数据分析

2021年3月29日 13:28

Excel （略）

数据透视表，vlookup，常用函数的使用，基础图表的制作

SQL

https://zhuanlan.zhihu.com/p/38354000

1. **简单查询**
- %表示任意字符串
- 利用count()统计数量

2. **汇总分析**
- 利用sum()计算总和
- DISTINCT 用于返回唯一不同的值。
- 分组group by
- 对分组结果指定条件 having
- 排序order by asc升序 desc降序

3. **复杂查询**
- 子查询
- limit 从查询结果中取出指定行，例如前n个，top(n)，排序后limit
- in 和 not in

4. **多表查询**

   联结

   select * from kemu left join score **on** kemu.id = score.id
- left join 就是“左连接”，表1左连接表2，以左为主，表示以表1为主，关联上表2的数据，查出来的结果显示左边的所有数据，然后右边显示的是和左边有交集部分的数据。
- right join 同理
- join，其实就是“inner join”，为了简写才写成join，两个是表示一个的，内连接，表示以两个表的交集为主，查出来是两个表有交集的部分，其余没有关联就不额外显示出来

   case表达式

   CASE WHEN <表达式> THEN <表达式>
   ELSE <表达式>
   END

5. **窗口函数**

   <窗口函数> over (partition by <用于分组的列名>
   order by <用于排序的列名>)

   partition by用来对表分组

   order by子句的功能是对分组后的结果进行排序

   group by分组汇总后改变了表的行数，一行只有一个类别。而partiition by和rank函数不会减少原表中的行数

<窗口函数>的位置，可以放以下两种函数：

1） 专用窗口函数，rank（1114），dense_rank（1112），row_number（1234）等专用窗口函数。

2） 聚合函数，如sum. avg, count, max, min等

窗口函数原则上只能写在select子句中

窗口函数解决3类问题：

1）分组排名2）topN问题3）每个组内比较

练习题：https://zhuanlan.zhihu.com/p/152233908

Python

- python基本语法
- 数据分析包 numpy pandas matplotlib
- 使用python操作结构化数据，进行数据清洗，数据抽取，数据可视化
- python操作数据库

Hive、spark

统计概率（概率论与数理统计）

1. 概率论的基本概念

https://www.zybuluo.com/catscarf/note/971426

a. 样本空间，随机事件

- 样本空间
  - 集合 S
- 随机事件
  - 集合 $A \subseteq S$
- 基本事件
  - 集合 A 只有一个元素
- 不可能事件
  - 集合 $A = \emptyset$

b. 事件的相互关系及运算

- 事件的关系
  1. 包含 $A \subseteq B$
  2. 相等 $A = B$
  3. 和事件 $A + B$
  4. 积事件 $A \cap B, AB$
  5. 不相容事件，互斥事件 $AB = \emptyset$
  6. 差事件 $A - B$
  7. 逆事件 $\overline{A}$
- 事件关系满足交换律，结合律，德摩根率
- 基本的运算规律
  1. $A + \overline{A} = 1$
  2. $A\overline{A} = \emptyset$
  3. $A - B = A \cap \overline{B} = A - AB$

c. 概率

- 直观定义：随机事件发生的稳定值，记为 $P(A) = p$
- 概率的性质（前三条为概率的公理化定义）
  1. 非负性 $P(\emptyset) = 0$
  2. 规范性 $P(A) = 1 - P(\overline{A})$
  3. 可列可加性
     - 若 A, B 两两互斥
     - $$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$
  4. $$P(B - A) = P(B) - P(AB)$$
  5. 概率的加法公式
     - $$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) - \sum_{1 \le i < j \le n} P(A_i A_j) + \sum_{1 \le i < j < k \le n} P(A_i A_j A_k) + \ldots + (-1)^{n-1} P(A_1 A_2 \ldots A_n)$$

d. 等可能概型（古典模型）

- 特点

  有限性、等可能性
- 组合数

$$C_N^n = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

- 放回抽样、不放回抽样
- 实际推断原理（概率小的事情在单次实验中几乎不会发生）

e. 条件概率

- 定义

$$P(B|A)=\frac{P(AB)}{P(A)},P(A)>0$$

- 乘法公式

$$P(AB)=P(A)P(B|A)$$

f. 全概率公式和贝叶斯定理

- 全概率公式
  - 若$B_1,B_2,B_3,\ldots,B_n$是S的划分（离散数学中的概念），则
  - $$P(A)=\sum_{j=1}^{n}P(B_j)P(A|B_j)$$
  - 关键在于能否构造一个合适的划分
  - 原理是分情况讨论

- 贝叶斯公式
  - $$P(B_i|A)=\frac{P(AB_i)}{P(A)}=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)}$$
  - A是后验概率，B是先验概率。贝叶斯公式描述了先验概率已知的情况下，后验概率对先验概率的修正。
  - **直观理解：**癌症检查中，已知一个人有患癌症的可能，那么后验概率（检查结果）对先验概率（检查前患癌症的可能）的修正，可以增加或减少这个人患癌症的概率。也即医院检查可以（一定概率上）确诊。
  - **作者拓展：**贝叶斯公式在推荐算法上（如搜索引擎排序）具有重要应用，它可以通过用户的点击修正推荐排序结果

g. 事件的独立性

- 事件的独立性常常通过实际情况来判断

  相互独立≠不相关

- 公理化定义
  - 对事件组 $A_1,A_2,\ldots,A_n$，若他们相互独立，则必有
  - $$\begin{aligned}&P(A_iA_j)=P(A_i)P(A_j)\\&P(A_iA_jA_k)=P(A_i)P(A_j)P(A_k)\\&\cdots\\&P(A_1A_2\ldots A_n)=P(A_1)P(A_2)\ldots P(A_n)\end{aligned}$$
  - 注意，若三个事件两两独立，不能推出三个事件相互独立
- 性质
  - 若 A,B 相互独立，则 $\overline{A},B$，$A,\overline{B}$，$\overline{A},\overline{B}$ 也相互独立
- 小概率事件

  小概率事件在一次实验中几乎不发生

  但在大规模重复实验中，至少有一次发生的概率非常高

2. 概率论的基本概念

a. 随机变量

- 定义

- 随机变量 $X(e)$，X 是 $S \to R$ 的函数，e 是样本点
- 自变量 $e \subset S$
- 随机事件 $A = \{e|X(e) = I\} = \{X = I\}$
- 如多次投掷骰子，随机事件 {6点在第3次出现} 可以记作 $X = 3$，X 是随机变量

- 随机变量

离散型随机变量，值的集合的基数小于等于阿列夫零（离散数学概念）

连续型随机变量

- 分布律

X	x_1	x_2	$\dots$	x_k	$\dots$
P	p_1	p_2	$\dots$	p_k	$\dots$

$$P(X = x_k) = p_k \ (k = 1, 2, \dots)$$

- 几何分布 Geometric Distribution
  - 多次投掷骰子，6 点第一次出现时投掷的次数

X	1	2	3	$\dots$	k	$\dots$
P	$\frac{1}{6}$	$\frac{5}{6} \cdot \frac{1}{6}$	$(\frac{5}{6})^2 \cdot \frac{1}{6}$	$\dots$	$(\frac{5}{6})^{k-1} \cdot \frac{1}{6}$	$\dots$

b. 离散型随机变量

- 0-1分布

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

X	0	1
P	$1-p$	p

  - 若X服从两点分布，则单次试验称为伯努利（Bernoulli）试验
  - 记为$X \sim 0-1(p)$
  - 也记为 $X \sim B(1, p)$，B 是Binomial的意思，两点分布可以看作Binomial分布的特例
  - $\sim$ 读作服从于
- 二项分布 Binomial Distribution

$$P(X = k) = C_n^k \cdot P^k \cdot (1-p)^{n-k}$$

  - n 重Bernoulli实验，事件发生次数 k 的统计规律
  - 记为$X \sim B(n, p)$
- 泊松分布 Poisson Distribution

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \ (k = 0, 1, 2, \dots)$$

  - 记为 $X \sim \pi(\lambda)$ 或 $x \sim P(\lambda)$

与二项分布的关系

- 当 n 很大 p 很小的时候

- 当 n 很大 p 很小的时候
- $$C_n^k P^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!} \quad \lambda = np$$

- 几何分布

  定义：在n次伯努利试验中，试验k次才得到第一次成功的机率。

  详细：前k-1次皆失败，第k次成功的概率。

  - $P(X=k) = p(1-p)^{k-1}$
  - 记为 $X \sim Geom(p)$
  - 实例：研究段誉多少次施展武功能成功的统计规律

c. 分布函数

- 定义
  - $F_X(x) = P(X \leq x)$
- 离散型的随机变量分布函数为阶梯函数
- 性质
  - $P(a < X \leq b) = F(b) - F(a)$
  - $P(a < X < b) = F(b-0) - F(a)$
  - $P(X = b) = F(b) - F(b-0)$
  - $F(x)$ 单调不减
  - $F(-\infty) = 0, F(+\infty) = 1$
  - $F(x)$ 右连续

d. 连续性随机变量及其概率密度

- 定义
  - $$F(x) = \int_{-\infty}^{x} f(t)dt$$
  - $F(x)$ 为连续型随机变量的分布函数
  - $f(t)$ 为连续型随机变量的概率密度函数
  - 若一个随机变量有概率密度函数则其一定为随机变量
- 性质

  1. $$f(x) \geq 0$$
  2. $$F(+\infty) = 1$$
  3. $$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(t)dt$$
  4. $$F'(x) = f(x)$$
  5. $$f(x)$$

     可以大于1
  6. 概率密度对 $>, \geq, <, \leq$ 不敏感，即对端点取值不敏感

e. 均匀分布和指数分布

- 均匀分布 Uniform Distribution
  - $$f(x) = \frac{1}{b-a} \quad a \leq x < b$$
  - $$F(x) = \frac{x-a}{\quad} \quad a < x < b$$

- $$f(x) = \frac{1}{b-a} \quad a \le x < b$$
- $$F(x) = \frac{x-a}{b-a} \quad a \le x < b$$
- 记为 $X \sim U(a,b)$ 或 $X \sim Unif(a,b)$

- 指数分布 Exponential Distribution
  - $$f(x) = \lambda e^{-\lambda x} \quad x > 0$$
  - $$F(x) = 1 - e^{-\lambda x} \quad x > 0$$
  - 记为 $X \sim E(\lambda)$ 或 $X \sim Emp(\lambda)$
  - 指数分布具有无记忆性（Memoryless Property）且在连续性随机变量的分布中，只有指数分布具有无记忆性
  - 实例：设旅客等待时间服从指数分布，则已知旅客已经等了20分钟，求旅客再等5分钟的概率，和旅客从头开始等5分钟的概率相同
  - 即 $P(X > 25|X > 20) = P(X > 5)$
  - 指数分布常用来表示独立随机事件发生的时间间隔，如中文维基百科新条目出现的时间间隔
  - 在排队论中，一个顾客接受服务的时间长短也可以用指数分布来近似

f. 正态分布

- 正态分布 Normal Distribution
  - $$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
  - 记为 $X \sim N(\mu, \sigma^2)$

- 性质
  - 关于 $x = \mu$ 对称
  - $$f_{max} = f(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$$
  - $$limf(x) = 0$$

- 参数性质
  - 改变 μ，$f(x)$ 只沿 x 轴评议
  - σ 越大，$f(x)$ 越矮胖，σ 称为尺度参数
- 实例：身高，体重，测量误差，多个随机变量的和
- 标准正态分布
  - $$Z \sim N(0,1)$$
  - $$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$
  - $$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$
  - $\Phi(z)$ 有标准正态分布函数表
- 一般正态分布转为标准正态分布

- 当 $X \sim N(\mu, \sigma^2)$ 时，$(x-\mu)/\sigma \sim N(0,1)$
- $$F_x(a) = P(x \le a) = P(\frac{x-\mu}{\sigma} \le \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$$

- 3δ标准
  - 当 x 落在 $(-3\sigma, 3\sigma)$ 的概率为 99.73%

g. 随机变量函数的分布

- 已知 X 的概率分布，已知 $Y = g(x)$，求 Y 的概率分布
  - 先给出 Y 的可能分布，再利用等价事件来给出概率分布
  - 离散型随机变量，直接利用分布律求解即可
  - 连续型随机变量，先利用分布函数找到等价事件，再利用概率密度函数即可
- 定理
  - 若 $Y = g(x)$，$g'(x) > 0$ 或 $g'(x) < 0$
  - $$f_Y(y) = f_x(h(y)) \cdot |h'(y)| \ \alpha < y < \beta$$
  - $h(y)$ 是 $g(x)$ 的概率密度函数的反函数
  - α 和 β 是根据 x 与 y 的对应关系求得的
- 一般的
  - 若 $X \sim N(\mu, \sigma^2)$，$Y = aX + b$，则 $Y \sim (a\mu + b, a^2\sigma)$

- **当前的所有分布**
  二项分布 Binomial Distribution
  泊松分布 Poisson Distribution
  几何分布 Geometric Distribution
  均匀分布 Uniform Distribution
  指数分布 Exponential Distribution
  正态分布 Normal Distribution

3. 二元随机变量及其分布

a. 二元随机变量，离散型随机变量分布律
b. 二元离散型随机变量边际分布律与条件分布律
c. 二元随机变量分布函数、边际分布函数及条件分布函数
d. 二元连续型随机变量，联合概率密度
e. 二元连续型随机变量边际概率密度
f. 二元连续型随机变量条件概率密度
g. 二元均匀分布，二元正态分布
h. 随机变量的独立性
i. 二元随机变量函数的分布
j. Z=X+Y的分布
k. MAX(X,Y)和MIN(X,Y)的分布

4. 随机变量的数字特征

a. 随机变量的数学期望
b. 随机变量函数的数学期望
c. 数学期望的性质
d. 方差定义和计算公式

e. 方差的性质
f. 协方差与相关系数
g. 不相关与独立
h. 矩，协方差矩阵，多元正态分布的性质

5. 大数定律及中心极限定理

a. 依概率收敛，切比雪夫不等式
b. 大数定律
c. 中心极限定理

6. 统计量与抽样分布

a. 总体，样本
b. 统计量，常用统计量
c. X^2分布
d. t分布， F分布
e. 单个正态总体的抽样分布
f. 两个正态总体的抽样分布

7. 参数估计

a. 点估计，矩估计
b. 极大似然估计
c. 估计量的评价准则，无偏性
d. 有效性，均方误差
e. 相合性
f. 置信区间，置信限
g. 枢轴量法
h. 单个正态总体均值的区间估计
i. 成对数据均值差，单个正态总体方差的区间估计
j. 两个正态总体参数的区间估计

8. 假设检验

a. 假设检验的基本思想
b. 单个正态总体参数假设检验（标准差已知，检验）
c. 单个正态总体参数假设检验（标准差未知，检验）
d. 单个正态总体参数假设检验（成对数据检验和参数的检验）
e. 两个正态总体参数假设检验(比较两个正态总体均值的检验)
f. 两个正态总体参数假设检验(比较两个正态总体方差的检验)
g. 拟合优度检验

9. 方差分析与回归分析（略）

a. 单因素方差分析
b. 单因素方差分析（参数估计及均值的多重比较）
c. 回归分析（参数估计）
d. 回归分析（模型检验与应用）

机器学习

- 分类算法：逻辑回归、贝叶斯、决策树、随机森林
  数据挖掘算法——常用分类算法总结
  https://blog.csdn.net/songguangfan/article/details/92581643
- 回归算法：线性回归
  【算法】七种常用的回归算法 https://cloud.tencent.com/developer/article/1102103
- 聚类算法：K-means

- **特征工程**
  深度了解特征工程 https://zhuanlan.zhihu.com/p/111296130
- **模型评价**
  模型的评价方法 https://zhuanlan.zhihu.com/p/53774460
- **交叉检验**
  Cross-Validation（交叉验证）详解 https://zhuanlan.zhihu.com/p/24825503

业务问题

- 数据分析思维
  数据驱动决策的13种思维方式 https://www.jianshu.com/p/5a8f01fe7f2a
  数据分析师必备的20种分析思维 https://zhuanlan.zhihu.com/p/83138160
  业务指标：漏斗思维、分类思维、平衡思维、A/Btest，金字塔原理

- **DAU下降**（日活跃用户数量）
  https://www.sohu.com/a/449047677_114819

- **漏斗思维**
  漏斗思维 https://www.jianshu.com/p/32414ca5895f
  五个经典漏斗模型，看漏斗思维穿透流程化的本质
  http://www.woshipm.com/operate/3269415.html

- **A/B test**
  (分割测试或桶测试）是一种将网页或应用程序的两个版本相互比较以确定哪个版本的性能更好的方法 https://cloud.tencent.com/developer/article/1496302
  分离式组间试验，也叫对照试验，科研领域（药物测试）中已广泛应用
  https://blog.csdn.net/m0_37773338/article/details/108561193

- **金字塔原理**
  金字塔原理（简析） https://zhuanlan.zhihu.com/p/44423303
  什么是金字塔原理，如何运用金字塔原理？ https://www.sohu.com/a/402879711_404038

- **相关性和因果关系的区别**