1.几种趋势的量度

(1).平均数

- 计算方法

  所有数字相加，除以个数。

- 使用场景

  在数据非常对称，且仅显示出一种趋势时使用。

(2).中位数

- 计算方法

  将所有数据按照升序顺序进行排列。如果有奇数个数值，则中位数为中间的数值；如果是偶数个数值，则中位数为两个中间的数值加和除以2得到的结果。

- 使用场景

  在数据由于异常值而发生偏斜时使用。

(3).众数

- 计算方法

  选出具有最大频数的一个或几个数值。如果数据可分为多组，则为每组找出一个众数。

- 使用场景

  众数是唯一能用于类别数据的平均数类型。当数据可分为两组或者多组时使用。

2.分散性和变异性的量度

(1).全距(也叫极差)

全距指出数据的扩展范围。用于度量数据集分散程度的一种方法。

- 计算方法

  用数据集中的最大值(上届)减去最小值(下届)。

- 缺点

  全距仅仅描述了数据的宽度，因此不能指出数据的真实形态以及数据是否包含异常值。

(2).四分位数

四分位数是这样一些数值，它们将数据一分为四(一共三个数)。最小的四分位数为"下四分位数"，最大的四分位数为"上四分位数"。中间的四分位数即中位数。

- 计算方法(下四分位数)

  首先计算(n/4)，如果结果是整数，则下四分位数位于(n/4)这个位置和下一个位置中间，取这两个位置上数值的平均值，即下四分位数。如果(n/4)不是整数，则向上取整，得到的结果就是下四分位数。

- 计算方法(上四分位数)

  首先计算(3n/4)，如果结果是整数，则上四分位数位于(3n/4)这个位置和下一个位置中间，取这两个位置上数值的平均值，即上四分位数。如果(3n/4)不是整数，则向上取整，得到的结果就是上四分位数。

(3).四分位距

一种不易受异常值影响的"迷你距"。四分位距是50%的中间数值形成的一个范围。

- 计算方法

  上四分位数 - 下四分位数

(4).百分位数

第k百分位数，即位于数据范围k%处的数值。

表现方式 - 箱线图(箱形图)

在同一张图上体现多个距和四分位数的一种方法。"箱"显示出四分位数和四分位距的位置，"线"则显示出上、下届。箱线图能在同一张图上体现多批数据，因此非常有利于比较。

(5).方差

方差是量度数据分散性的一种方法。

- 计算方法

  先计算均值，然后计算"数值与均值距离的平方数"，最后取平均值。

  *平方解决了数值与均值差异的负数问题。但也随之带来的问题是不太直观，因此引入了"标准差"的概念。

(6).标准差

标准差是描述典型值与均值距离的一种方法。标准差越小，数值离均值越近。

- 计算方法

  先计算方差，然后取平方根。

(7).标准分

标准分是将几个数据集转换为一个理论上的新分布。这个分布的均值为0，标准差为1，这是一种可用于进行比较的通用分布。标准分将你的数据有效地转化为符合这个模型的数据，同时确保数据的基本形状不变。标准分可取任意值，这些值表示相对于均值的位置。正的表示高于均值，负的表示低于均值。数据大小体现了数值与均值的距离。

- 计算方法

  通过整个数据集的均值和标准差，可求出一个特定数值的标准分。

  z = (特定数值 - 均值) / 标准差

3.概率计算

(1).概率

概率，是量度某事发生几率的一种数量指标。

- 计算方法

  发生某事件的可能数目 / 所有可能的结果

概率/样本空间

表示所有可能结果的一种简便表示法。

条件概率

量度与其他事件的发生情况有关的某个事件的概率。

- 贝叶斯定理

  该定理提供了一种计算逆条件概率的方法。在无法预知每种概率的情况下，十分有用。

4.离散概率分布的运用

5.排列与组合

(1).排位

- 计算方法(排位)

如果要求n个对象的可能排位方式的数目，则计算n!。

- 计算方法(圆形排位)

如果要求n个对象做圆形排位，其可能的排位方式的数目为(n-1)!。

- 计算方法(按类型排位)

如果要对n个对象排位，其中包括第一类对象k个，第二类对象j个，第三类对象m个...，则排位方式数据的计算式为n!/(k!j!m!)

6.几何分布、二项分布、泊松分布

(2).排列

排列，是指从一个较大(n个)对象群体中取出一定数目(r个)对象进行排序，并得出排序方式总数目。

对比 - 排列 vs 组合

- 排列，是指从一个群体中选取几个对象，在考虑这几个对象的顺序的情况下，求出这几个对象的选取方式的数目。排列，与顺序有关。
- 组合，是指从一个群体中选取几个对象，在不考虑着几个对象的顺序的情况下，求出这几个对象的选取方式的数目。在不需要知道每个位置的确切占位情况时，组合是比排列更通用的算法。组合，与顺序无关。

计算方法

- 计算方法(排列)

  如果从n个对象中选取r个对象，则排列数目为 n!/(n-r)!
- 计算方法(组合)

  如果从n个对象中选取r个对象，则组合数目为n!/r!(n-r)!

6.几何分布、二项分布、泊松分布

(1).几何分布

进行多次相互独立的试验时可使用几何分布，每一次试验都存在成功或失败的可能，而你感兴趣的是为了取得第一次成功需要试验多少次。

计算方法

P(X=r) = q^(r-1)p

*p代表单次成功的概率，q为失败的概率1-p，上述表示第r次成功的概率。这个公式叫做概率的几何分布。

分布特征

- 随着r的增大，P(X=r)逐渐下降。
- 任何几何分布的众数都永远是1。因为1是最具有最大概率的数。

扩展 - 不等式

$P(X>r) = q^r$

*上述公式指的是为了取得第一次成功需要试验r次以上的概率。为了让需要进行的试验次数大于r，意味着前r次试验必须以失败告终。

$P(X<=r) = 1 - q^r$

*为了取得一次成功而需要尝试r次或r次以下的概率。

简明表达式

$X \sim Geo(p)$

*如果一个变量X的概率符合符合分布，且单次试验的成功概率为p，则可用上述公式表达。

期望公式

$E(X) = \sum xP(X \leqslant x) = 1/p$

*随着X增加，累计总和越来越接近于一个特定的值。例如，单次试验的成功概率为0.2，可理解为5次尝试中有一次尝试趋向于成功，即可期望尝试5次即获得成功。

方差公式

$Var(X) = q/p^2$

(2).二项分布

进行次数固定的独立试验，每一次试验都存在成功和失败的可能时可使用二项分布，你感兴趣的是成功或失败的次数。

计算方法

$P(X=r)=n!/r!(n-r)! * p^r * q^(n-r)$

*每次答对概率为p，而每道题答错概率为1-p，也就是q。答对n个问题中的r个问题的概率上述公式。这类问题，称为二项分布。

简明表达式

$X \sim B(n,p)$

分布特征

根据n与p的不同数值，二项分布的形状会发生变化，p越接近0.5，图形越对称。一般情况下，当p小于0.5时，图形向右偏斜；当p大于0.5时，图形向左偏斜。

期望公式

$E(X) = np$

方差公式

$Var(X) = npq$

区别 - 几何分布vs二项分布

两者有共同之处，处理的都是独立试验，每次试验要么成功、要么失败。差别在于实际要求的结果。

- 如果试验次数固定，求成功一定次数的概率，则需要使用二项分布；使用二项分布还可以求出在n次试验中能够期望取得的成功次数。
- 如果感兴趣的是在取得第一次成功之前需要试验多少次，则需要使用几何分布。

(3).泊松分布

在遇到独立事件时，若已知λ(即给定时间区间内的事件平均发生次数)且感兴趣的事一个特定区间内发生次数，这是可使用泊松分布。

简明表达式

X ～Po(λ)

用X表示给定区间内的事件发生次数。如果X符合泊松分布，且每个区间内平均发生λ次。

计算方法

P(X=r) = e^(-λ)λ^r/r!

期望

E(X) = λ

方差

Var(X) = λ

分布特征

- 泊松分布的形状随着λ的数值发生变化。λ小，则分布向右偏斜，随着λ变大，分布逐渐变得对称。
- 如果λ是一个整数，则有两个众数，λ和λ-1。如果λ不是整数，则众数为λ。

组合泊松变量

如果X和Y都符合泊松分布，则X+Y也符合泊松分布。也就是说，可利用X和Y的分布情况求出X+Y的概率。

伪装泊松分布

泊松分布的一个用途，可在特定的条件下近似代替二项分布。当n很大且p很小时，可以用X～Po(np)近似代替X ～B(n,p)。

7.正态分布 - 连续型概率分布

(1).概述

离散概率vs连续概率

对于离散概率分布来说，关心的是取得一个特定数值的概率；而对于连续概率分布来说，关心的是取得一个特定范围的概率。

概率密度函数

- 描述连续随机变量的概率分布。通过它可以求出一个数据范围内的某个连续变量的概率，它向我们指出该概率分布的形状。
- 概率密度函数下方总面积必须等于1。

计算方式 - 面积

连续随机变量的概率通过面积表示。位于函数图形下方且介于这个特定数值范围之间的面积，就是这个特定数值范围的概率。不能通过把数值范围内的每一个数值的概率相加得出连续概率分布的概率，原因是数值个数无穷无尽。唯一方法就是算出由连续概率函数形成的曲线下方的面积。

(2).正态分布(也叫高斯分布)

正态分布具有钟形曲线，曲线对称，中央部分的概率密度最大。越是偏离均值，概率密度减小。均值和中位数均位于中央，具有最大概率密度。

表示方法

X～N(μ,σ^2)

*μ指出曲线的中央位置，σ指出分散性。如果一个连续随机变量X符合均值为μ，标准差为σ的正态分布，则可用上面方法描述。

图形特点

- μ为均值，σ为标准差(σ^2即为方差)
- σ^2越大，正态曲线越扁平、越宽。
- 无论图形画多大，概率密度永远不会等于0。可理解为事件越来越不可能发生，但微小的发生机会却永远存在。

(3).标准分

对概率分布进行标准化，从而令X ～N(μ,σ^2)变为Z～N(0,1)。我们要做的是为需要求概率的数值找出数值范围，然后求出这个范围的限值得标准分，最后通过正态分布表查找求得标准分的概率。一个特定数值的标准分还说明了数值与均值相距多少个标准差，可由此获悉该数值与均值的相对接近程度。

原理

对原来的正态分布进行标准化时，一切比例保持相同。整个区间既没有增大，也没有缩小。由于代表概率的是面积，因此概率也保持不变。

计算公式

z = (x - μ)/σ

(4).组合正态分布

X、Y为独立变量，即它们相互之间对对方的概率没有影响。

X～N(μ,σ^2)，Y ～N(μ,σ^2)

=>

X-Y～N(μ,σ^2)

*μ=μx-μy σ^2=σx^2+σy^2

(5).正态分布代替二项分布

可以使用正态分布近似计算二项分布。但需要注意，需要进行连续性修正。

如何选择：正态分布or泊松分布作为二项分布的近似

- 如果X～B(n,p)，当np>5且nq>5时，则使用正态分布近似代替二项分布。
- 如果n>50且p<0.1时，则使用泊松分布近似代替二项分布。

8.统计抽样的运用 - 抽取样本

(1).如何设计样本

样本的作用是用它来判定总体情况。为了确保得到正确结果，需要明智地选择样本，以便让样本尽量具有代表性。

确定目标总体

这里的目标总体指的是正在研究、并且打算为其采集结果得群体。目标总体要尽可能精确，这样能更为容易地得出尽可能代表总体的样本。

确定抽样单位

决定要抽取哪一类对象。通常，要抽样的对象类型就是在确定目标总体时所描述的对象类型。

确定抽样空间

列一张表，表中列出目标总体范围内的所有抽样单位。这张表就被称为抽样空间。

(2).样本偏倚

无偏样本

可以代表目标总体，即该样本与总体样本具有相似特性，可以利用这些相似特性对总体本身做出判断。

偏倚样本

无法达标目标总体，由于样本与总体的特性不相似，无法根据样本对总体做出判断。

(3).如何选择样本

简单随机抽样

做法

- 重复抽样

在选取一个抽样单位并记录下这样抽样单位的信息后，再讲这个单位放回总体中。这样做的结果是某个抽样单位有可能被选取不止一次。

- 不重复抽样

不再将抽样单位放回总体。

方法

- 抽签
- 随机编号

分层抽样

将总体分割为几个相似的组，每个组具有类似的特性。这次组被称为层。对每一层进行简单随机抽样，确保最终样本中具有每个组的代表。

整群抽样

将总体划分为几个群，其中每个群都尽量与其他群相似，可通过简单随机抽样抽取几个群，然后用这些群中的每一个抽样单位形成样本。

系统抽样

按照某种顺序列出总体名单，然后每k个单位进行一次调查。这种方式有个缺陷，如果总体中存在某种循环模式，则样本将会有偏倚。

(4).评估

评估总体均值

- 点估计量，根据样本数据得出的对你所认为的总体均值的最佳猜测值。一般说来，样本越大，点估计量越准确。
- 样本均值，被称为总体均值的点估计量。

评估总体方差

用样本方差评估总体方差，评估结果会偏低。差别程度取决于样本数值的大小。一般通过另一种方法评估总体方差，即取样本中的每一个数值，减去样本均值，所得之差取平方数；然后将所有平方值加起来，除以样本数减1。除以n-1比除以n能得出精确性稍微高一点的结果。因为样本数值的方差很可能略小于总体方差。

预测总体比例

用样本成功比例作为总体成功比例的点估计量。

9.置信区间

(1).概念

(2).构建步骤

- 首先选择用于构建置信区间的总体统计量。
- 接着求其抽样分布。
- 随后确定用于构建置信区间的置信水平。
- 最后必须求出置信区间的置信上下限。

10.假设检验与X2分布

(1).假设检验

粗略步骤

- 确定要进行检验的假设
- 选择检验统计量
- 确定用于做决策的拒绝域
- 求出检验统计量的p值
- 查看样本结果是否位于拒绝域内
- 作出决策

统计量 - 显著性水平

显著性水平，表明你希望在观察结果得不可能程度达到多大时拒绝原假设。

(2).X2分布

X2提供了一种观察频数和期望频数之间的差异进行度量的办法。X2的数值越小，观察频数和期望频数之间的总差值越小。

主要用途

- 检验拟合优度

  也就是检验一组给定的数据与指定分布的吻合程度。
- 检验变量独立性

  检验两个变量的独立性。也就是说检查变量之间是否存在某种关联。

11.相关与回归

(1).最佳拟合

最佳拟合线，即能最准确预测出所有点真实值的线。

计算方法 - 最小二乘回归法

最小二乘回归法，是一种数学方法，可用一条最佳拟合线将一组二变量数据拟合，通过将公式为y=a+bx的

一条直线与一组数值相拟合，使得误差平方和最小。

回归线

直线y=a+bx称为回归线。

使用限制

线性回归法，只是根据已有信息的估算，它体现了各个数据点之间的关系，这并不代表它也适用于数据限值以外的范围。

相关系数

反映直线拟合度的一个指标。是介于-1和1之间的一个数，描述了各个数据点与直线的偏离程度，通常用字母r来表示。如果r=-1，表示数据为完全负线性相关；如果r=1，表示数据为完全正线性相关；如果r=0，则不存在相关性。