本篇文章，主要介绍数据分析常用语：

Source：https://mp.weixin.qq.com/s/VsxF84HTbg-tjsJgp7Go-w

**1、绝对数和相对数**

绝对数：是反应客观现象总体在一定时间、一定地点下的总规模、总水平的综合性指标，也是数据分析中常用的指标。比如年GDP，总人口等等。

相对数：是指两个有联系的指标计算而得出的数值，它是反应客观现象之间的数量联系紧密程度的综合指标。相对数一般以倍数、百分数等表示。相对数的计算公式：

相对数=比较值（比数）/基础值（基数）

**2、百分比和百分点**

百分比：是相对数中的一种，它表示一个数是另一个数的百分之几，也称为百分率或百分数。百分比的分母是100，也就是用1%作为度量单位，因此便于比较。

百分点：是指不同时期以百分数的形式表示的相对指标的变动幅度，1%等于1个百分点。

**3、频数和频率**

频数：一个数据在整体中出现的次数。

频率：某一事件发生的次数与总的事件数之比。频率通常用比例或百分数表示。

**4、比例与比率**

比例：是指在总体中各数据占总体的比重，通常反映总体的构成和比例，即部分与整体之间的关系。

比率：是样本(或总体)中各不同类别数据之间的比值，由于比率不是部分与整体之间的对比关系，因而比值可能大于1。

**5、倍数和番数**

倍数：用一个数据除以另一个数据获得，倍数一般用来表示上升、增长幅度，一般不表示减少幅度。

番数：指原来数量的2的n次方。

**6、同比和环比**

同比：指的是与历史同时期的数据相比较而获得的比值，反应事物发展的相对性。

环比：指与上一个统计时期的值进行对比获得的值，主要反映事物的逐期发展的情况。

**7、变量**

变量来源于数学，是计算机语言中能储存计算结果或能表示值抽象概念。变量可以通过变量名访问。

**8、连续变量**

在统计学中，变量按变量值是否连续可分为连续变量与离散变量两种。在一定区间内可以任意取值的变量叫连续变量，其数值是连续不断的，相邻两个数值可作无限分割，即可取无限个数值。如:年龄、体重等变量。

**9、离散变量**

离散变量的各变量值之间都是以整数断开的，如人数、工厂数、机器台数等，都只能按整数计算。离散变量的数值只能用计数的方法取得。

**10、定性变量**

又名分类变量：观测的个体只能归属于几种互不相容类别中的一种时，一般是用非数字来表达其类别，这样的观测数据称为定性变量。可以理解成可以分类别的变量，如学历、性别、婚否等。

**11、均值**

即平均值，平均数是表示一组数据集中趋势的量数，是指在一组数据中所有数据之和再除以这组数据的个数。

**12、中位数**

对于有限的数集，可以通过把所有观察值高低排序后找出正中间的一个作为中位数。如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。

**13、缺失值**

它指的是现有数据集中某个或某些属性的值是不完全的。

**14、异常值**

指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。

**15、方差**

是衡量随机变量或一组数据时离散程度的度量。概率论中方差用来度量随机变量和其数学期望（即均值）之间的偏离程度。统计中的方差（样本方差）是每个样本值与全体样本值的平均数之差的平方值的平均数。在许多实际问题中，研究方差即偏离程度有着重要意义。方差是衡量源数据和期望值相差的度量值。

**16、标准差**

又常称均方差，是离均差平方的算术平均数的平方根，用σ表示。标准差是方差的算术平方根。标准差能反映一个数据集的离散程度。平均数相同的两组数据，标准差未必相同。

**17、皮尔森相关系数**

皮尔森相关系数是用来反映两个变量线性相关程度的统计量。相关系数用r表示，其中n为样本量，分别为两个变量的观测值和均值。r描述的是两个变量间线性相关强弱的程度。r的绝对值越大表明相关性越强。

**18、PV（Page View）页面浏览量**

指某段时间内访问网站或某一页面的用户的总数量，通常用来衡量一篇文章或一次活动带来的流量效果，也是评价网站日常流量数据的重要指标。PV可重复累计，以用户访问网站作为统计依据，用户每刷新一次即重新计算一次。

**19、UV（Unique Visitor）独立访客**

指来到网站或页面的用户总数，这个用户是独立的，同一用户不同时段访问网站只算作一个独立访客，不会重复累计，通常以PC端的Cookie数量作为统计依据。

**20、Visit 访问**

指用户通过外部链接来到网站，从用户来到网站到用户在浏览器中关闭页面，这一过程算作一次访问。

**21、Bounce Rate 跳出率**

指用户通过链接来到网站，在当前页面没有任何交互就离开网站的行为，这就算作此页面增加了一个“跳出”，跳出率一般针对网站的某个页面而言。

跳出率=在这个页面跳出的用户数/PV

**22、退出率**

一般针对某个页面而言。指用户访问某网站的某个页面之后，从浏览器中将与此网站相关的所有页面全部关闭，就算此页面增加了一个“退出“。

退出率=在这个页面退出的用户数/PV

**23、Click 点击**

一般针对付费广告而言，指用户点击某个链接、页面、banner的次数，可重复累计。比如我在PC端看到一则新闻链接点进去看了一会就关了，过了一会又点进去看了一遍，这就算我为这篇新闻贡献两次点击。

**24、avr.time 平均停留时长**

指某个页面被用户访问，在页面停留时长的平均值，通常用来衡量一个页面内容的质量。

avr.time=访客数量/用户总停留时长

**25、CTR 点击率**

指某个广告、Banner、URL被点击的次数和被浏览的总次数的比值。一般用来考核广告投放的引流效果。

CTR=点击数（click）/被用户看到的次数

**26、Conversion rate 转化率**

指用户完成设定的转化环节的次数和总会话人数的百分比，通常用来评价一个转化环节的好坏，如果转化率较低则需优化该转化环节。转化率=转化会话数/总会话数

**27、漏斗**

通常指产生目标转化前的明确流程，比如在淘宝购物，从点击商品链接到查看详情页，再到查看顾客评价、领取商家优惠券，再到填写地址、付款，每个环节都有可能流失用户，这就要求商家必须做好每一个转化环节，漏斗是评价转化环节优劣的指标。

**28、投资回报率（ROI：Return On Investment ）**

反映投入和产出的关系，衡量我这个投资值不值得，能给到我多少价值的东西（非单单的利润），这个是站在投资的角度或长远生意上看的。

其计算公式为：投资回报率（ROI）=年利润或年均利润/投资总额×100%，通常用于评估企业对于某项活动的价值，ROI高表示该项目价值高。

**29、重复购买率**

指消费者在网站中的重复购买次数

**30、流失分析（Churn Analysis/Attrition Analysis）**

描述哪些顾客可能停止使用公司的产品/业务，以及识别哪些顾客的流失会带来最大损失。流失分析的结果用于为可能要流失的顾客准备新的优惠。

**31、顾客细分&画像（Customer Segmentation & Profiling）**

根据现有的顾客数据，将特征、行为相似的顾客归类分组。描述和比较各组。

**32、顾客的生命周期价值 (Lifetime Value, LTV)**

顾客在他/她的一生中为一个公司产生的预期折算利润。

**33、购物篮分析（Market Basket Analysis）**

识别在交易中经常同时出现的商品组合或服务组合，例如经常被一起购买的产品。此类分析的结果被用于推荐附加商品，为陈列商品的决策提供依据等。

**34、实时决策（Real Time Decisioning, RTD）**

帮助企业做出实时（近乎无延迟）的最优销售/营销决策。比如，实时决策系统（打分系统）可以通过多种商业规则或模型，在顾客与公司互动的瞬间，对顾客进行评分和排名。

**35、留存/顾客留存（Retention / Customer Retention)**

指建立后能够长期维持的客户关系的百分比。

**36、相关性分析(Correlation analysis)**

是一种数据分析方法，用于分析变量之间是否存在正相关，或者负相关。

**37、生存分析（Survival Analysis）**

估测一名顾客继续使用某业务的时间，或在后续时段流失的可能性。此类信息能让企业判断所要预测时段的顾客留存，并引入合适的忠诚度政策。

**38、算法(Algorithms)**

可以完成某种数据分析的数学公式。

**39、商业智能（Business Intelligence）**

分析数据、展示信息以帮助企业的执行者、管理层、其他人员进行更有根据的商业决策的应用、设施、工具、过程。

**40、分类分析(Classification analysis)**

从数据中获得重要的相关性信息的系统化过程;这类数据也被称为元数据(meta data),是描述数据的数据。

**41、聚类分析(Clustering analysis)**

它是将相似的对象聚合在一起，每类相似的对象组合成一个聚类(也叫作簇)的过程。这种分析方法的目的在于分析数据间的差异和相似性。

**42、对比分析(Comparative analysis)**

在非常大的数据集中进行模式匹配时，进行一步步的对比和计算过程得到分析结果。

**43、数据分析(Data Analysis)**

是指根据分析目的，用适当的分析方法及工具，对数据进行处理与分析，提取有价值的信息，形成有效结论的过程。

**44、数据处理（Data Processing）**

数据处理是指根据数据分析的目的，将收集到的数据进行加工、整理，形成适合数据分析的样式，它是数据分析前必不可少的阶段。

**45、数据挖掘（Data mining)**

数据挖掘是通过使用复杂的模式识别技术，从而找到有意义的模式，并得出大量数据的见解。

**46、数据清洗(Data cleansing)**

对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

**47、数据质量（Data Quality）**

有关确保数据可靠性和实用价值的过程和技术。高质量的数据应该忠实体现其背后的事务进程，并能满足在运营、决策、规划中的预期用途。

**48、数据建模(Data modelling)**

使用数据建模技术来分析数据对象，以此洞悉数据的内在涵义。

**49、数据集(Data set)**

大量数据的集合。

**50、判别分析(Discriminant analysis)**

将数据分类，按不同的分类方式，可将数据分配到不同的群组，类别或者目录。是一种统计分析法，可以对数据中某些群组或集群的已知信息进行分析，并从中获取分类规则。

**51、探索性分析(Exploratory analysis)**

在没有标准的流程或方法的情况下从数据中发掘模式。是一种发掘数据和数据集主要特性的一种方法。

**52、机器学习(Machine learning)**

人工智能的一部分，指的是机器能够从它们所完成的任务中进行自我学习，通过长期的累积实现自我改进。

**53、网络分析(Network analysis)**

分析网络或图论中节点间的关系，即分析网络中节点间的连接和强度关系。

**54、异常值检测(Outlier detection)**

异常值是指严重偏离一个数据集或一个数据组合总平均值的对象，该对象与数据集中的其他它相去甚远，因此，异常值的出现意味着系统发生问题，需要对此另加分析。

**55、模式识别(Pattern Recognition)**

通过算法来识别数据中的模式，并对同一数据源中的新数据作出预测

**56、预测分析(Predictive analysis)**

大数据分析方法中最有价值的一种分析方法，这种方法有助于预测个人未来(近期)的行为，例如某人很可能会买某些商品，可能会访问某些网站，做某些事情或者产生某种行为。通过使用各种不同的数据集，例如历史数据，事务数据，社交数据，或者客户的个人信息数据，来识别风险和机遇。

**57、回归分析(Regression analysis)**

确定两个变量间的依赖关系。这种方法假设两个变量之间存在单向的因果关系(译者注：自变量，因变量，二者不可互换)。

**58、路径分析(Routing analysis)**

针对某种运输方法通过使用多种不同的变量分析从而找到一条最优路径，以达到降低燃料费用，提高效率的目的。

**59、情感分析(Sentiment Analysis)**

通过算法分析出人们是如何看待某些话题。

**60、SQL**

在关系型数据库中，用于检索数据的一种编程语言。

**61、时序分析(Time series analysis)**

分析在重复测量时间里获得的定义良好的数据。分析的数据必须是良好定义的，并且要取自相同时间间隔的连续时间点。

**62、文本挖掘（Text Mining）**

对包含自然语言的数据的分析。对源数据中词语和短语进行统计计算，以便用数学术语表达文本结构，之后用传统数据挖掘技术分析文本结构。

**63、可视化(Visualization)**

只有正确的可视化，原始数据才可被投入使用。这里的“可视化”并非普通的图型或饼图，可视化指是的复杂的图表，图表中包含大量的数据信息，但可以被很容易地理解和阅读。

**64、仪表板(Dashboard)**

使用算法分析数据，并将结果用图表方式显示于仪表板中。