公式 总数是 N，索引从 0 开始 需要先 sort

Central Tendency

Median = { N 是 odd，则 median 是在[(N-1)/2] idx 上的值 ; N 是 Even，则 median 是在(N/2)-1 和(N/2) idx 上 2 个值的 mean }

Dispersion **Percentiles** =10th percentile is item at index 0.1 * (N - 1)

IQR=Q3-Q1；其中 Q1 表示 25 percentile，Q3 表示 75

Variance = $\frac{\sum(x_i - mean)^2}{N-1}$

Stdev: Square root of variance

对于 interval，ratio 来说，如果 index 不是整数怎么办？这里以 5.4 为例子

- Percentile_Value = value at index 5 + 0.4 * (value at index 6 - value at index 5)

对于 ordinal 来说，如果 index 不是整数怎么办？这里以 5.4 我们选择 idx6 的 value 作为 Percentile_Value

Lower inner fence – Q1–1.5*IQR **Outer fence**: Q1/3 +/- 3*IQR

Euclidean Distance (L2) — q=2

$$d(i,j) = \sqrt[2]{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \ldots + (x_{ip}-x_{jp})^2}$$

Manhattan Distance — q=1

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \ldots + |x_{ip}-x_{jp}|$$

Minkowski Distance

$$d(i,j) = \sqrt[q]{(x_{i1}-x_{j1})^q + (x_{i2}-x_{j2})^q + \ldots + (x_{ip}-x_{jp})^q}$$

$$H(S) = I(S) = -\sum_i P_i log_2 P_i$$

TF = |occurrences of term in doc|

IDF(t, D) = log $\frac{\text{Total number of documents in corpus D}}{\text{Number of documents containing term t}}$

TFIDF = TF * IDF

Reduces the weight of common words across multiple documents

a term appears in fewer documents, it is more likely to be meaningful

Laplace correction

$$p(w|c) = \frac{\text{count(w,c)}+1}{\text{count(word,c)}+\text{count(word)}}$$

- count(w,c) = Count of word w in class c
- count(word,c) = Count of all words in class c
- count(word) = Count of all distinct words in the dataset

Distinctness: = ≠ **Nominal** (distinctness) Examples: ID, eye color, zip codes

Order: < > **Ordinal** (distinctness, order) Examples: ranking, grades, height, educational levels

Differences are meaningful : + - **Interval** (distinctness, order & differences are meaningful) No true zero 0 并不表示"无"

Examples: calendar dates, temperatures in Celsius or Fahrenheit

Ratios are meaningful * / **Ratio** (all 4 properties) Examples: temperature in Kelvin, length, counts, elapsed time

	Nominal	Ordinal	Interval	Ratio
Countable	✓	✓	✓	✓
Order defined		✓	✓	✓
Difference defined (addition, subtraction)			✓	✓
Zero defined (multiplication, division)				✓

Central Tendency 可以看到只有 interval 和 ratio 支持 mean 的计算

	Nominal	Ordinal	Interval	Ratio
Mode	✓	✓	✓	✓
Median		✓	✓	✓
Mean			✓	✓

Dispersion 只有 Interval 和 Ratio 支持 Stdev

	Nominal	Ordinal	Interval	Ratio
Counts / Distribution	✓	✓	✓	✓
Minimum, Maximum		✓	✓	✓
Range		✓	✓	✓
Percentiles		✓	✓	✓
Standard deviation, Variance			✓	✓

Normalization 缩放到 [0,1] 敏感

$$x_{norm} = \frac{x - min(X)}{max(X) - min(X)}$$

Standardization 均值=0，方差=1 不敏感（更鲁棒）

$$x_{stand} = \frac{x - \mu(X)}{\sigma(X)}$$

Data Cleaning Type and name conversion. Filtering of missing or inconsistent data. NA

Unifying semantic data representations. 避免情绪用词 Matching entries from different sources. Rescaling and optional dimensionality reduction.比如 PCA,在 Week8

Pearson's Capture linear relationships 1 indicates a perfect positive linear relationship 0 indicates no linear relationship

Spearman's Capture both linear and nonlinear monotonic relationships 1: 一个变量上升时另一个也总是上升（单调递增）0: 没有单调关系, no monotonic

Plot Properties Bar: Count frequency of each category (Nominal, Ordinal) Pie: 显示部分与整体的比例关系。Histogram: Count frequency of numerical (Interval, Ratio)

Box: 展示 numeric 的分布特征，包括中位数，四分位数，异常值等 Scatterplot: The most suitable to visualize and analyze the correlation between two variables

Ideal Relation unique table name column with unique names All rows have the same structure Every column is atomic — known as 1NF

Every row is unique The order of the rows is immaterial.

Real word Relation — RDBMS (在 ideal 上的扩展) allows "null" support ordering tuples and attributes **allow duplicated rows**

PRIMARY KEY 唯一性 不允许 NULL CANDIDATE KEY 唯一性 最小性（Minimality）可以包含 NULL FOREIGN KEY 允许 NULL Allow duplicate

OLTP maintaining dynamic relationships OLAP large volumes of historical data Time variant: Large volume of historical data

Star schema 1 Fact table denormalization fast query Higher redundancy due to denormalization Integrated: integrating multiple

Snowflake schema 不止一个 Fact table **normalized** slow query lower redundancy due to normalization Non-volatile: Updates infrequent

Observational Study No intervention typically **only establish correlation**

Experimental Study 有 intervention establish **causality**

Independent Variable The variable that is **manipulated**

Dependent Variable The **measure of interest**

Controlled Variables Conditions or factors **kept the same**

Hypothesis Testing Q: Asks whether the **independent variable** has an effect on the

H0: Assumes that there is **no effect. initial assumption.**

H1: Assumes that there **is an effect** or **a difference.**

p-value < α	Strong evidence against the null hypothesis (H_0). 代表极端值几乎不可能发生	✓ Yes

p-value 在原假设 H0 为真的前提下，观察到当前样本数据(一般为极端结果)的概率。比如公平硬币连续 10 次都是 back 的概率是 0.5^10 = 0.00098.

Significance Level (α) the probability of (wrongly) rejecting H0 given that it is true (**Type I error rate**, i.e., false positive)

p-hacking If we perform many tests, we are likely to get false positives

p-value > α	Weak evidence against the null hypothesis (H_0)	✗ No

Type of Error One-side : H1 可以是大于或者小于 Two-side : H1 只能是不等于

	Accept H_0	Reject H_0
H_0 is True	✓ Correct Decision *(No difference)*	✗ Type I Error *(False positive)*
H_0 is False (H_1 is true)	✗ Type II Error *(False negative)*	✓ Correct Decision *(Difference exists)*

Guess \ True Answer	A	B	C	...
A	TP	FN_B		
B	FP_B	TP_B	FP_B	FP_B
C		FN_B	TP	
...		FN_B		TP

Recall for B = $\frac{\text{Cases gussed as B that are B}}{\text{Cases that are B}}$ = $\frac{TP_B}{TP_B + FN_B}$

第一个表示是否正确 FP 表示错误的预测为 Positive

第二个表示预测的结果 FN 表示错误的预测为 Negative

求 P-Value 的方法

Pearson's Correlation — parametric H_0: 两个变量之间没有线性相关关系。

变量是连续的且近似服从正态分布。变量间的关系是线性的。数据是成对的独立观测。

Paired Student's t-test (pairwise t-test) — parametric H0: two population means are equal

The samples are paired (e.g. before and after treatment). Standard deviations are equal.

Populations are normally distributed.

Wilcoxon Signed-Rank Test — nonparametric H0: two related paired samples come from the same **distribution**.

The samples are paired

Unpaired Student's t-test — parametric H0: two population **means** are equal

The samples are independent. Standard deviations are equal (by default).

Populations are normally distributed.

Mann-Whitney U test — nonparametric H0: the **distribution** underlying sample x is the same as that of sample y

The two samples are independent. Does not require normal distribution.

2 组以上

Analysis of Variance (**ANOVA**) — **parametric** H0: two or more groups have the same population **mean**.

the samples are independent. Populations are normally distributed. Standard deviations are equal.

H0: two or more groups have the same population **mean**.

Kruskall-Wallis H-test — nonparametric H0: the population **medians** of all groups are equal

the samples are independent.

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$	Precision = $\frac{TP}{TP+FP}$ 不希望错判 不希望把负类错判成正类
Recall = $\frac{TP}{TP+FN}$ 不希望漏判 不希望把正类错判成负类	F-Score = $\frac{2PR}{P+R}$

Apriori Principle Bottlenecks Generate huge candidate frequent sets (2^N)-1 Multiple scans of database

FP Tree (Frequent-Pattern Tree) — unsupervised learning Avoid costly repeated database scans

Partitional clustering K-Means $O(n \times k \times i \times d)$ **Hierarchical clustering** Agglomerative (bottom-up) divide-and-conquer methodology

Intra-cluster **min**

Inter-cluster **max**

Measures of Cluster Validity

$$SSE = \sum_{i=1}^{k} \sum_{x \in K_i} dist(x, c_i)^2$$

Silhouette coefficient s = 1–a/b if a<b b 是 average distance of x to points in the next nearest cluster

The closer to 1, the better; -1 表示非常差 s = b/a-1 if a>=b a 是 average distance of x to points in its cluster

可以自己设置一个 threshold，如果所有 cluster 的 Silhouette Coefficient 达标，则说明 k 选的很好。

∝ **cohesion** ⊵ **separation**

PCA) — unsupervised

particularly useful when the variables within the data set are highly correlated redundancy in the data

Covariance Matrix PCA Matrix eigenvalues,

off-diagonal values are **different from zero**. = 47.9% of the overall variability

经过 PCA 后，可以发现 PCA component1 占到 ∫ 1.651/3.448

Simple Linear Regression $Y = \alpha + \beta X + \varepsilon$ Ordinary Least Squares (**OLS**) $\hat{\beta} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{cov(x,y)}{Var(x)} = \frac{r(x,y)\cdot sd(y)}{sd(x)}$ $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ SST=SSR+SSE

Multiple linear regression $MSE = J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$ $\theta_j \leftarrow \theta_j - \alpha \cdot \frac{\partial}{\partial \theta_j} J(\theta)$

Add non-linearity

$Y = h_\theta(x) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \ldots + \theta_d x^d = \sum_{j=0}^{d} \theta_j x^j$ $\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$

Coefficient Determination (R^2) $= 1 - \frac{SSE}{SST}$ =SSR/SST

Ranges from 0 to 1

higher values indicating better fit

negative R^2 表示 model is very poor

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ **Smaller** $SSE = \sum_i (y_i - \hat{y}_i)^2$ **Higher**

Logistic Regression $g(Z) = h_\theta(x) = \frac{1}{1 + e^{-(W^T X^{(i)} + b)}}$ $Z = W^T X^{(i)} + b \geq 0 \rightarrow g(Z) \geq 0.5 \rightarrow \hat{y} = 1$ $Z = W^T X^{(i)} + b < 0 \rightarrow g(Z) < 0.5 \rightarrow \hat{y} = 0$

不一定 converge to global min

Cross-Entropy (即 cost function)

$$L = f(w) = -\frac{1}{n}\sum_{i=1}^{n}[y_i log\sigma(W^T X^{(i)} + b) + (1 - y_i) log[1 - \sigma(W^T X_i + b)]]$$

正态分布 (Normal distribution)

Mean = Median = Mode

Right-skewed

Mode < Median < Mean

Left-skewed

Mean < Median < Mode

```
SELECT COUNT(value),
       MIN(value),
       Max(value),
       MAX(value) - MIN(value)                                  AS Range,
       AVG(value)                                               AS Mean,
       MODE()  WITHIN GROUP (ORDER BY value)                    AS Mode,
       PERCENTILE_DISC(0.5) WITHIN GROUP (ORDER BY value)       AS Median,
       PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value)      AS Percentile25,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)      AS Percentile75,
       PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY value)
       - PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY value) AS IQR
FROM Measurement WHERE sensor='temp';
```

Generalization Error: 这是模型在未见过的数据（测试集）上的 误差，也就是模型对新样本的预测能力

Data drift (non-stationary data) retrain Online/incremental learning 动态适应

Indicator features 如果 token 在文档中出现，赋值 1，否则赋值 0

Assumptions to check:

- Normality of residuals
- Homogeneity of variance
- Independence of observations

FROM
WHERE
GROUP BY
HAVING
SELECT
ORDER BY

1. Set minimum support and confidence thresholds
2. Generate frequent 1-itemsets
3. For k ≥ 2:
   a. Generate candidate k-itemsets
   b. Prune candidates using Apriori property
   c. Calculate support for remaining candidates
   d. Keep itemsets meeting minimum support
4. Generate rules from frequent itemsets
5. Keep rules meeting minimum confidence

R [INNER] JOIN S ON join-condition

R ~~[INNER] JOIN~~ S USING (list-of-attributes)

R NATURAL JOIN S

gender CHAR CHECK (gender IN ('M','F','T'))

CREATE TABLE NAME(…);

DROP TABLE NAME CASCADE;

INSERT INTO table (list-of-columns) VALUES (list-of-expression)

UPDATE table SET column = expression WHERE search_condition;

DELETE FROM table WHERE search_condition;

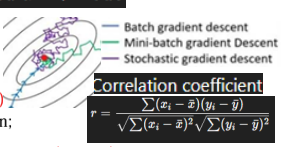
FOREIGN KEY (lecturer_id) REFERENCES Lecturer(lecturer_id)

UNKOWN OR TRUE = True

UNKOWN AND FALSE = False

Support(X) = $\frac{frequency(X)}{total\ transactions}$

Confidence($X \rightarrow Y$) = $\frac{Support(X \cup Y)}{Support(X)}$



Correlation coefficient

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Laplace correction

$$p(w|c) = \frac{\text{count(w,c)}+1}{\text{count(word,c)}+\text{count(word)}}$$

- count(w,c) = Count of word w in class c
- count(word,c) = Count of all words in class c
- count(word) = Count of all distinct words in the dataset

Infrequent itemset → All supersets infrequent

Frequent itemset → All subsets frequent

EXTRACT(DAY FROM event_duration) AS days,

EXTRACT(HOUR FROM event_duration) AS hours,

EXTRACT(MINUTE FROM event_curation) AS minutes

'2012-04-01' + **INTERVAL** '36 HOUR'

DATE '2012-03-01'

CURRENT_DATE

CURRENT_TIME

SQL Type	Example	Description
DATE	'2012-03-26'	A date (some systems incl. time)
TIME	'16:12:05'	A time, often down to nanoseconds
TIMESTAMP	'2012-03-26 16:12:05'	Time at a certain date: SQL Server: DATETIME
INTERVAL	'5 DAY'	A time duration