转写正确率的计算和估计

# 样本量的确定介绍：

在统计推断中，样本量的大小直接影响统计推断的精度。在现实中，样本量大小的选取需要在满足统计推断精度的情况下，尽可能少地减少需要收集的观测数据从而节约的成本。因此，本文中样本容量估算目的为：**在确保结局具有一定可靠性条件下，需要的最小样本容量。**

首先要说明在什么情况下，可以认为结局具有一定的可靠性：

1. 目标方差小：目标方差也就是 统计量估计值 的方差。统计量估计值由抽取的样本决定，而样本的抽取存在随机性，因此不同次的抽样将会得出不同的估计值。在目标方差足够小的情况下，可以认为不同次抽样的结果差距较小。在这样的情况下就保证了估计值的可靠性。
2. **在相同的置信度下，有较小的置信区间。**
3. 假设检验固定第一类错误概率的条件下，犯第二类错误的概率较小。

在所研究的语音转述正确性问题中，语音数据库大小都在数十万，因此样本较容易获得。同时转述正确率为样本均值，因此在样本量较大时，由中心极限定律，转述正确率的估计近似服从正态分布。因此上述三条可靠性评价在正太性假设下近乎等价。

# 样本量的确定：

从3614个台语语音采样中，得到的转述正确率方差为136，其95%置信区间为[127.7532, 144.9352] [3] . 另外在5528总体的描述性分析中，方差估计为126.5 [4]

* 计算式：

变量定义：

n：样本容量。

：样本标准差。

：容许误差（在本问题中可取为1）。

：统计量分布的上 分位数。

* 加入总体数量进行修正：（已知方差）

其中为样本量，为正态分布上分位数，为总体方差，为绝对误差，N为总体个数。

未确定样本量而使用的方差估计：从3614个台语语音采样中，得到的转述正确率方差为136，其95%置信区间为[127.7532, 144.9352] [3] . 另外在5528总体的描述性分析中，方差估计为126.5 [4].

由于样本量的选择较多依据实际情况和抽样方案而定. 过多地抽样会在一定程度上增大成本的同时也能提高准确性，而不足的抽样在试验准确度不足的情况下，也可再次抽样扩大样本量. 在统计学中，抽样一般选取可能的最大样本量. 因此在最初抽样的样本量设计中，可以通过 代入上式进行样本量的确定.

抽样方法

进行分层抽样，同时在不同的层中进行简单随机抽样。

在确定总的样本量下，以如下方法抽取各分层中样本并计算各层中均值。

* 分层抽样最小方差方法：

在分层抽样中，不同的分层中数据有着不同的方差。在方差较大的层中，选取较多的数据计算该层中的均值是降低方差的有效方法。

如果有M个分层，每个分层中选取的样本量是, m=1, 2, 3,..., M, 有.

在此使用Neyman's 最优分配[2] 可使分层抽样错误率估计方差最小化：

其中： ; 为m 层中的元素个数； 为该层均值估计.

各层抽取样本量计算：

为m 层标准差，K 为常数使得上两式同时成立。

统计推断

# 转写正确率均值

1. **均值统计量的分布：**

在的正态分布,方差未知时，可构造相应的t 统计量（此处缺乏正太性假设）:

当样本量n 充分大时，自由度为n-1 的t 分布将近似服从正态分布。

在方差已知时，由中心极限定理可知：

在大样本时， 近似于

[5]

1. **置信区间估计**

有 的几率处于置信区间

1. **假设检验**

在实际中，为了检验转写正确率是否大于某一必要的下限值, h, 将进行假设检验:

原假设与备择假设如下：

*：；:*

为实际的转写正确率， 为转写正确率的下限值。

P值的计算：

对于给定的显著性水平下

|  |  |
| --- | --- |
|  | 接受原假设 |
|  | 拒绝原假设 |

# 转写正确率方差 [3]

1. **方差估计的统计量的分布：**

转写正确率需要有正太假设（待讨论）：

也可利用如下统计量估计

依据

1. **转写正确率方差的置信区间**

的置信区间为(有正态假设)：

的置信区间为(无正态假设)：

(宽松的估计，取相关系数为1)

二次抽样数

首次抽样数

上界方差估计所需样本量

下界方差估计所需样本量

已抽取样本量

样本量估计

真实方差所需的样本量

二次抽样数

首次抽样数

下界方差估计所需样本量

已抽取样本量

样本量估计

真实方差所需的样本量

方差上方估计所需样本量

时间

样本量估计

真实方差所需的样本量

1. **转写正确率的运用**

**工作质量的检查指标**

对一个或者一群员工在一段时间里工作质量的检查，可以选取他们在一段时间内的语音转写作为调查对象。进行抽样，然后对他们的转写正确率进行估计，同时可以运用假设检验来判断他们的转写正确率是否高于某个标准。

每次抽查的样本数的确定，可由之前抽取的所有样本的方差作为计算抽样语句数的依据。

**可以得到的结果：**

1. 每个人当天的转写正确率及区间估计都能得到相应的计算。
2. 当天转写正确率的方差及其区间估计

此区间估计的上届可以用为下一次抽样数量的依据

1. 假设检验转写正确率是否高于某一个数值（可以设定为公司的正确率最低要求）
2. https://wenku.baidu.com/view/0ebf328d49649b6648d7479d.html
3. https://en.wikipedia.org/wiki/Sample\_size\_determination#Estimation\_of\_a\_mean
4. 转写正确率检验分析报告（from 前任工作者）
5. 样本量确定的实验（from 前任工作者）
6. http://blog.csdn.net/hyacinth\_yuan009/article/details/77599882