上下文：客户做了线下问卷调查，把针对其中女性用户累积N天的游戏消费金额作为一个随机变量，需要关注三个和统计相关的问题：

1. **给定置信度比如95%的情况下，如何计算消费额平均值统计量（注意这里是消费额平均值的统计，不是针对消费额统计）的置信区间**。

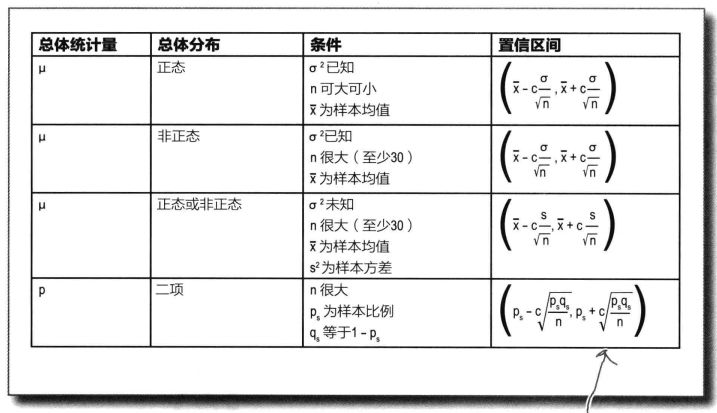
**关于随机变量的分布以及抽样的随机变量的均值的分布是两个概念，比如随机变量A（个人的收入）可以是zipf分布，但是该随机变量A抽样后的均值的分布就是符合正态分布（采样的数量大于30）或者符合t分布（采样的数量小于30）**。

**（为什么是**30，这个可能是经验值，文章中都这样说！）

现在可以利用给定的置信度来计算当前这轮采样的置信区间了（注意：每次采样都会得到不同的置信区间）。简单理解，置信度的含义就是说100回的采样（举个例子，每回的采样都会采样比如300个样本），得到了100个置信区间，那么有大概95个置信区间会包含该随机变量A的真实的均值。而关于这个置信区间如何计算的问题，可以参考（**强烈推荐阅读，写的很清晰**）：<https://zhuanlan.zhihu.com/p/36206276> ，对应的图如下：



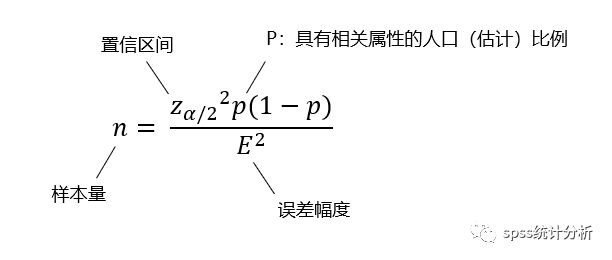
**注意：上图中有笔误，不是计算置信水平/置信度，而是计算置信区间**。样本数量的不同以及观察的随机变量本身的分布以及标准差是否已知的条件下，计算方法都不同，如下图：



对于客户这个场景，需要使用上图中的第三行，也就是标准差未知，总体分布是非正态分布的情况。

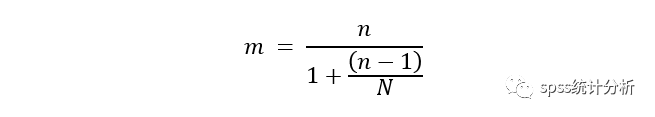
1. **给定置信度的情况下，如何计算最少需要的样本量？**

参考link（**写的很清楚，推荐阅读**）：http://www.360doc.com/content/19/0802/10/50903246\_852532871.shtml ，参考下图：



上图中的Za/2是给定的置信度（同样上面是笔误）对应的Z分数，p是采样出来的人口比例（这里专门针对的就是调查问卷），E是可以接受的人为需要输入的误差比例比如5%。

另外，如果研究的总体样本空间不大，可以对上面的结果进一步修正，如下：



N就是研究的总体样本空间大小。

1. **对于女性消费额这个随机变量，如何求该分布的累积分布值或者互补累计分布值？**

**两种方案可以选择**：

1. Python fitter + python scipy的组合方式**:**

**Fitter库用来同时拟合多个可选择的分布**，并最后选择一个拟合效果最好的分布；得到这个分布以后，利用scipy对应概率分布的CDF API简单就可以得出结果了。

1. 单独一个Python powerlaw库：

对于给定的一个zipf分布或者幂律分布，计算该随机变量的累积分布CDF值P(x<a)或者互补累积分布CCDF值P(x>a)的一个简单的python库 powerlaw（简单pip install就可以使用），可以参考link（**针对随机变量是离散的或者连续的都可以处理**）：<https://www.163.com/dy/article/GD21QOLF0511D05M.html>