如何选择抽样大小（Sample Size）

Accela Zhao 2020/02/13

有测量才有改进，今天的软件开发离不开数据驱动。分析设计需要以数据为基础，上线效果需要用数据来衡量，安全稳定需要数据来监控，预期数据与实际的缺口预言改进的方向。

然而产线数据往往规模大、流速快，采集过多数据影响性能，长期存储占用空间。退而求其次，抽样是个常用的方法。

多大的样本集能够代表原始数据呢？这却不易回答。既然是用样本集代替原始数据进行统计，那么

- **样本集的统计特征（平均值、方差、百分位数等）应该与原始数据足够接近**

下面，我们来逐项研究，目录……

样本集的平均值（Sampling Distribution of Sample Mean）

先找样本集最简单的统计特征——平均值（Mean）——来研究。

首先设定符号表示，

- X：用随机变量 X 表示原始数据
- σ：表示 X 的标准差，即原始数据的标准差
- S：样本集 S，大小为 n
- X_i：第 i 个抽样用 X_i 表示，$S = \{X_0, X_1, .. , X_i, .. X_{n-1}\}$
- M_s：表示样本集的平均值，$M_s = (X_0 + X_1 + .. + X_i + X_{n-1}) / n$

另外，一些常见的公式（更多公式[25]），

- E(Y)：表示对随机变量 Y 求期望
- Var(Y)：表示对随机变量 Y 求方差，标准差 σ(Y) = sqrt(Var(Y))
- E(Y1 + Y2) = E(Y1) + E(Y2)
- Var(Y1 + Y2) = Var(Y1) + Var(Y2)，如果 Y1 和 Y2 相互独立
- $Var(a*Y1 + b) = a^2 * Var(Y1)$

样本集的平均值 M_s 是一个随机变量。可以想象我们进行了多次抽样，每次都会得到不同的样本集 S 和各自的平均值 M_s。M_s 会形成某种概率分布，是什么分布呢？

- 如果原始数据符合正态分布，则样本集平均值 M_s 符合**正态分布**。（详细[26]）
- 通常 M_s 接近**正态分布**；只要原始数据不是非常偏斜（Skewed），并且样本集足够大（通常 n > 30）。

上述就是鼎鼎大名的中心极限定理（Central Limit Theorem [1]）。**通常可以认为，样本集平均值 M_s 符合正态分布**。下面来计算值 M_s 分布的期望和方差:

$$E(M_s) = E\left(\frac{\sum_{i=0}^{n-1} X_i}{n}\right) = \frac{\sum_{i=0}^{n-1} E(X_i)}{n} = E(X)$$

$$Var(M_s) = Var\left(\frac{\sum_{i=0}^{n-1} X_i}{n}\right) = \frac{1}{n^2} Var\left(\sum_{i=0}^{n-1} X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$\sigma(M_s) = \sqrt{Var(M_s)} = \frac{\sigma}{\sqrt{n}}$$

可以看到，样本集平均值 M_s 符合这样一个正态分布，其期望等于原始数据的期望（废话）。**其标准差等于原始数据的标准差除以样本集大小的开方**。这侧面说明， 样本集越大，其统计特征越稳定。

可信的平均值所需的样本集大小

从上文可见，最有意义的量是样本集的平均值的标准差 σ (M_s)；它有个特殊名字，叫**平均值的标准误**（Standard Error of the Mean [2]）。相比“标准差”通常指原始数据，“标准误”用来指样本集。

如何决定多大的样本集合适呢？即是需要样本集的平均值与原始数据的平均值足够接近。也就是说，让平均值的标准误足够小。也就是说，通过让样本集的大小 n 足够大，**以使平均值的标准误 σ (M_s) = σ / sqrt(n) 足够小**。

样本集平均值 M_s 终归是随机的，我们需要引入一些新概念以描述其随机性

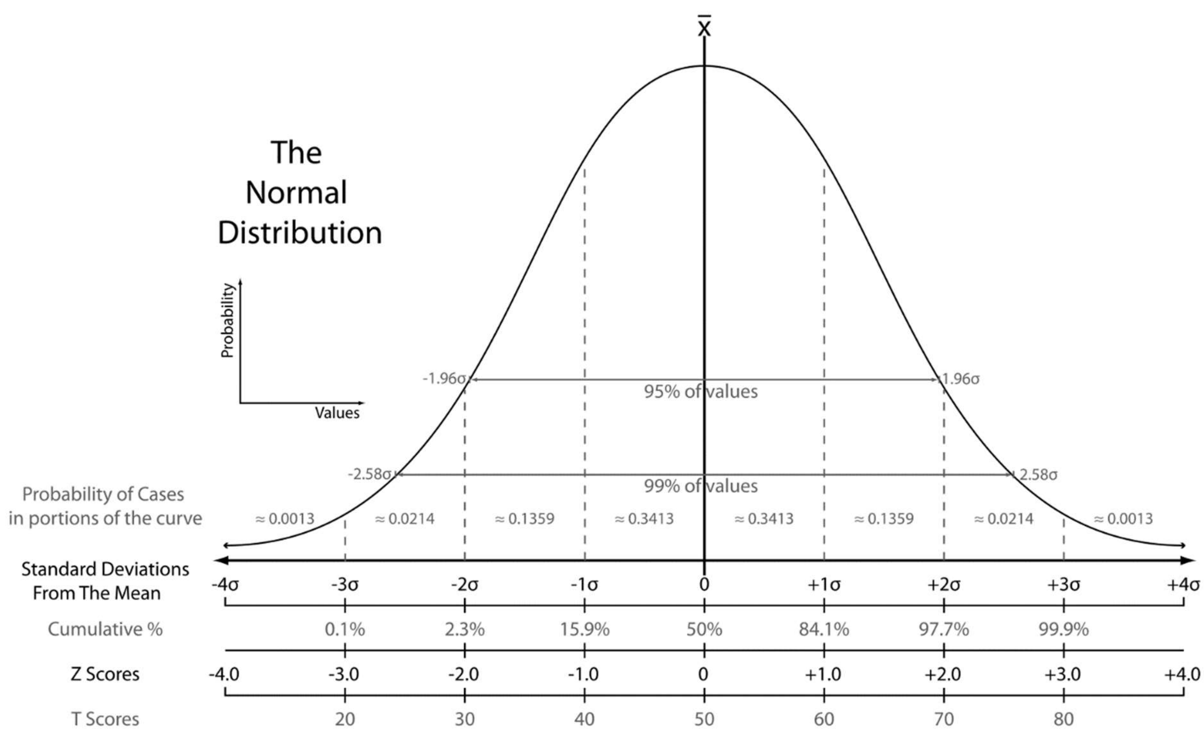
- **误差范围（Margin of Error）**：指样本集平均值与原始数据平均值之差——$|M_s - E(X)|$——的最大值。我们希望样本集足够大，以使误差范围足够小。误差范围也称作置信区间（Confidence Interval）。

- **置信水平（Confidence Level）**：有多大概率，M_s 会落在我们指定的误差范围内；常选 99%，95%，90%

对于符合正态分布的 M_s，有没有更好的方法表示误差范围呢，将其标准差也考虑进去？有，就是

- **Z-Score [3]**：Z-Score := $(x - \mu) / \sigma$。在我们讨论的 M_s 的上下文中，Z-Score = Margin of Error / $\sigma(M_s)$。

置信水平和 Z-Score 是一一对应的，在正态分布中可以通过查表[4]得到。置信水平指样本值落在 $\mu \pm$ Margin of Error 区间的概率，它是正态分布图像与 $x = \mu \pm$ Margin of Error 围成的面积。见下图[5]。



举个例子，我们希望有 95%的概率，样本集平均值 M_s 会落在误差范围内。这就是说，置信水平是 95%。通过查表，我们发现对应的 Z-Score 是 1.96。那么，Margin of Error / σ (M_s) = 1.96。

继续，给定我们想要的误差范围（Margin of Error），就可以通过 σ (M_s) = σ / sqrt(n)推算出样本集的大小 n 了。可以看出原始数据的标准差 σ 越大，所需样本集越大。但我们可以简化定义，**令 Margin of Error = 10% * σ**。最终算出**样本集大小 n = 384。**

如下，我们有了**通过样本集的平均值来决定样本大小的方法**：

1. 设定置信水平，比如 99%。
2. 通过置信水平查表得到 Z-Score，比如 2.576
3. 设定误差范围对原始数据标准差的比例，由 Margin of Error / σ 表示，比如 1%
4. 用如下公式计算样本集大小，**比如结果是 n = 66358**

$$\mathrm{n} = Z\text{-}Score^2 * \left(\frac{Margin\ of\ Error}{\sigma}\right)^{-2}$$

如上，**计算出来的 n 可以如下解释**：

- 如果样本集大小大于 66358，那么有 99%的概率，我们计算出的样本集的平均值与实际的偏差，小于原始数据标准差的 1%。

上述方法也见于 StatisticsHowTo [6]（“Known population standard deviation“一节）；如其标题，要求事先知道原始数据的标准差 σ，不然无法独立控制误差范围（Margin of Error）。Wikipedia [7]（“Estimation of a mean”一节）也记录了上述方法。

样本集的方差（Sampling Distribution of Sample Variance）

上文中，用样本集的平均值 M_s 决定样本集大小有不足：误差范围（Margin of Error）无法甩掉原始数据的标准差 σ。**我们想用样本集的标准差估算原始数据的标准差 σ，多大的样本集合适？**

先设定新符号：

- **σ_s^2：表示样本集的方差**，$\sigma_s^2 = \sum(X_i - M_s)^2 / (n - 1)$。（样本集除以 n - 1 而不是 n，[因为 χ^2 分布自由度[8]]）

- σ_s：表示样本集的标准差，$\sigma_s = sqrt(\sigma_s^2)$ （废话）

和 M_s 类似，样本集的标准差 σ_s 是一个随机变量。想象我们多次抽样，每个样本集都有各自的标准差 σ_s。**σ_s 符合什么概率分布呢？**也类似 M_s 的思路，确定概率分布后，我们就能确定合适的样本集大小 n。

计算 σ_s^2 的技巧在于

- 与计算 $Var(M_s)$不同，$Var(\sigma_s^2)$包含平方。X_i 暗示正态分布，而 X_i 的平方暗示 χ^2 分布

- 多出来 M_s 也是个随机变量，直觉告诉我们，应该变形等式，将 X_i 的平方和 M_s 分开。

$$
\begin{aligned}
(1)\quad (X_i - M_s)^2 &= \underbrace{(X_i - M_s)}_{\text{一次方的(Xi - Ms)容易处理，将容易处理的部分“提纯”}}(X_i - M_s) \\
&= (\underbrace{(X_i - E(X)) - (M_s - E(X))}_{\text{用常量E(X)对齐}X_i\text{和}M_s})((X_i - E(X)) + (M_s - E(X)) - 2\underbrace{(M_s - E(X))}_{\text{对i求和时是不变量}}) \\
&= (X_i - E(X))^2 - \underbrace{(M_s - E(X))^2}_{\text{对i求和时是不变量}} - 2\underbrace{(X_i - M_s)}_{\text{对i求和正好是零}}(M_s - E(X))
\end{aligned}
$$

$$
\begin{aligned}
(2)\quad \sum_{i=0}^{n-1}(X_i - M_s)^2 &= \sum_{i=0}^{n-1}(X_i - E(X))^2 - \sum_{i=0}^{n-1}(M_s - E(X))^2 - \sum_{i=0}^{n-1}2(X_i - M_s)(M_s - E(X)) \\
&= \sum_{i=0}^{n-1}(X_i - E(X))^2 - n(M_s - E(X))^2
\end{aligned}
$$

$$
(3)\quad \frac{{\sigma_s}^2}{\sigma^2} = \frac{1}{n-1}\Bigg(\underbrace{\sum_{i=0}^{n-1}\Big(\underbrace{\frac{X_i - E(X)}{\sigma}}_{\text{N(0, 1)（如果X符合正态分布）}}\Big)^2}_{\text{式子(4)：n个N(0, 1)的平方和，}\chi^2\text{(n)分布 （如果X符合正态分布）}} - \underbrace{\Big(\underbrace{\frac{M_s - E(X)}{\sigma/\sqrt{n}}}_{M_s\text{的标准化的正态分布，N(0, 1)（当n足够大）}}\Big)^2}_{\text{式子(5)： N(0, 1)的平方，}\chi^2\text{(1)分布（当n足够大）}}\Bigg)
$$

如上图中标注，**结果式(3)描述了样本集的方差 σ_s^2 的分布**。它由两个部分，式(4)和式(5)，组成。式(4)和式(5)都被常量 E(X)和σ标准化了。

- 式(5)中，根据中心极限定理，我们知道 M_s 趋近正态分布。式(5)趋近正态分布 $N(0,1)$的平方

- 式(4)中，核心部分是$(X_i - E(X)) / \sigma$。若原始数据 X 符合正态分布，则核心部分符合正态分布 $N(0,1)$。式(4)的形式类似于正态分布的平方和

1. 当原始数据 X 符合正态分布

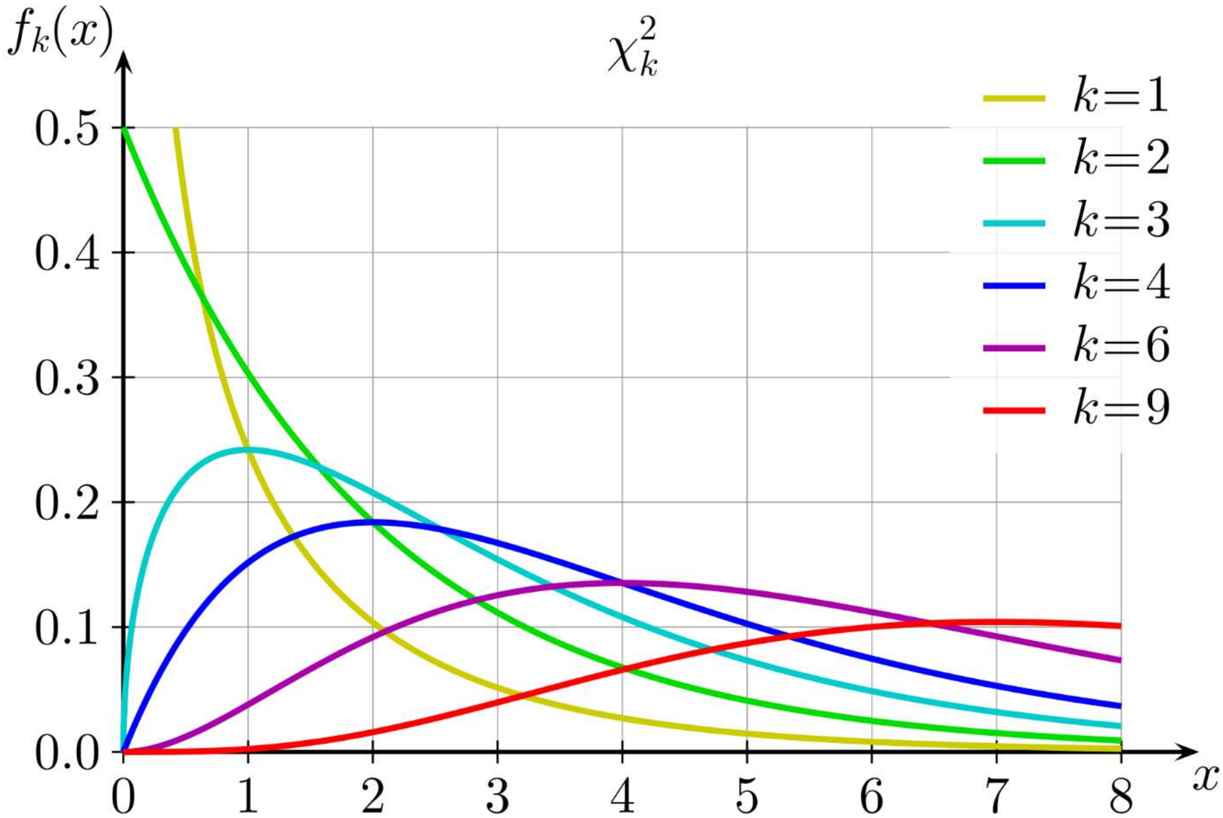
如果原始数据 X 符合正态分布，一切都变得简单。式子(3)是正态分布的平方和减去正态分布的平方。首先，正态分布的平方以及平方和是什么分布呢?

那就是大名鼎鼎的 **χ^2 分布（Chi-square distribution，卡方分布）**，插图来自 Wikipedia [9]:

- 如果随机变量 Y 符合标准正态分布 N(0, 1)，那么 Y^2 符合自由度为 1 的 χ^2 分布，记作 $\chi^2(1)$

- m 个互相独立的标准正态分布 N(0, 1)，其平方和符合自由度为 m 的 χ^2 分布，记作 $\chi^2(m)$

- 互相独立的 χ^2 分布相加，结果还是 χ^2 分布，后者的自由度是前者自由度之和。详细证明[10]用到 Moment-Generating Function



所以，式(4)符合自由度为 n 的 $\chi^2(n)$，式(5)符合自由度为 1 的 $\chi^2(1)$。两者相减，得到自由度为 n-1 的 $\chi^2(n-1)$。**所以，结论是，样本集的方差 σ_s^2 符合自由度为 n – 1 的卡方分布：$\chi^2(n-1)$。**

$$\frac{\sigma_s{}^2}{\sigma^2} \sim \frac{\chi^2(n-1)}{n-1}$$

（详细严谨的证明[23]见这里；应该只用随机变量相加而不是相减，还需要证明 σ_s^2 与 M_s 相互独立）

从上面式子中 $\chi^2(n-1)$的自由度 n - 1，可以理解为什么定义 σ_s^2 时，是除以 n - 1 而不是 n。从 χ^2 的期望和方差[9]，我们可以得到：

- $E(\sigma_s^2/\sigma^2) = 1$，这意味着 σ_s^2 是无偏估计（Unbiased Estimator）。（废话）
- **$Var(\sigma_s^2/\sigma^2) = 2 / (n - 1)$**，这意味着随着样本集大小 n 增大，样本集方差 σ_s^2 的波动逐渐减小

2. 当原始数据 X 不符合正态分布

式(5)没问题，随着样本集大小 n 增大，根据中心极限定理，其仍然趋近 $\chi^2(1)$。

式(4)有问题。中心极限定理仅对随机变量之和有效；很遗憾，**对随机变量的平方和没有[11]中心极限定理**。

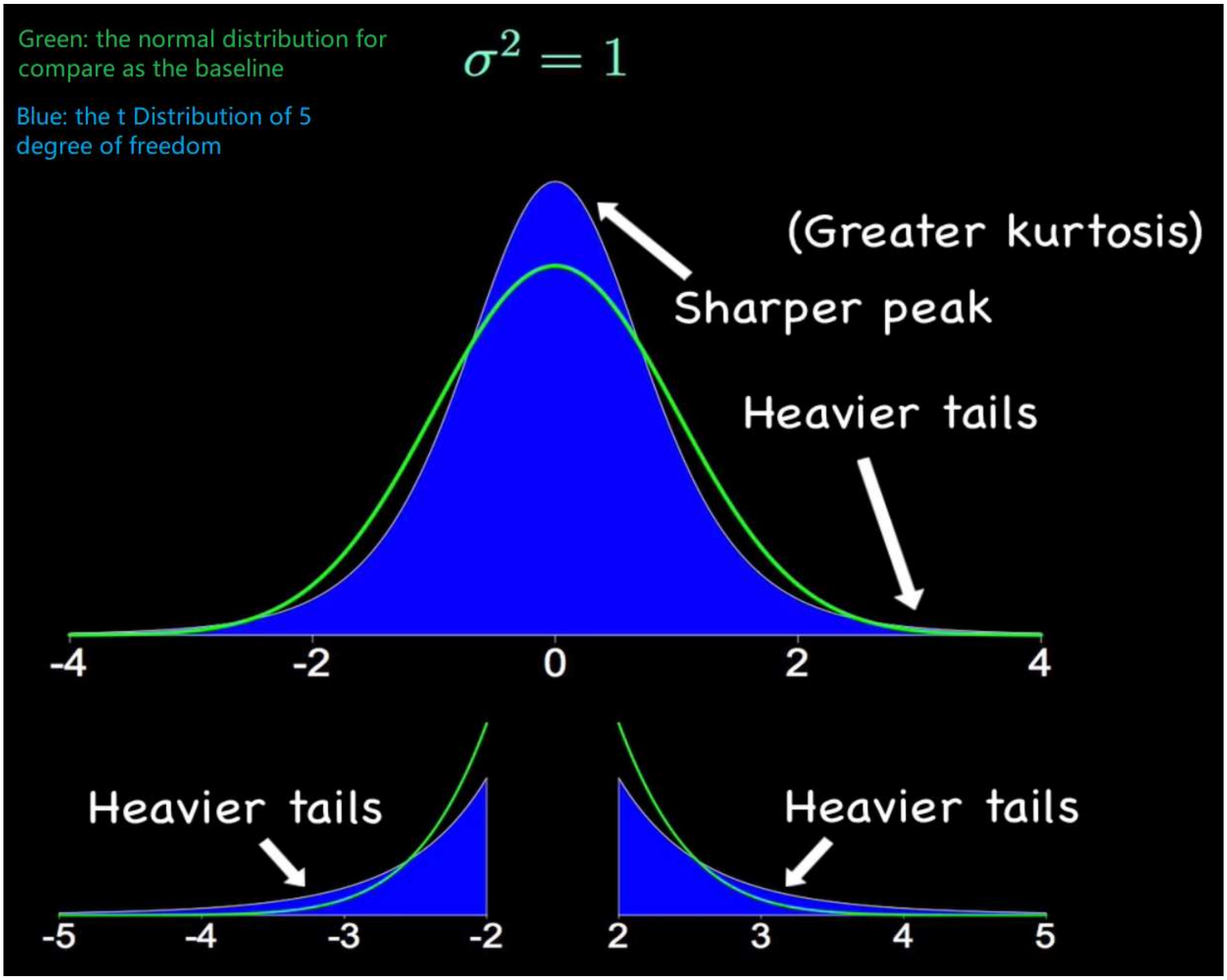
那么，当样本集大小 n 足够大时，样本集方差 σ_s^2 是否趋近 χ^2 分布呢？jbstatistics [12] 的模拟方法精彩地揭示了：

- 当 n 非常大时，σ_s^2 会接近正态分布
- σ_s^2 分布与 χ^2 分布有偏差；**即使 n 很大，并不能将偏差消除**
- σ_s^2 与原始数据的真实方差 σ^2 有偏差；**即使 n 很大，并不能将偏差消除**

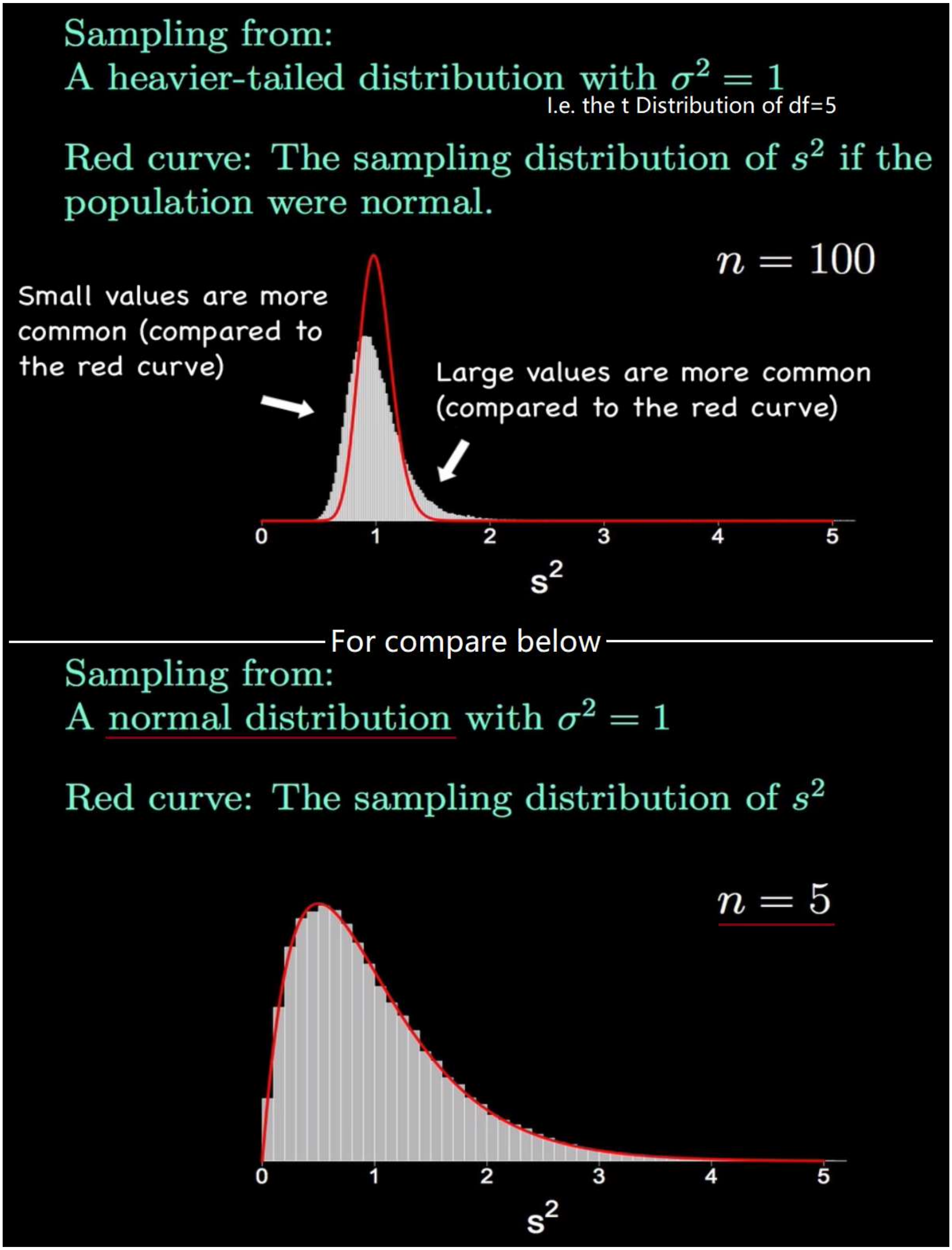
更进一步，**偏差来源于原始数据分布的峰度（Kurtosis）[13]**。正态分布的峰度为零，峰度大表示分布有长尾（Long-tailed），峰度小表示分布更集中在中心。

- 无论峰度 > 0 或峰度 < 0，σ_s^2 分布与 χ^2 分布都有偏差
- 峰度 > 0 时，σ_s^2 与真实 σ^2 有偏差；峰度 < 0 时，σ_s^2 更容易接近真实 σ^2

下图例子来自 jbstatistics [12]（我加了一些注释）。用自由度为 5 的 t Distribtuion [14]（蓝色）与正态分布（绿线）对比，两者方差都为 1。t Distribution 有更大的峰度，即长尾上概率密度更大；也有更高的中心峰。



下图中，对 t Distribution 进行多次抽样，样本集大小 n = 100。计算样本集的方差，累计绘出柱状图（白色），与正态分布抽样（红线）对比。可以看见，偏差很明显；**即使 n = 100 时偏差仍很大**。作为对比的正态分布抽样，在 n = 5 时偏差已小到难以看见。



可信的方差所需的样本集大小

多大的样本集，其方差足够接近原始数据的真实方差？同样， 我们需要区分原始数据 X 是否符合正态分布

1. 当原始数据 X 符合正态分布

如果原始数据 X 符合正态分布，或者有自信非常接近正态分布，那么 σ_s^2 / σ^2 符合 $\chi^2(n-1) / (n-1)$分布；后者期望为 1，方差随 n 减小。

类似于用样本集平均值 M_s 决定样本集大小 n 的过程， 我们定义：

- **误差范围（Margin of Error）**： σ_s^2 / σ^2 相对于 1 的偏差，即 $\sigma_s^2 / \sigma^2 = 1 \pm$ Margin of Error

- **置信水平（Confidence Level）**： 有多大概率 σ_s^2 / σ^2 会落在误差范围 $1 \pm$ Margin of Error 内

通过在线 χ^2 计算工具[15]，我们可以逐渐增大自由度，也就是样本集大小；直到 $1 \pm$ Margin of Error 区间内的概率密度面积超过置信水平。此时我们可以说，有“置信水平”这么高的概率，样本集的方差偏离真实方差小于“误差范围”。

我们已经预先计算好了一些不同设定下的样本集大小 n：

- 99%置信水平，误差范围±1%：**要求 n >= 140000**

- 99%置信水平，误差范围±5%：**要求 n >= 5500**

- 95%置信水平，误差范围±5%：**要求 n >= 3499**

2. 当原始数据 X 不符合正态分布

别担心， 我们计算样本集的方差，终归是为了作为输入，通过平均值方法决定样本集大小。即使方差不准确，我们只需把方差适当上调，便可以保证样本集大小不被低估。

另一方面，软件开发中的数据量通常较大，样本集大小 n 为几万几十万并不少见，此时误差已经较小。前文 jbstatistics [12] 例子中的偏差，只是以 $n = 100$ 举例。

还可以使用工程中常见的渐进迭代方法：逐步调大 n，直到样本集的方差 σ_s^2 的波动小于设定的阈值。

样本的“比例”（Sampling Distribution of Sample Proportion）

（英文原文是 Proportion，翻译成“比例”不太确切。）

这一节所讲的是，举个例子：原始数据 X 中有 p 比例是坏鸡蛋，其余是好鸡蛋。抽出 n 个样本，通过样本集的坏鸡蛋的比例 p_s，来推测 p。

问题：样本集大小 n 需要多大，才能确信样本集的“比例”p_s，足够接近原始数据的真实“比例”p？

另外，这一节是样本百分位（Percentiles）的基础。首先我们定义符号：

- p：表示原始数据集中，有 p 比例是我们想要找的“坏鸡蛋”
- p_s：表示样本集中，观察到有 p_s 比例是“坏鸡蛋”

那么，p_s 的分布是什么？我们可以转化一下问题[16]，从 X 定义一个新的随机变量 Z：

- $Z_i := 1$ 如果 X_i 是坏鸡蛋，反之 0。那么，Z 满足伯努利分布（Bernoulli Distribution [17]）；即 Z 是从固定比例中抽样，要么中，要么不中。
- 可以发现，$p_s = \sum Z_i / n$，即求和每个样本的 Z_i。抽样变成一个重复 n 次的伯努利试验；**p_s 符合二项分布（Binomial distribution [18]）：$p_s \sim B(n, p) / n$**。
- $E(p_s) = p$；$Var(p_s) = p(1 - p) / n$；**$\sigma(p_s) = sqrt(p(1 - p) / n)$**。由此可见，随着 n 增大，$p_s$ 分布的方差减小，p_s 与真实 p 逐渐接近。

还记得中心极限定理吗？二项分布是随机变量之和，这意味着，当 n 足够大时，二项分布接近正态分布。即，**样本集的“比例”p_s 近似符合正态分布：**$p_s \sim N(\mu=np, \sigma^2=np(1-p)) / n$ = **N(p, p(1-p)/n)**。实践中经常当正态分布处理[19] 。

(另 KhanAcademy [20] 有举例讲解；jbstatistics [21] 有讲解和模拟。)

可信的“比例”（Proportion）所需的样本集大小

既然样本集的“比例”p_s 符合正态分布 N(p, p(1-p)/n)，那么计算样本集大小 n 的方法与“**可信的平均值所需的样本集大小**”一节**一模一样**。公式如下：

$$n = Z\text{-}Score^2 * \frac{p(1-p)}{Margin\ of\ Error^2}$$

$$\leq Z\text{-}Score^2 * \frac{0.25}{Margin\ of\ Error^2} \Big|_{p=0.5}$$

可以看到，p = 0.5 时，p(1 - p)取得最大值；通常在此处计算 n。注意误差范围（Margin of Error）是绝对量：例如取 p = 1%，Margin of Error = 0.05%，含义是 p_s 在 0.01 ± 0.0005 范围内浮动。

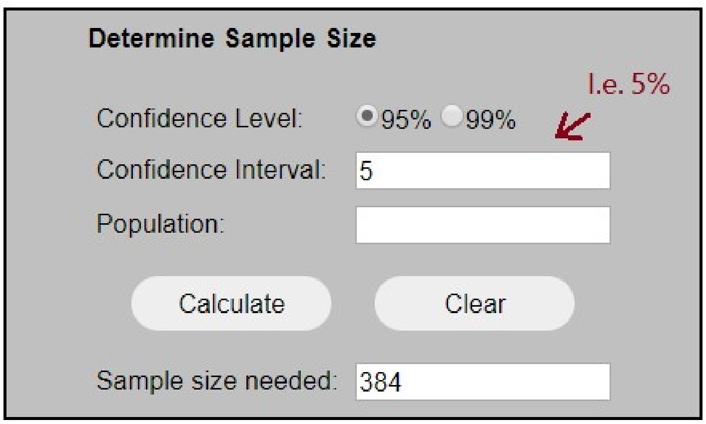
我们已经预先计算好了一些不同设定下的样本集大小 n：

- 95%置信水平，误差范围±5%：**要求 n >= 385**
- 95%置信水平，误差范围±1%：**要求 n >= 9604**
- 95%置信水平，误差范围±0.1%：**要求 n >= 960400**

- 99%置信水平，误差范围±0.1%：要求 n >= 1658944
- 99%置信水平，误差范围±0.05%：要求 n >= 6635776
- 99%置信水平，误差范围±0.01%：要求 n >= 165894400

上述方法也见于 StatisticsHowTo [22]，有详细步骤。Wikipedia [19]（“Estimation of a proportion”一节）也记录了上述方法。

相比“可信的平均值所需的样本集大小”，上述方法**不需要知道原始数据 X 的方差**，因而也更常用。网上的有不少样本集大小计算器[24]（除了没告诉你为什么）：



可信的百分位数（Percentiles）所需的样本集大小

百分位数，例如 P99，指数据排序后位于第 99%位置的数据的值。P99 经常用于衡量系统延迟；例如 P99 = 1s，说明 99%的用户延迟小于 1s，1%用户的延迟大于等于 1s。对于千万用户的云系统，1%意味着大量用户；云系统非常强调 P99 这样长尾部分的表现，而不仅仅是平均值。

有了**样本的“比例”（Proportion）**为基础，我们可以研究样本集的百分位数。首先定义符号：

- 百分位 p：例如 P99 = 1s 中，p = 0.99，表示其百分位、位置

- 百分位数 Q_p：例如 P99 = 1s 中，$Q_p = 1s$，表示**原始数据** X 在百分位 p 处的值

- 百分位数 $Q_{p,s}$: 表示样本集在百分位 p 处的值；$Q_{p,s}$ 是我们能观测的，Q_p 是想推测的

原始数据的百分位数 Q_p 将原始数据分成了两个部分：小于 Q_p 的位于[0, p)范围的原始数据，和大于等于 Q_p 的位于[p, 1]范围的原始数据。前者占总比例 p，后者占总比例 1 - p。

第一步：百分位“比例”划分，确定 $Q_{p,s}$ 下界

按照**样本的“比例”（Proportion）**中的方法抽样，我们视落入原始数据的[p, 1]范围的样本，为“坏鸡蛋“。样本集大小为 n。设我们使用的误差范围（Margin of Error）为 M_1，置信水平（Confidence Level）为 C_1。则，

- 样本集 S 分成互补的两个部分：S_1 来自原始数据的[0, p)范围，S_2 来自原始数据的[p, 1]范围

- 有 C_1 的概率，$|S_1| / |S| = p \pm M_1$；即 S_1 和 S 元素个数之比为 p，再加上误差。另，$|S| = |S_1| + |S_2| = n$

- S_1 中元素的最大值 < Q_p，因为 S_1 来自原始数据的[0, p)范围。同理，S_2 中元素的最小值 >= Q_p

一般我们希望高估百分位数而不是低估，例如用 P99 衡量系统延迟。**下面按照希望高估的情形计算**。我们从 S_2 中选取最小值作为样本集的百分位数 $Q_{p,s}$，有 $Q_{p,s} >= Q_p$。

我们希望 $Q_{p,s}$ 不要高估 Q_p 太多。样本集 S_2 是对[p, 1]范围的原始数据的随机抽样， 其个数$|S_2|$越少，$Q_{p,s}$ 越容易高估。**我们选最坏情况$|S_1| / |S| = p + M_1$；此时$|S_2| = n*(1 - p - M_1)$。**

现在，我们可以通过元素个数关系来将样本集划分，真正地确定 S_1 和 S_2。从而可以从 S_2 中选取最小值作为 $Q_{p,s}$。我们一定在让 S_2 个数偏少，从而在高估 Q_p。

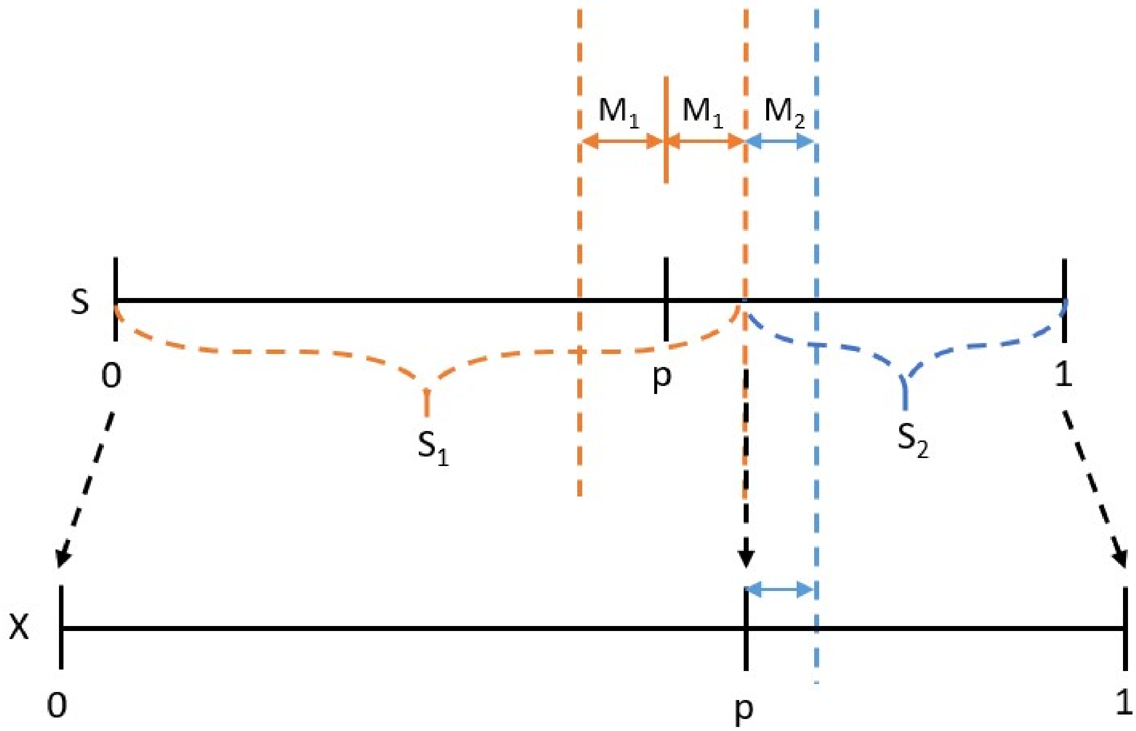
我们可以说：

- **有至少 C_1 的概率，样本集的 $Q_{p,s}$ 大于等于原始数据 X 位于百分位 p 处的百分位数 Q_p**

第二步：二项分布抽样，确定 $Q_{p,s}$ 上界

现在，如下图所示，S_2 占样本集的$[p + M_1, 1]$的范围，S_2 中元素来自原始数据的$[p, 1]$范围。

我们希望 S_2 中至少有一个元素，e，来自原始数据的$[p, p+M_2)$范围，设该事件发生的概率为 C_2。那么，$Q_{p,s} <= e < Q_{p+M2}$，由此我们确定了 $Q_{p,s}$ 高估的上界。



我们可以说：

- **有 C_2 的概率，样本集的 $Q_{p,s}$ 小于原始数据位于百分位 $p+M_2$ 处的百分位数 Q_{p+M2}**

如何计算 C_2？由于 S_2 可看作对原始数据 $[p, 1]$范围的随机抽样，C_2 是单次概率为 $M_2 / (1 - p)$的重复实验，可以用二项分布计算。（下文有公式）

第三步：约束方程，确定样本集大小

为了描述问题，我们先定义置信水平和误差范围；它们与前文的各种定义是一致的：

- **误差范围（Margin of Error）**：我们得到的样本集的百分位数 $Q_{p,s}$，位于原始数据 X 的 p 百分位到 p + Margin of Error 百分位之间；即 $Q_{p,s}$ 来自于原始数据的[p, p + Margin of Error]范围。

- 注意，**我们控制的是百分位的误差范围，而不是百分位数的范围**。注意，**我们确保 $Q_{p,s}$ 一定不会低于原始数据的真实百分位数 Q_p**。如果原始数据在百分位 p 附近有跳跃，即使百分位的误差范围很小，百分位数的变动也可能很大

- **置信水平（Confidence Level）**：有多大概率，$Q_{p,s}$ 会落在上述误差范围内

另，为了方便表达：

- Confi(Z-Score)函数：正态分布中，Z-Score 对应的置信水平。它是 $x = \mu \pm \text{Z-Score} * \sigma$ 与正态分布围成的区域的面积，包括 $x = \mu$ 的左右两半边

结合前文两个步骤各自的置信水平 C_1 和 C_2，以及各自的误差范围 M_1 和 M_2，我们可以得出如下图方程，来约束样本集大小 n：

$$(1)\quad \begin{cases} Confidence\ Level = C_1 * C_2 \\ Margin\ of\ Error = M_2 \end{cases}$$

$$(2)\quad C_1 = Confi(Z\text{-}Score) = Confi\left(\frac{M_1}{\sqrt{\frac{p(1-p)}{n}}}\right)$$

$$(3)\quad C_2 = 1 - \left(1 - \frac{M_2}{1-p}\right)^{n(1-p-M_1)}$$

方程中，Confidence Level 和 Margin of Error 是用户输入，M_1 是可调节变量。最终需要让 n 足够大，以满足用户指定的 Confidence Level。C_1 和 C_2 对 n 都是单调递增的。

如何确定合适的 M_1? 以 M_1 为自变量，对 $C_1 * C_2$ 求导，可以发现：$C_1 * C_2$ 在(0, 1 - p)区间上首先单调递增，然后单调下降，有且仅有一个极值点。

设：

$$\begin{cases} Z\text{-}Score = \dfrac{M_1}{\sqrt{p(1-p)/n}} \\ f(x) = \dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad \Rightarrow Confi(\text{Z-Score}) = 2\displaystyle\int_0^{\text{Z-Score}} f(x)dx \quad \text{// 标准正态分布N(0, 1)} \\ g(M_1) = (1 - \dfrac{M_2}{1-p})^{n(1-p-M_1)} \end{cases}$$

$$(C_1 * C_2)_{M_1}{}' = \frac{2f(Z\text{-}Score)}{\sqrt{p(1-p)/n}}\left(1 - \left(1 + n\left(-\ln\left(1 - \frac{M_2}{1-p}\right)\right)\sqrt{p(1-p)/n}\frac{F(Z\text{-}Score)}{f(Z\text{-}Score)}\right)g(M_1)\right) = 0$$

$$(4) \quad \Rightarrow \left(1 + n\left(-\ln\left(1 - \frac{M_2}{1-p}\right)\right)\sqrt{p(1-p)/n}\frac{F(Z\text{-}Score)}{f(Z\text{-}Score)}\right)g(M_1) = 1$$

上图中，式(4)给出了使 $C_1 * C_2$ 最大的 M_1 极值点所在位置。如果知道 Confidence Level 和 Margin of Error 的具体数值，可以用数值方法暴力求解该 M_1。通过此 M_1，能够解得最小的 n。

但为了简单，下文中，我们**强行要求 M_1 = Margin of Error**。这样，通过逐步增大样本集大小 n，$C_1 * C_2$ 会逐步增大，直到满足置信水平（Confidence Level）。这样，**我们求得了样本集大小 n 值**。

第四步：总结

通过样本集估计原始数据在百分位 p 处的百分位数 Q_p，如何确定合适的样本集大小？

- **用户输入：百分位 p（例如 0.99），误差范围（Margin of Error），置信水平（Confidence Level）**

接下来，带入上文方程(1)，(2)，(3)。令 M_1 = Margin of Error。M_2 就是用户输入的 Margin of Error。关于如何计算正态分布中的 Confi 函数、Z-Score，参考“**可信的平均值所需的样本集大小**”一节。

- **逐步增大样本集大小 n，直到计算出的 Confidence Level = C_1 * C_2 满足用户输入**

现在已经确定了样本集大小 n。抽样后，如何确定样本集的百分位数 $Q_{p,s}$ 呢？

- **将样本集从小到大排序，取百分位 p + Margin of Error 处的数据值作为样本集的百分位数 $Q_{p,s}$**

样本集的百分位数 $Q_{p,s}$ 是对原始数据 X 在 p 处的百分位数 Q_p 的估计。我们采用的方法偏好并且保证高估 Q_p（例如测量系统延迟需要）。如需低估偏好可参考前文依葫芦画瓢稍作改造。

现在，我们可以说：

- **有 Confidence Level 的概率，我们测得的样本集的百分位数 $Q_{p,s}$，位于原始数据 X 的 p 百分位到 p + Margin of Error 百分位之间：$Q_p <= Q_{p,s} < Q_{p+\text{Margin of Error}}$**

我们已经预先计算好了一些不同设定下的样本集大小 n：

- P99 百分位，99%置信水平，误差范围±0.1%：要求 n >= 70000

- P99 百分位，99%置信水平，误差范围±0.05%：要求 n >= 300000

- P99 百分位，99%置信水平，误差范围±0.01%：要求 n >= 7000000

- **P99 百分位，95%置信水平，误差范围±0.1%：要求 n >= 40000**

- P99 百分位，95%置信水平，误差范围±0.05%：要求 n >= 160000

- P99 百分位，95%置信水平，误差范围±0.01%：要求 n >= 4000000

- **P50 百分位，99%置信水平，误差范围±1%：要求 n >= 17000**
- P50 百分位，99%置信水平，误差范围±0.1%：要求 n >= 1700000
- P50 百分位，99%置信水平，误差范围±0.05%：要求 n >= 7000000
- P50 百分位，99%置信水平，误差范围±0.01%：要求 n >= 170000000

- P50 百分位，95%置信水平，误差范围±1%：要求 n >= 10000
- P50 百分位，95%置信水平，误差范围±0.1%：要求 n >= 1000000
- P50 百分位，95%置信水平，误差范围±0.05%：要求 n >= 4000000
- P50 百分位，95%置信水平，误差范围±0.01%：要求 n >= 100000000

- **P10 百分位，99%置信水平，误差范围±1%：要求 n >= 6000**
- P10 百分位，99%置信水平，误差范围±0.1%：要求 n >= 600000

- P10 百分位，99%置信水平，误差范围±0.05%：要求 n >= 2400000
- P10 百分位，99%置信水平，误差范围±0.01%：要求 n >= 60000000

- P10 百分位，95%置信水平，误差范围±1%：要求 n >= 3500
- P10 百分位，95%置信水平，误差范围±0.1%：要求 n >= 350000
- P10 百分位，95%置信水平，误差范围±0.05%：要求 n >= 1400000
- P10 百分位，95%置信水平，误差范围±0.01%：要求 n >= 35000000

- P999 百分位，99%置信水平，误差范围±0.01%：要求 n >= 700000
- P999 百分位，99%置信水平，误差范围±0.005%：要求 n >= 2700000
- P999 百分位，99%置信水平，误差范围±0.001%：要求 n >= 70000000

- **P999 百分位，95%置信水平，误差范围±0.01%：要求 n >= 390000**
- P999 百分位，95%置信水平，误差范围±0.005%：要求 n >= 1540000
- P999 百分位，95%置信水平，误差范围±0.001%：要求 n >= 39000000

总结

以“**样本集的统计特征（平均值、方差、百分位数等）应该与原始数据足够接近**”为出发点，根据我们要求样本集保持的统计特征不同，能够得出不同的样本集大小需求。

大部分情况，n = 10K 大小的样本集足够了；测量 P99 需要 n = 40K。这些对今天的软件系统通常不是问题。

如果抽样测量的值涉及原始数据的具体数值，比如平均值，那么会受到原始数据方差的影响。如果抽样测量的值只涉及比例（Proportion），那么可以巧妙地规避原始数据方差的影响。

怀念曾经可以天真简单地使用抽样方法的日子……面纱之下，一入统计深似海……与之相比，工程里常用的暴力模拟和迭代逼近也不失为好方法。

（全文完）

文中资料链接

[1] Central Limit Theory: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Probability/BS704_Probability12.html

[2] KhanAcademy: Standard Error of the Mean: https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/standard-error-of-the-mean

[3] StatisticsHowTo: Z-Score: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/z-score/

[4] Z-Score Table: http://www.sjsu.edu/faculty/gerstman/StatPrimer/z-two-tails.pdf

[5] Wikipedia: Z-Score: https://en.wikipedia.org/wiki/Standard_score

[6] StatisticsHowTo: Find Sample Size - Known population standard deviation: https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/find-sample-size/#CI2

[7] Wikipedia: Determine Sample Size – Estimation of a mean: https://en.wikipedia.org/wiki/Sample_size_determination#Estimation_of_a_mean

[8] TowardsDataScience: Why sample variance divided by n - 1: https://towardsdatascience.com/why-sample-variance-is-divided-by-n-1-89821b83ef6d

[9] Wikipedia: Chi-square Distribution: https://en.wikipedia.org/wiki/Chi-squared_distribution

[10] Stat.psu.edu: STAT 414: Sums of Chi-Square Random Variables: https://online.stat.psu.edu/stat414/node/171/

[11] Ocw.mit.edu: Statistical Thinking and Data Analysis: Central Limit Theorem: https://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/lecture-notes/MIT15_075JF11_chpt05.pdf#page=3

[12] jbstatistics: Sampling Distribution of Sample Variance: https://www.youtube.com/watch?v=V4Rm4UQHij0

[13] OnlineStatBook: Kurtosis: http://onlinestatbook.com/2/glossary/kurtosis.html

[14] OnlineStatBook: t Distribution: http://onlinestatbook.com/2/estimation/t_distribution.html

[15]: Divms.uiowa.edu: Chi-square Calculator: https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

[16] Stat.StackExchange: General method for deriving the standard error: https://stats.stackexchange.com/questions/89154/general-method-for-deriving-the-standard-error

[17] Wikipedia: Bernoulli Distribution: https://zh.wikipedia.org/zh-hans/伯努利分布

[18] Wikipedia: Binomial Distribution: https://zh.wikipedia.org/zh-hans/二項分佈

[19] Wikipedia: Determine Sample Size – Estimation of a proportion:
https://en.wikipedia.org/wiki/Sample_size_determination#Estimation_of_a_proportion

[20] KhanAcademy: Sampling Distribution of Sample Proportion: https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-proportion/v/sampling-distribution-of-sample-proportion-part-1

[21] jbstatistics: Sampling Distribution of Sample Proportion: https://www.youtube.com/watch?v=fuGwbG9_W1c

[22] StatisticsHowTo: Determine Sample Size: https://www.qualtrics.com/experience-management/research/determine-sample-size/

[23] Stat.psu.edu: STAT 414: Sampling Distribution of Sample Variance https://online.stat.psu.edu/stat414/node/174/

[24] SurveySystem: Sample Size Calculator: https://www.surveysystem.com/sscalc.htm

[25] Ocw.mit.edu: Expectation, Variance and Standard Deviation: https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading6a.pdf

[26] Stat.yale.edu: Sample Means: http://www.stat.yale.edu/Courses/1997-98/101/sampmn.htm