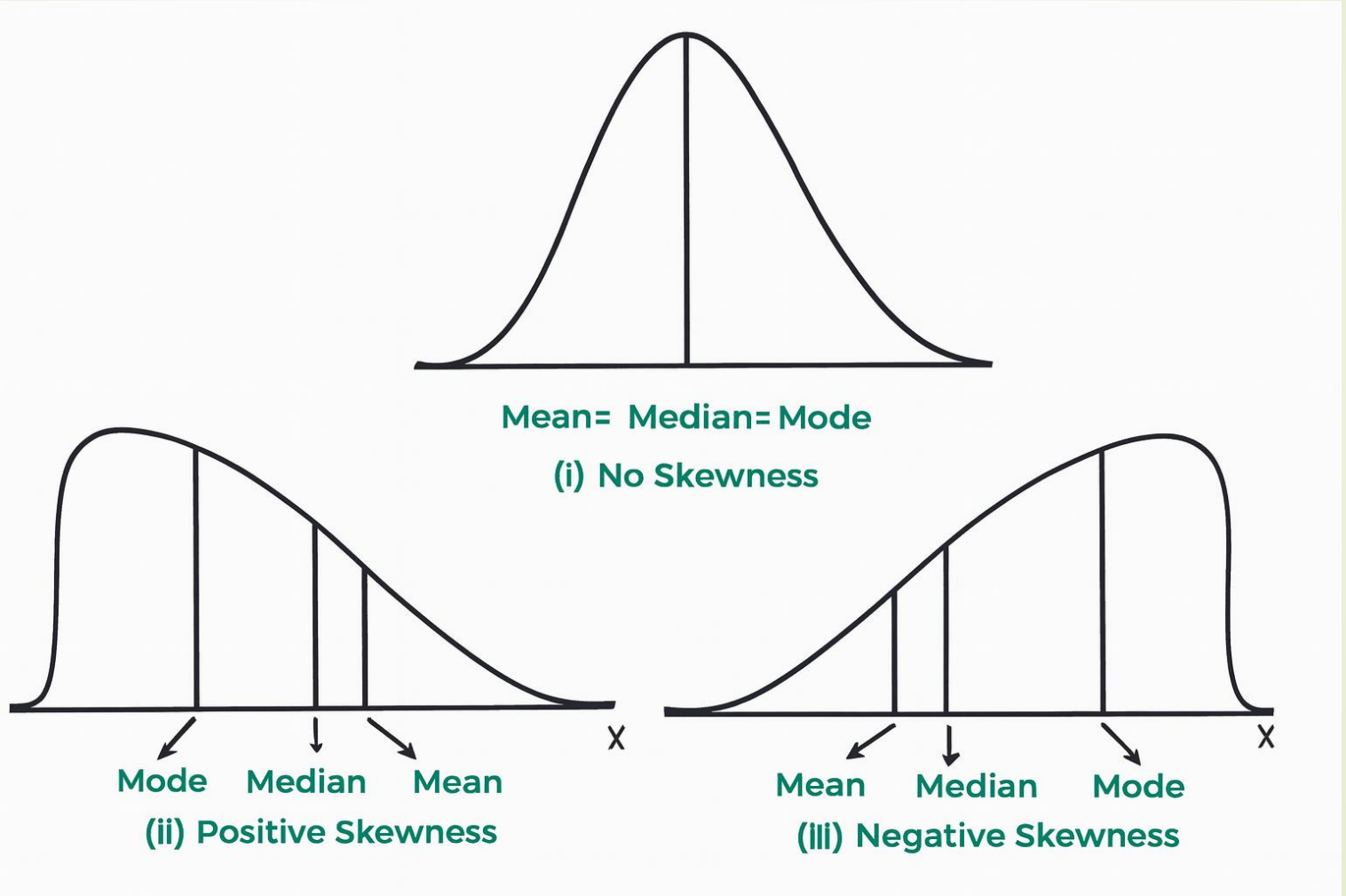
در این توزیع، احتمال وقوع هر کدام از رویدادها، شبیه به هم و برابر یک مقدارِ عددی خاص در بازه‌ای مشخص است.

توزیع یکنواخت به دو دسته تقسیم میشود:

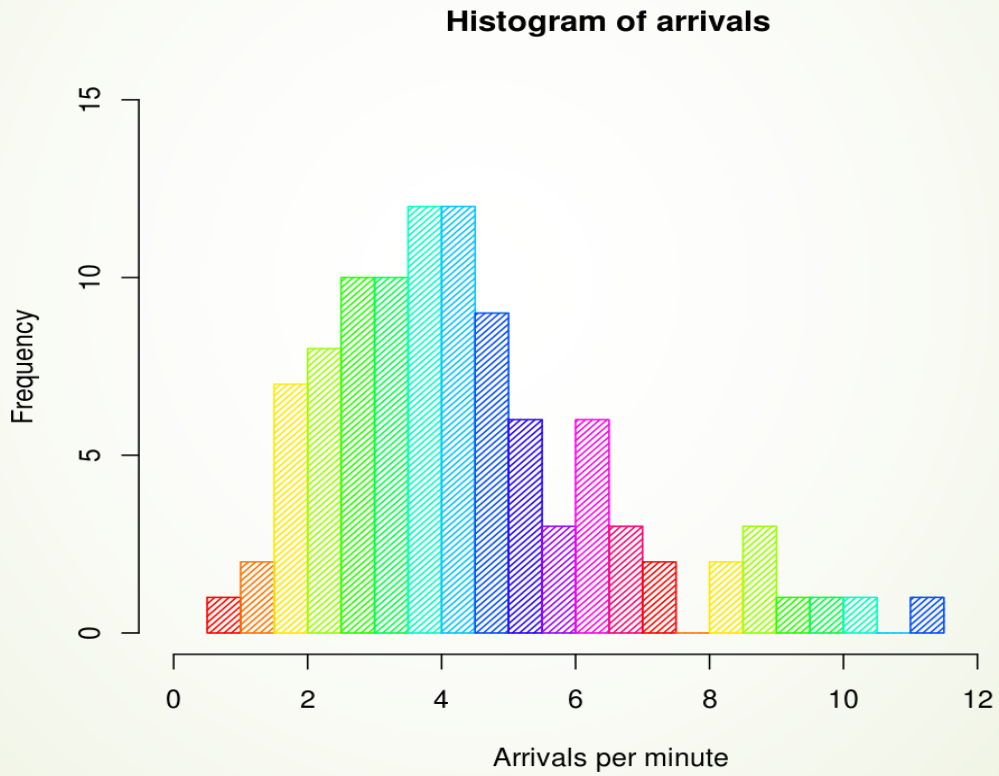
- گسسته: احتمال وقوع هر مقدار از مجموعه‌ای محدود از مقادیر برابر است.
- پیوسته: احتمال وقوع هر مقدار در یک بازه پیوسته از اعداد حقیقی برابر است.



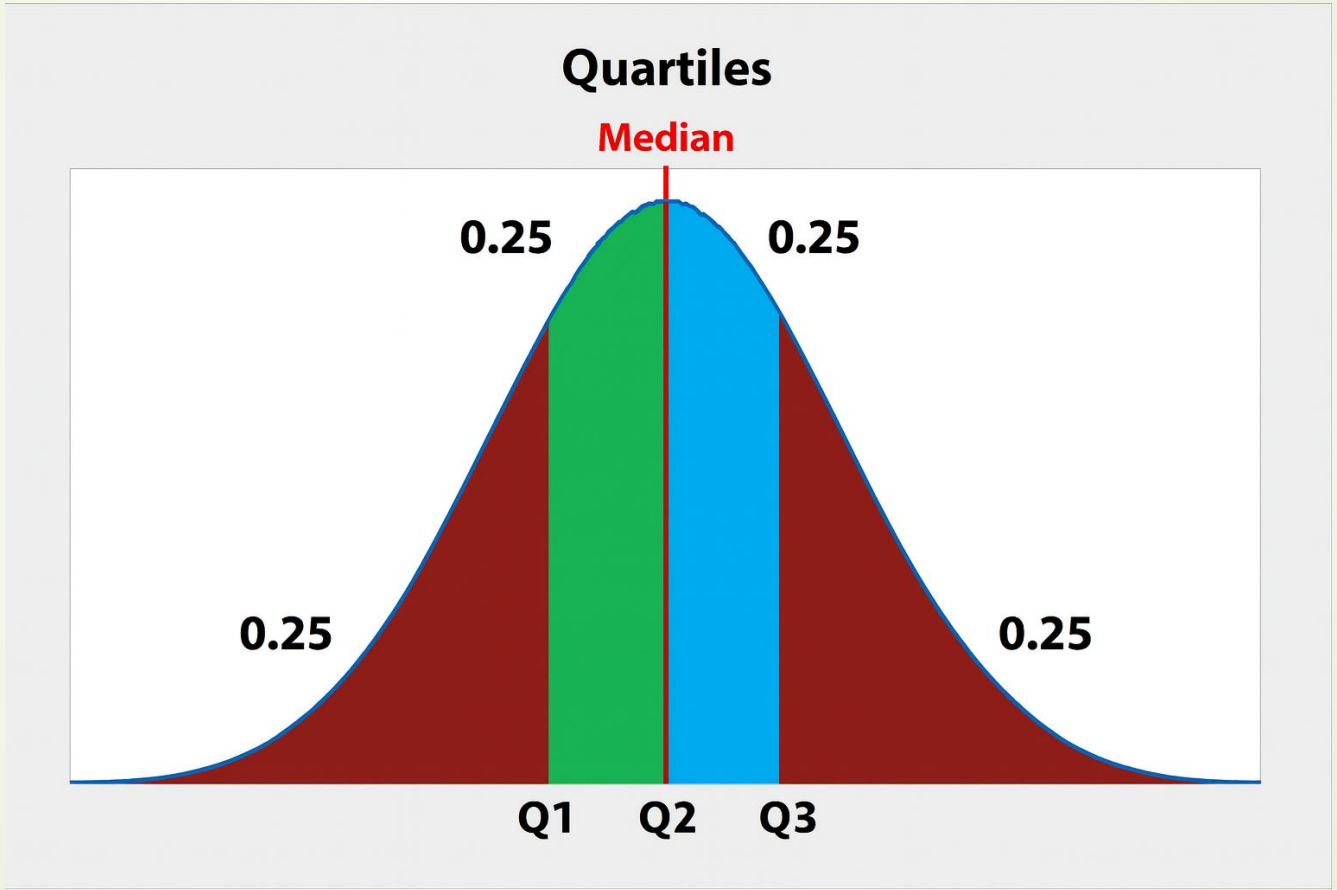
۱.۱۳ چولگی


Mean= Median= Mode
(i) No Skewness
X
Mode
Median
Mean
(ii) Positive Skewness
X
Mean
Median
Mode
(iii) Negative Skewness

۱.۱۴ هیستوگرام



۱.۱۵ چارک (Quartile)



۲. امار استنباطی

- آزمون فرضیه: آزمون فرضیه یعنی یک روش ساده و منطقی برای تصمیم‌گیری درباره‌ی یک ادعا با استفاده از داده‌ها.

کاربرد ها:

- برای بررسی اعتبار ادعاها (مثل داروی جدید مؤثر است، میزان تولید افزایش یافته)
- برای تصمیم‌گیری تحت شرایط عدم قطعیت
- برای نتیجه‌گیری درباره جامعه بر اساس نمونه
- در تحقیقات علمی، کنترل کیفیت، پزشکی، علوم اجتماعی و...

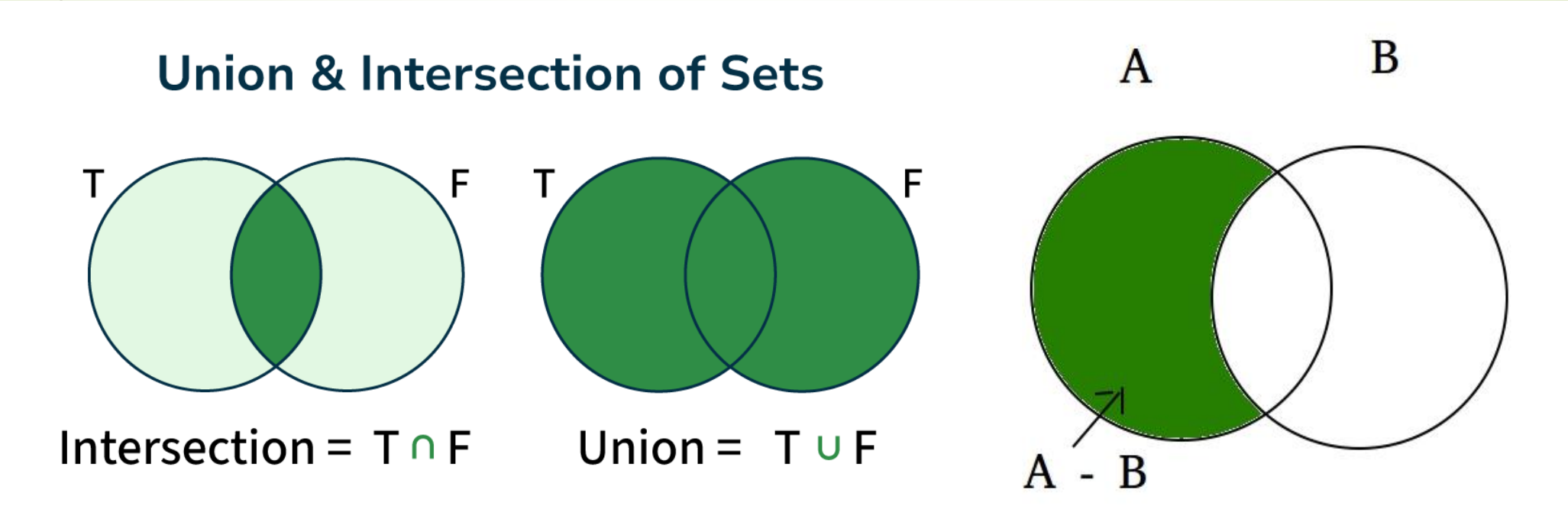
۳. احتمال

- آزمایش تصادفی: آزمایشی که نتیجه آن از قبل در دسترس نیست.
مانند: پرتاب تاس و سکه
- فضای نمونه: نتایج آزمایش تصادفی
- پیشامد: هر زیر مجموعه از فضای نمونه

- مثال: فضای نمونه پرتاب تاس و پیشامدِ اعداد زوج را بیان کنید.

- مثال: فرض میکنیم یک کیسه داریم که ۳ مهره سفید و ۲ مهره قرمز دارد. ۲ مهره را بصورت تصادفی خارج میکنیم، فضای نمونه آن را مشخص کنید.

۳.۱ اجتماع و اشتراک و تفاضل پیشامد ها



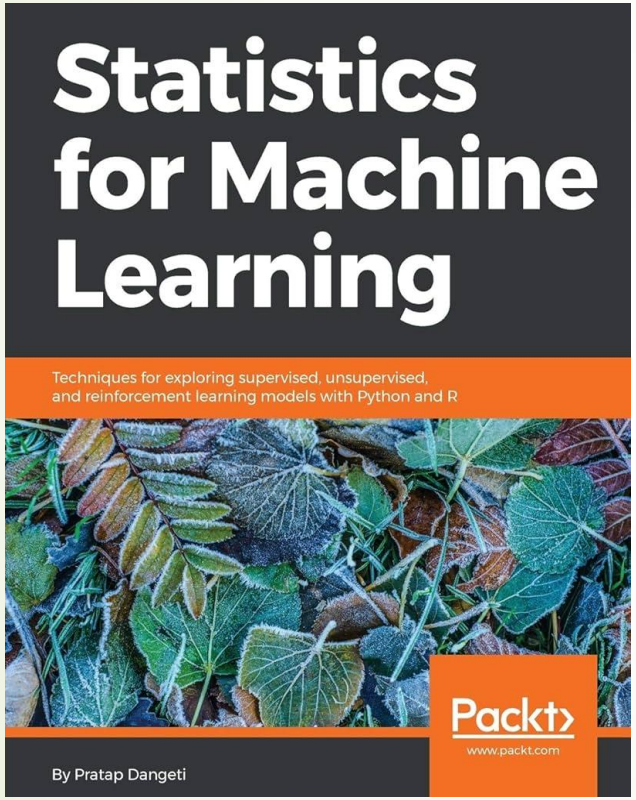
۳.۲ احتمال شرطی

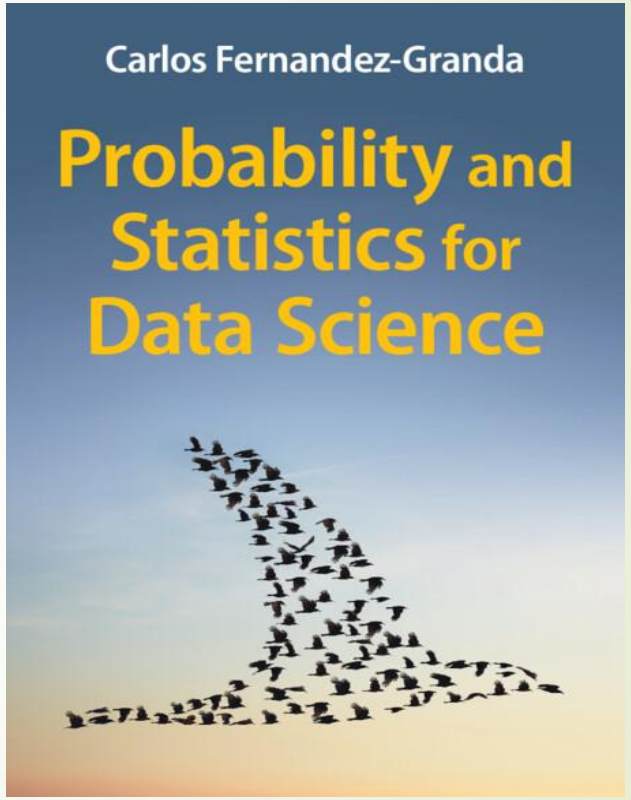
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

۳.۳ قضیه بیز

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

منابع


Statistics for Machine Learning
Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R
Packt
www.packt.com
By Pratap Dangeti


Carlos Fernandez-Granda
Probability and Statistics for Data Science