推断统计学的几个相关概念

推断性统计是研究如何利用样本数据来推断总体特征的统计方法，主要的内容包括参数估计和假设检验。

1. 参数估计

参数估计（parameter estimation）是指在抽样及抽样分布的基础上，根据样本统计量来推断所关心的总体参数，即利用样本数据推断总体特征参数。其中由样本统计量所构造的总体参数的估计区间称为**置信区间**（confidence interval）。

置信区间的最小值称为置信下限，最大值称为置信上限。

置信区间与普通意义上理解的区间不同之处在于统计量和参数的关系，我们在初级课程中提到：抽样推断下，往往根据样本统计量去估计总体的参数（例如某批灯泡的使用寿命），即参数≈统计量。统计学家为了更加严谨的阐述事件发生的情况，增加了一个置信水平（也称为置信度、置信概率或置信系数）的概念。

置信水平（confidence level）-Pr：衡量总体均值落在置信区间的可能性大小。该数值一般由人为定义一个未落在置信区间的可能性大小，称为显著性水平，记作α，那么落在置信区间的可能性（置信水平 Pr）就等于 1-α，即：

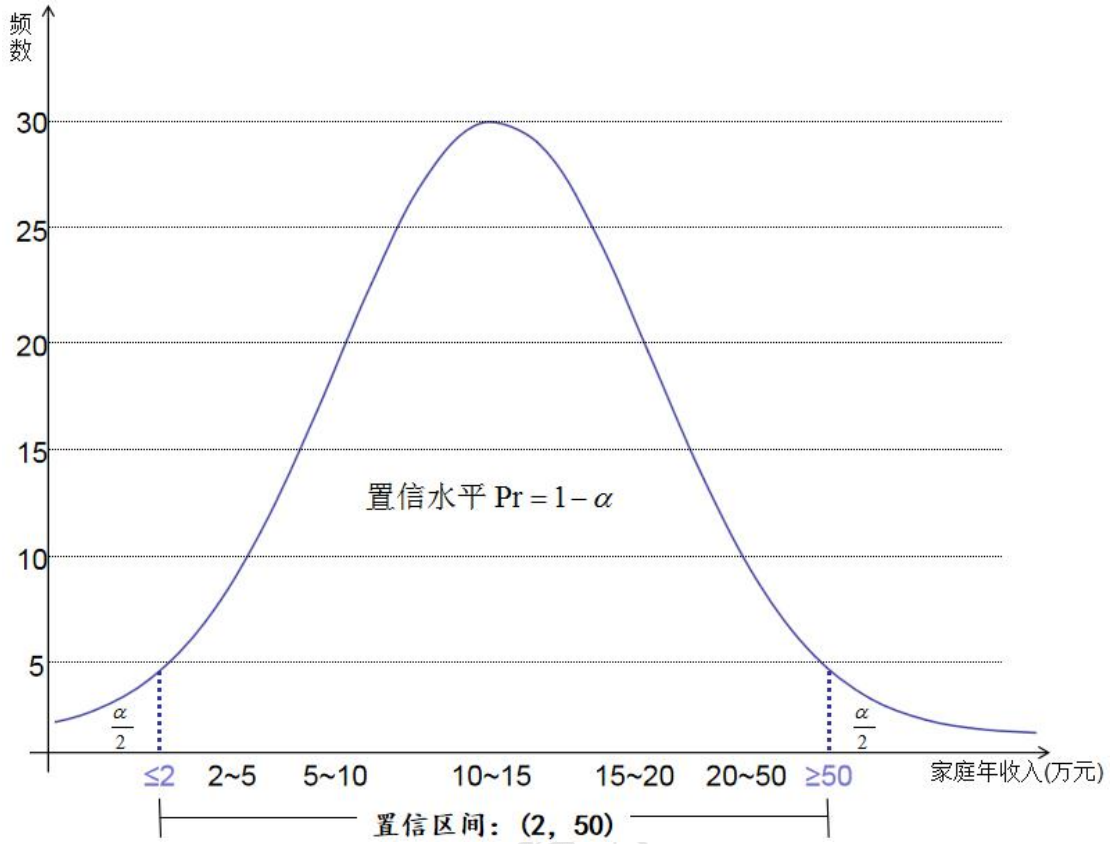
$$Pr=1-\alpha$$

例如推断某市家庭年收入，抽取了 100 个家庭收入情况如图表所示：

收入区间（万元）	频数（人数）	收入区间（万元）	频数（人数）
≤2	5	15~20	20
2~5	10	20~50	10
5~10	20	50 以上	5
10~15	30		

注：定义显著性水平α=0.1

小知识：统计学家在某种程度上确信这个区间会包含真正的总体参数，所以称该区间为置信区间。



注：为方便分析，我们假设 100 户人的家庭收入服从标准正态分布。

100 个家庭的统计量可以推断出：我们构造了该市置信水平为 90%的置信区间（2，50），即真实的总体参数有 90%的概率落在这一区间内，有 5%的概率≤2，5%的概率≥50，如此通过样本统计量来估计总体的参数是不是就更严谨了。

除 Pr=90%、α=0.1 以外：

常用置信水平值还有 95%、99%，置信水平值越大越好；

对应的显著性水平为 0.05、0.01，显著性水平越小越好。

2. 假设检验

参数估计和假设检验是推断统计的两个组成部分，他们都是利用样本对总体进行某种推断，但推断的角度不同：参数估计讨论的是用样本统计量估计总体参数的方法；而假设检验则是先对参数提出一个假设，然后利用样本信息去检验这个假设是否成立。

假设检验是利用样本数据判断对总体的假设是否成立，进行假设检验主要分为两步：

①假设：首先对总体参数或分布形式作出某种假设（原假设和备择假设）。

原假设（null hypothesis）和**备择假设**（alternative hypothesis）是对总体参数或分布形式提出的两种假设。

例如今年某地新生婴儿的平均体重这一参数 μ_0，根据去年的数据 3190 克提出：

$$H_0：\mu_0=3190（g）$$

这里 H_0 表示一个事件，即今年某地新生婴儿的平均体重为 3190 克，如果该假设（H_0）不成立，则需要提出另一种假设：

$$H_1：\mu_1\neq3190（g）$$

该假设称：备择假设，即今年某地新生婴儿的平均体重不为 3190 克。

我们可以看出，原假设与备择假设互斥，肯定原假设，意味着放弃备择假设，否定原假设，意味着接受备择假设。

小知识：

原假设一般均由 H_0 表示，它的下标为 0，所有也被称为“零假设”；

而假设检验是围绕着对原假设是否成立而展开的，所有备择假设也被称为“替换假设”，表明当原假设不成立时的替换。

②检验：然后利用样本信息来判断该假设是否成立（α错误和β错误）。

检验时采用反证法，即假定原假设成立，然后对样本值与原假设的差异进行分析，如果有充分的理由推翻这一命题，则否定原假设，反之则肯定原假设。

α错误（α error）和**β错误**（β error）是检验结果错误的两种情况。

小知识：

这里样本值与原假设的差异指的是置信区间外的取值，这些值如果是由样本的随机性引起的属于正常情况，该取值较大时，我们描述差异是显著的，同时否定原假设，反之则肯定。

对于原假设提出的命题，我们需要作出判断，这种判断可以用“原假设正确”或“原假设错误”来表述：

α错误：当原假设 H_0 为真却被我们拒绝了，犯这种错误的概率用α表示，所以α错误也被称为弃真错误；

β错误：当原假设 H_0 为伪我们却没有拒绝，犯这种错误的概率用β表示，所以β错误也被称为取伪错误；

很多模型的显著性检验和模型中系数的显著性检验的基本思想都是假设检验。其核心思想是：如果对总体所作的某种假设是真的，那么样本统计量与原假设出现显著性差异的概率是很小的。如果在某一次随机抽样中，显著性差异竟然出现了，我们就有理由怀疑这一假设的真实性，拒绝这一假设。

补充小思考：

上述今年某地新生婴儿的平均体重例子中α错误和β错误分别表示：

α错误：原假设：H_0=3190（g）真，我们认为是假，即今年某地新生婴儿平均体重与去年无差异，但我们检验的结果却不是3190克；

β错误：原假设：H_0=3190（g）假，我们认为是真，即今年某地新生婴儿平均体重与去年有差异，但我们检验的结果却是3190克。

内容简介

本节主要内容是在数据准备后，建立最终的模型。其次还会涉及一些推断统计学的相关概念和线性回归模型的相关知识，从整体上对回归分析进行介绍。

1. 主要内容

- 推断统计学的几个相关概念和线性回归模型的相关知识；
- 根据数据准备后形成的数据集构建最终的回归模型。

2. 学习目标

学完本节，能解决以下问题：

- 推断统计学主要包含哪两方面内容？其中都包含哪些基本概念？
- 线性回归和逻辑回归模型的联系是什么？其中线性回归模型涉及了哪些统计学评估指标？
- $l1$正则化和$l2$ 正则化比较有哪些区别？
- 如何使用交叉验证分割好的数据集的进行超参数的网格搜索？
- 如何查看最终模型是否存在过拟合的情况？