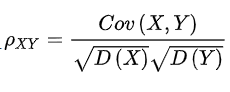
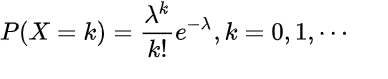
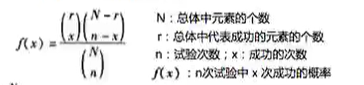
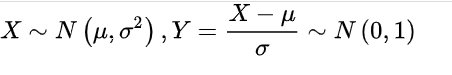
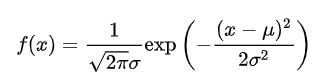
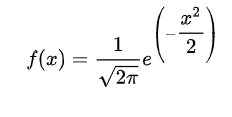
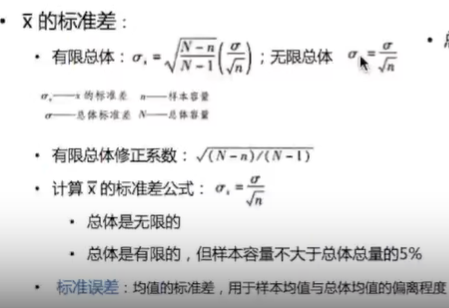
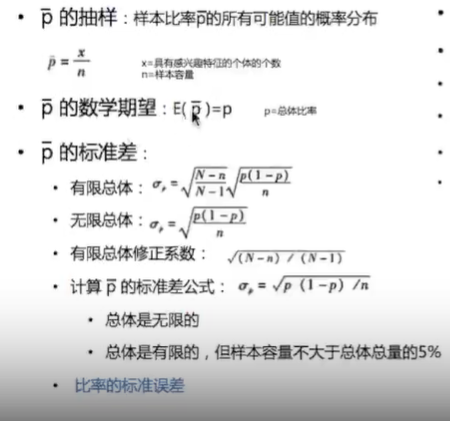
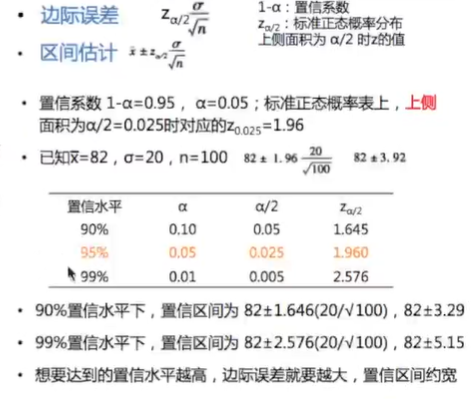
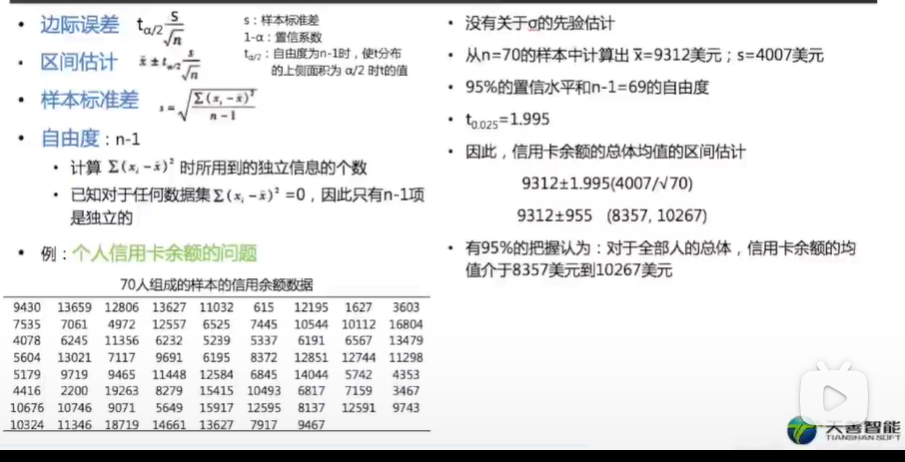
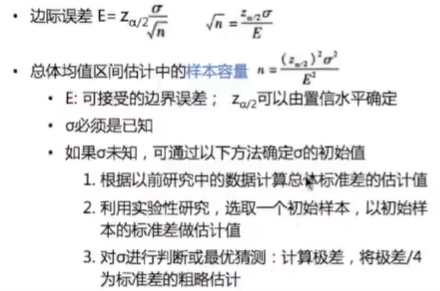
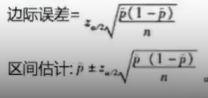
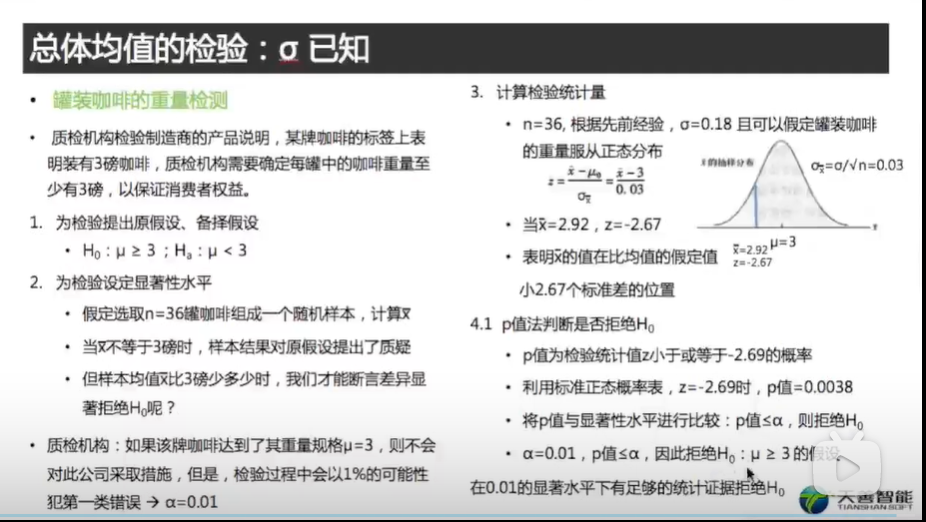
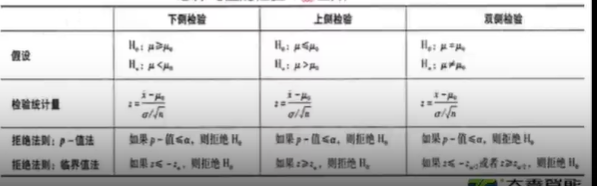
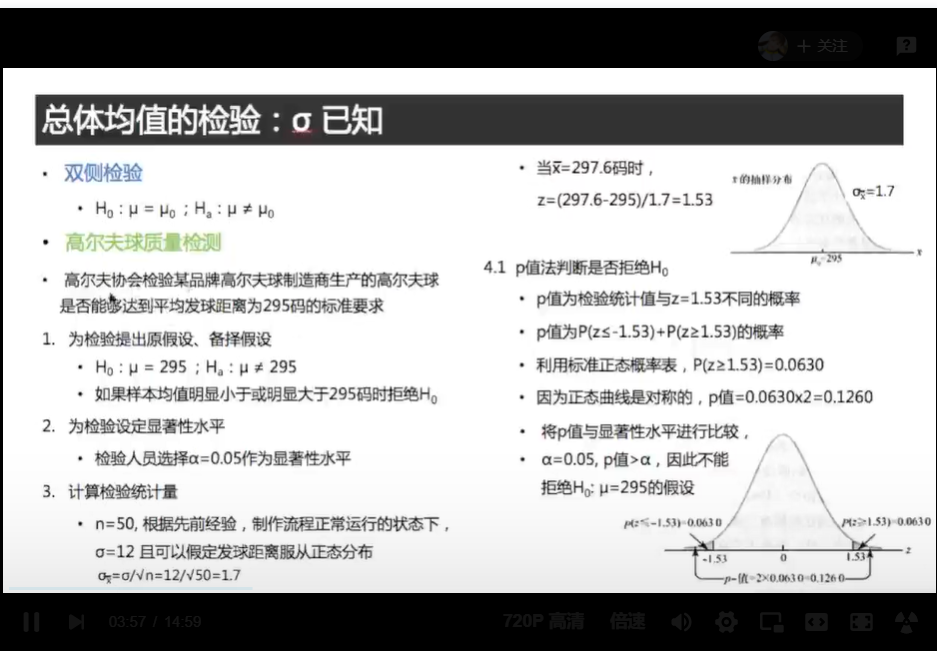
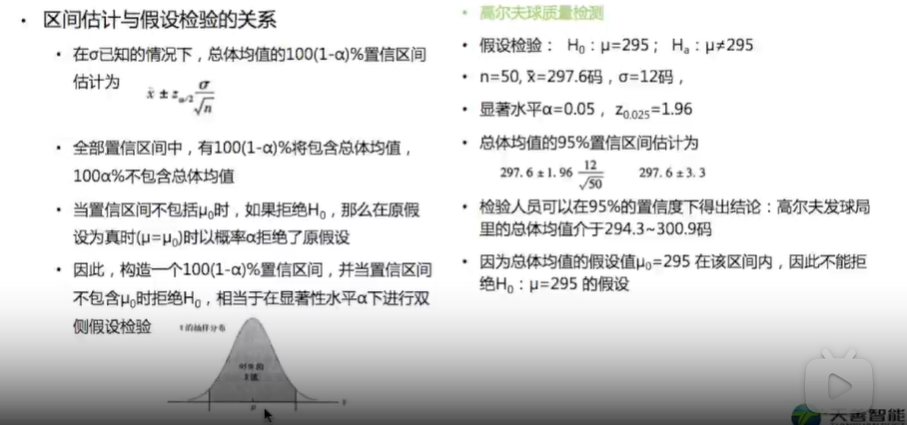
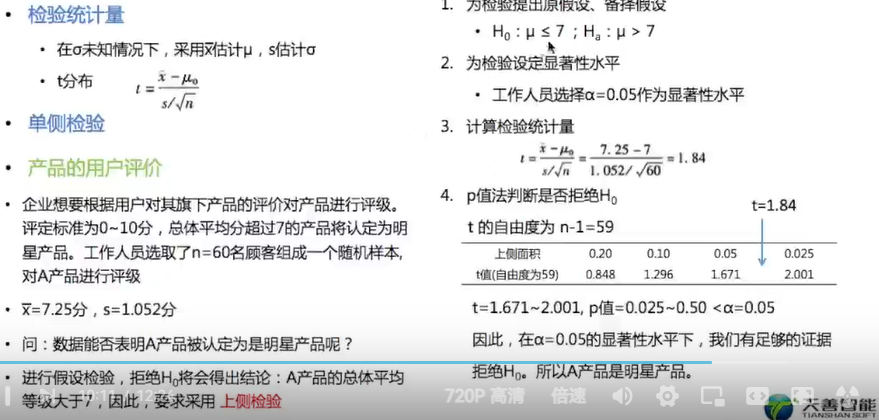
1. **概念（一个变量）：**
   1. **极差=最大值-最小值。标准差系数=标准差/平均数\*100%。**
   2. **平均数>中位数，偏度为正；平均数<中位数，偏度为负。**
   3. **衡量相对位置-z分数：(x-平均数)/标准差。**
   4. **切比雪夫定理：**任意一个数据集中，落在其平均数±m个标准差范围内的数据的占比至少为1－1/m2
   5. **对正态分布，68%的数据落在与平均数相差1个标准差之内，95%的数据落在与平均数相差1个标准差之内。**
   6. **极端值：方法1：z分数绝对值大于3，异常值。方法2：小于第一四分位数-1.5\*（三分位数-一分位数）或大于第三四分位数+1.5（三分位数-一分位数）。**
2. **概念（两个变量）：**
   1. **协方差：cov(X,Y)=E[(X-E[X])(Y-E[Y])]=E(XY)-E[X]E[Y]。期望：样本协方差为sum((X-E[X])(Y-E[Y]))/(N-1)，整体协方差为sum((X-E[X])(Y-E[Y]))/N。协方差大于0，正相关；协方差小于0，负相关；协方差等于0，相互独立。**
   2. **协方差与方差：D(X+Y)=D(X)+D(Y)+2cov(X,Y)；D(X-Y)=D(X)+D(Y)-2cov(X,Y)**
   3. **皮尔森相关系数：**
   4. **ρxy=1，完全正相关(Y=aX+b)；ρxy=-1，完全负相关；ρxy=0，完全不相关**
3. **离散随机变量：二元分布概率表，二元分布期望E(X)=np(次数\*成功概率)，方差Var(X)=np(1-p)**
4. **泊松概率分布：**
   1. **已知某事的平均发生概率，求在特定时间或空间中某事件发生的次数的预估。当二项分布的n很大（>20），p很小(<0.05)时，泊松分布可作为二项分布的近似。**
   2. **泊松分布的平均值和方差均为**λ，即这段时间内事件平均发生次数，λ=np。
5. **超几何概率分布：每次实验概率不同，**

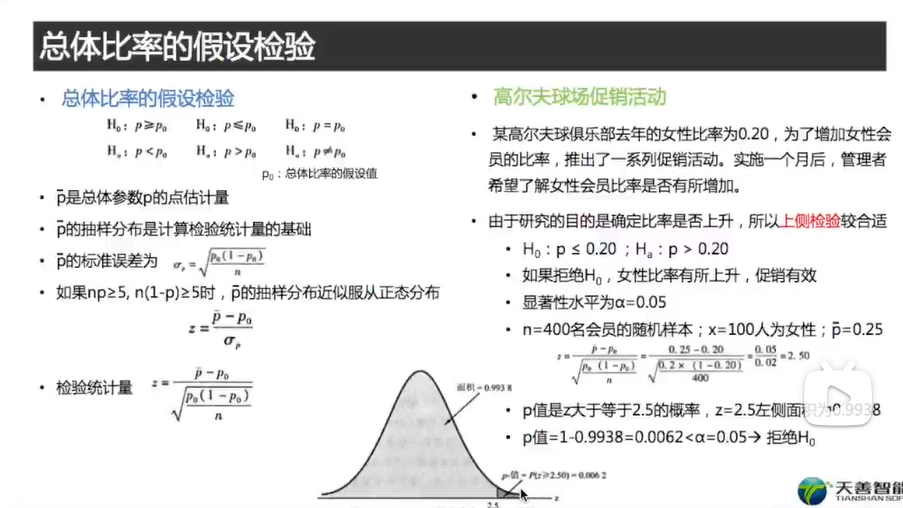
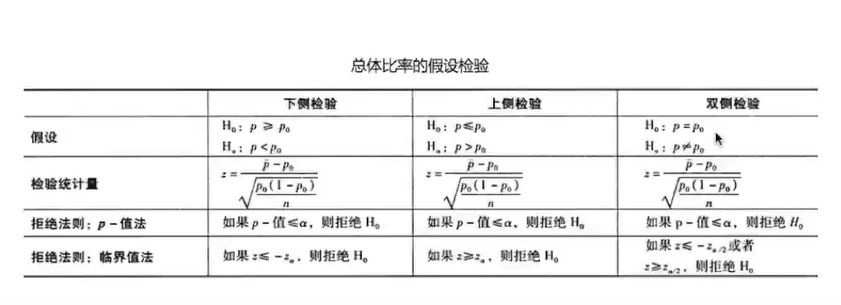


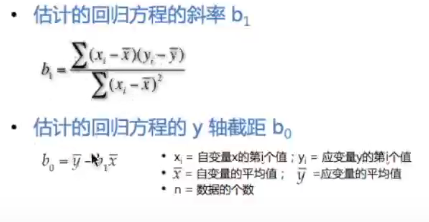
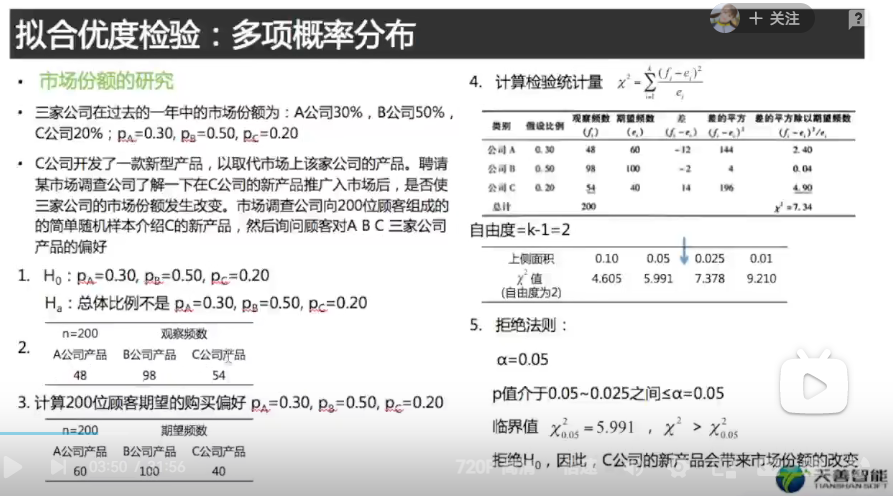
1. **正态分布：**
   1. **又称常态分布、高斯分布。**N(μ，σ2)，代表期望、方差。μ=0，ρ=1时称标准正态分布。
   2. **转化为标准正态分布：**
   3. **一维正态分布：**
   4. **标准正态分布：**
   5. **正态随机变量有68.3%的概率在均值±1个标准差的范围内；95.4%的概率在均值±2个标准差的范围内；99.7%的概率在均值±3个标准差的范围内。**
2. **抽样分布：**样本均值的抽样分布即所有样本均值的可能取值形成的概率分布。
   1. **E(样本均值)=整体均值**
   2. 
   3. **中心极限定理：从总体中抽取n个简单随机样本，当n(>30)很大时，样本均值的抽样分布近似服从正态概率分布。**
   4. **样本比率的抽样分布：**
   5. **np>5且n(1-p)，则p的分布类似正态分布**
3. **区间估计**
   1. **点估计量：样本均值【x拔】是整体均值μ的点估计量**
   2. **区间估计：总体均值的区间估计=点估计±边际误差（x拔±置信区间对应的正负标准差个数\*样本标准差）——样本标准差=整体标准差/根号下样本数，样本容量大于30即可**
   3. **例：标准差已知，z检验，用样本均值、总体标准差、算得的样本标准差，来估计整体均值。**
      1. **顾客消费，历史标准差=20美元，现在取100个抽样，发现x拔=82元。求历史平均值95%的区间估计。**
      2. **解：抽样标准差=历史标准差/(根号下n)=20/10=2。μ低=82-抽样标准差\*1.96（正态分布95%的概率落在μ正负1.96个标准差以内）=78.08；μ高=82+2\*1.96=85.92。**
      3. **答：历史消费平均值有95%的概率在78.08元-85.92元之间。**
      4. **Z分数=(x拔-μ)/样本方差**



* 1. **总体标准差未知：t检验，即用样本均值、样本标准差来估计总体标准差和总体平均值。**
     1. **T检验：区间估计=点估计±边际误差（两个参数都需要算）**
     2. **自由度越小，t曲线越平坦；自由度越大，t曲线越接近正态分布，即z分布；t>100，则非常类似**
  2. 样本数量的估计：
  3. 总体比例(p)区间估计：
  4. 
  5. 总体比例的样本数量的估计：区间估计把n挪到一边

1. **区间检验方式**
   1. **Z检验：面向大样本数据（>30）、标准差已知。先转化为z分数序列(个体数据-全部均值)/全部标准差，转化为正态的标准z序列。常见场景：两样本均值是否差异显著、一样本均值与某常数是否差异显著。**
   2. **T检验：面向小规模抽样，或标准差未知。实际是从正态中抽取少部分样本来实验。(抽样个体数据-抽样均值)/抽样标准差。**
2. **假设检验：**
   1. **原假设H0（出现几率大，一般是预想结果的反面也就是不希望成立的）、备择假设H1（与原假设完全对立，出现几率小，一般是预想结果也就是希望成立的）**
   2. **单侧检验（H0：μ>=μ0，H1：μ<μ0）、双侧检验（H0：μ=μ0，H1：μ≠μ0）**
   3. **两类错误：**
      1. **第一类：原假设为真，但被拒绝了**
      2. **第二类：原假设为假，但被接受了**
   4. **显著性水平：犯第一类错误的概率，用α表示。1-α为置信度或置信水平。**
      1. **一般第二类错误的概率很难控制，为了避免第二类错误，要从第一类错误的水平入手。**
   5. **例1：总体均值的检验，总体标准差已知，单侧检验。**
   6. **例2：总体均值的检验，总体标准差已知，双侧检验**
   7. **置信区间与假设检验：**
   8. **例3：总体均值未知**
   9. **总体比率的假设检验：**



1. **卡方检验：**
   1. **统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为0，表明理论值完全符合。**
2. **线性回归**
   1. **简单线性回归方程vs估计的简单线性回归方程**
   2. **最小二乘法：**
   3. 
   4. **残差：估计值和观测值之间的差**
   5. **是否很好地拟合了？拟合优度的度量：**
      1. **判定系数：r^2=SSR/SST**
      2. **SSR：回归平方和，估计值和平均数的平方和。**
      3. **SST：总平方和，观测值和平均数的平方和。**
      4. **SSE：残差平方和（误差平方和），估计值和观测值的平方和。**
      5. **SST=SSR+SSE**
      6. **相关系数r，介于-1~1之间。值为根号下r^2，符号同斜率的符号。**
3. **多元线性回归**
   1. **系数用软件算**
   2. **如果有分类变量，则用x=0或1表示，体现为截距的变化。**
4. **时间序列分析**
   1. **平均绝对误差MAE：误差的绝对值的平均值。**
   2. **均方误差MSE：绝对误差的平方和。**
   3. **平均绝对百分比误差MAPE：(误差绝对值/实际值)\*100%的平均数**
   4. **朴素预测法：跟上次的值一样**
   5. **过去数值平局法：历史数据的平均值作为下个日期的预测值**
   6. **趋势推测法：线性趋势回归、Holt线性指数平滑、非线性趋势回归**
5. **拟合优度检验：**
   1. **多项概率分布：**
   2. 
6. **实验设计&方差分析**
   1. **完全随机化实验：例，某公司想比较三种生产方法的效率。**
   2. **因子：生产方法**
   3. **处理：不同的生产方法A、B、C**
   4. **单因子实验：仅涉及一种方法的实验。**
   5. **响应变量：这种方法生产出来的数量**
   6. **实验单元：选出来操作实验的工人**
7. **因果分析：**
   1. **ABtest：完整AB实验很麻烦，数据要求高，时间久。**
   2. **观察性研究：省时间，误差大。自然状态，按特征分组，观察研究。无法控制样本随机性，也无法排除其他变量影响。可以给出参考，而非论证，容易出现辛普森悖论。**
   3. **倾向性得分匹配（propensity score matching）：匹配（一组受到干预A1，另一组没受到干预A2）。对于每个A1中的样本，都在A2中找一个跟他匹配，A1和A2只有目标因素是不同的(exact matching)。这种matching很难找，随着变量维度数增加，难度几何级数上升。所以选择倾向性得分，计算倾向性得分值，一般用逻辑回归（操作简单，结果易理解），用是否受到干预的概率当做倾向性得分。用最邻近匹配法对得分结果做匹配，将实验组和对照组评分最接近的个体匹配**
   4. **双重差分法：第一个差分：同一地区，推广活动前后的差分。第二个差分：另一个非常相似的地区，未举办推广活动，无影响。前提条件：两个地区本身特点相似、趋势相似。**
8. **辛普森悖论：**分组比较中都占优势的一方，在总评中有时反而是[失势](https://baike.baidu.com/item/%E5%A4%B1%E5%8A%BF/7439867)的一方
   1. **总数：**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **人数** | **转化率** | **转化人数** |
| **30岁以上** | **1000** | **3.00%** | **30** |
| **30岁以下** | **3000** | **4.00%** | **120** |
| **总计** | **4000** | **3.75%** | **150** |

* 1. **拆分(1.50%>1.20%,9.00%>4.56%)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **人数** | **转化率** | **转化人数** |
| **30岁以上男性** | **800** | **1.50%** | **12** |
| **30岁以上女性** | **200** | **9.00%** | **18** |
| **30岁以下男性** | **500** | **1.20%** | **6** |
| **30岁以下女性** | **2500** | **4.56%** | **114** |

* 1. **原因**

样本不均匀，小样本\*高转化率+大样本\*低转化率<大样本\*中转化率+小样本\*低转化率。

质\*量，分组评价只考虑了质，没考虑量，加总以后可能被反超。

启示：分组时数量要相近，每组特征要均匀。