统计模型

在某种意义上，统计学的方向与概率论的方向恰恰相反.在概率论中，我们使用给定的概率分布来计算某些事件的概率.相反，在统计学中，我们观察实验结果，但基础的概率分布是未知的，必须通过结果来推导.

当然，实际情况中, 对于那个基础概率并非完全未知，我们希望在有所不知的情况下(如果场景中知道所有信息我们就不需要统计学方法了, 或者说并非是"概率论"起作用的场景) 利用所有已知的信息来构建最优的统计模型.

那么接下来我们先对“统计模型”给出一个正式定义.

> **统计模型**
>
> 统计模型是给定样本空间上一组概率分布的集合.

统计模型的解释是：观察值 X 的所有可能的概率分布的集合.通常，观察值由“子观察值”组成，$X=(X_1,\ldots,X_n)$ 是一个随机向量.当变量 $X_1,\ldots,X_n$ 对应于相同实验的独立重复时，我们称之为样本.这时，变量 $X_1,\ldots,X_n$ 是独立同分布的，它们的联合分布完全由相同的边际分布决定.在这种情况下，$X=(X_1,\ldots,X_n)$ 的统计模型可以通过子观察值 $X_1,\ldots,X_n$ 的一组（边际）概率密度函数来描述(因为每一个都是独立同分布, 一定要用联合分布可能显得麻烦了).

“统计模型”的概念只有通过例子才能真正变得清晰.虽然从数学上“统计模型”的定义表达得很简单，但在给定实际情况中进行统计建模的过程却是复杂的.统计研究的结果依赖于构建一个好的模型.

例 1 样本

在一个包含 N 人的大型群体中，某个特征 A 的比例为 p；我们希望“估计”这个比例 p.检查群体中的每个人是否具有特征 A 确实是一个非常好的方法, 但是工作量和相应带来的成本过大.相反，我们从群体中随机选择 n 个人（有放回）.我们观察随机变量 $X_1,\ldots,X_n$ 的实现，其中

$$X_i=\begin{cases}0 & \text{如果第 } i \text{ 个人没有特征 } A\\ 1 & \text{如果第 } i \text{ 个人有特征 } A\end{cases}$$

由于实验设置（有放回抽样），我们事先知道 $X_1,\ldots,X_n$ 是独立的，并服从伯努利分布.也就是说，

$$P(X_i=1)=1-P(X_i=0)=p$$

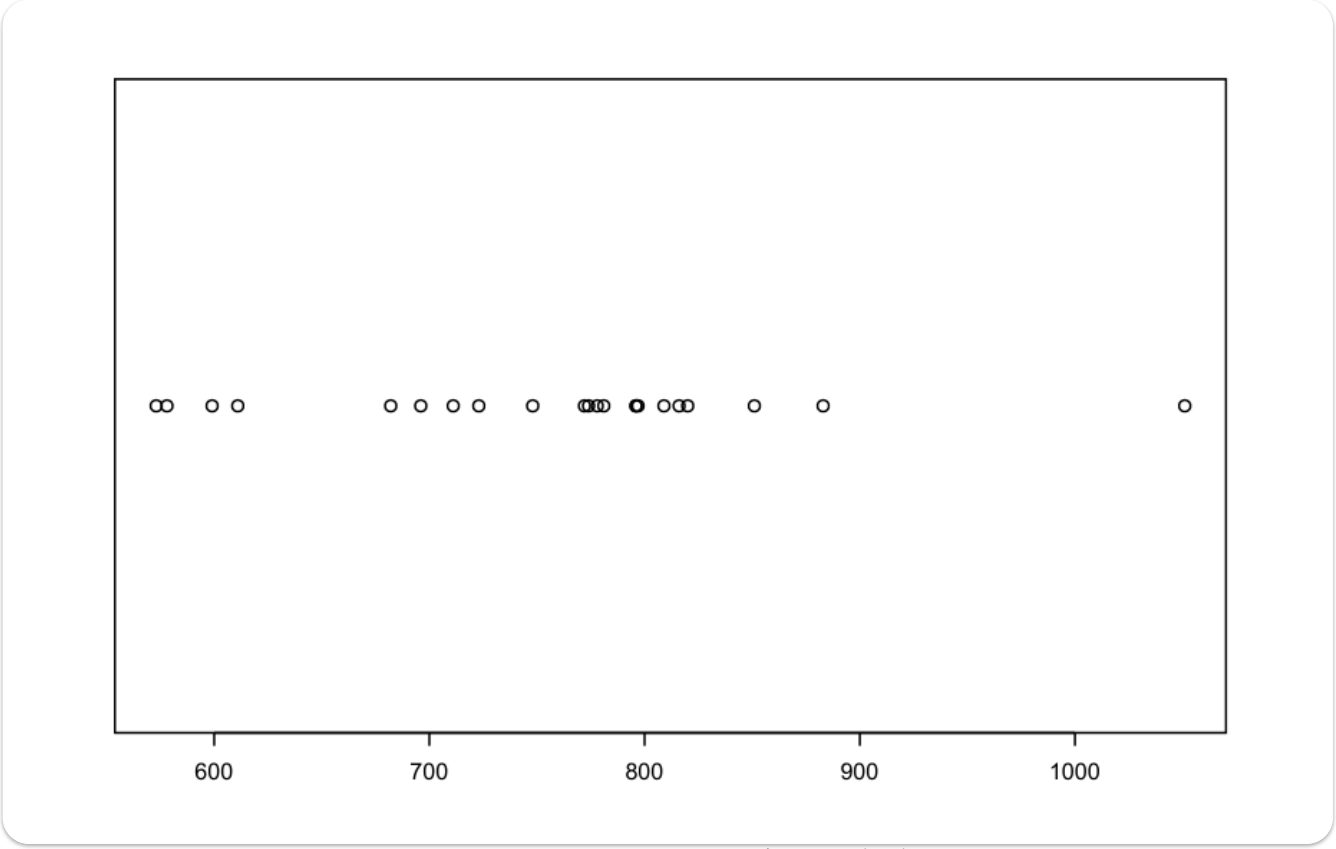
对于 $i=1,\ldots,n$.关于参数 p 没有先验知识，除了 $0\le p\le 1$.观察值是向量 $X=(X_1,\ldots,X_n)$.X 的统计模型由 X 的所有可能（联合）概率分布组成，这些分布的坐标 $X_1,\ldots,X_n$ 是独立的并且服从伯努利分布.对于 p 的每一个可能值，统计模型中包含 X 的一个唯一的概率分布, 也就是说当一个参数 p 固定的时候 X 的概率分布也唯一确定了.

看起来自然的方式是用具有特征 A 的人数比例，即 $n^{-1}\sum_{i=1}^n x_i$，作为对未知 p 的估计值，其中 x_i 是根据该人是否具有特征 A 而取值 1 或 0.在之后，我们会更精确地定义“估计”, 并且使用刚才描

述的模型来量化该估计值与 p 之间的差异，使用“置信区间”.总体和样本比例几乎不会完全相等.置信区间为调查结果中常提及的“误差范围”赋予了精确定义.我们还将确定，当我们研究 1000 人的样本时，误差范围有多大，这在一份调查内是一个常见的样本量.

例 2 测量误差

如果一位物理学家通过实验反复确定一个常数 μ 的值，他不会每次都得到相同的结果.例如，下图显示了 Michelson 在 1882 年对光速的 23 次测定结果.问题是如何从一系列观测值 $x_1, \ldots, x_n$ 中“估计”未知的常数 μ.对于图 1.1 中的观测值，这个估计值将落在 $700 - 900$ 的范围内，但我们不知道确切位置.统计模型可以帮助我们回答这个问题.概率模型首次在 18 世纪末期被应用于此类情境中，Gauss 大约在 1810 年“发现”了正态分布，正是为了深入了解此类情况.



图： Michelson 于 1882 年对光速的 23 次测量结果.水平轴上的刻度给出了测量的光速（以 $\mathrm{km/s}$ 为单位）减去 $299000\ \mathrm{km/s}$ 后的值.

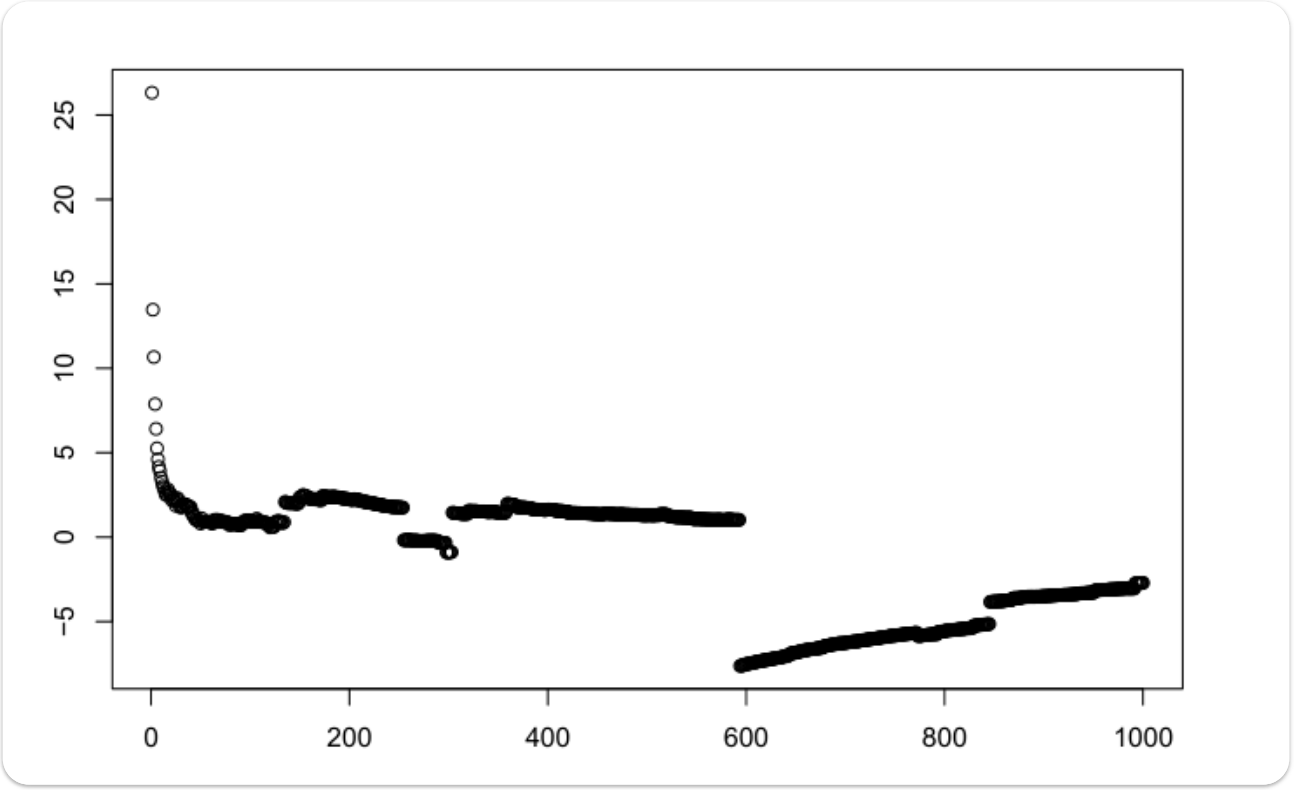
如果所有测量都是在相同的条件下进行的，并且与过去的测量结果独立，那么合理的做法是将这些数值视为独立同分布的随机变量 $X_1, \ldots, X_n$ 的实现.测量误差 $e_i = X_i - \mu$ 也是随机变量.一个常见的假设是测量误差的期望值为 0，即 $\mathrm{E}e_i = 0$，在这种情况下 $\mathrm{E}X_i = \mathrm{E}\left(e_i + \mu\right) = \mu$.由于我们假设 $X_1, \ldots, X_n$ 是独立随机变量并且它们具有相同的概率分布，因此 $X = (X_1, \ldots, X_n)$ 的模型是由 X_i 的统计模型确定(独立同分布时候的边际分布).对于 X_i，建议使用以下模型：具有有限期望 μ 的所有概率分布.对于 X 的统计模型是：所有可能的 $X = (X_1, \ldots, X_n)$ 的概率分布，其中坐标 $X_1, \ldots, X_n$ 是独立同分布且期望为 μ.

物理学家通常相信他们拥有更多的先验信息，并对模型做出更多假设.例如，他们假设测量误差服从期望为 0 且方差为 σ^2 的正态分布，换句话说，观察值 $X_1, \ldots, X_n$ 服从期望为 μ 且方差为 σ^2 的正态分布.此时，统计模型是：$X = (X_1, \ldots, X_n)$ 的所有概率分布，其中坐标是独立的且服从 $N(\mu, \sigma^2)$ 分布.

最终的目标是对 μ 做出一些推断.
在第二个模型中，我们知道得更多，因此我们应该能够更“确定”地说出一些关于 μ 的信息.
另一方面，第二个模型有更高的“概率”是错误的(框定了一个更小的范围,那就更容易出错)，这种情况下，确定性的增加可能是一种虚幻的"更好的方法".

值得一提的是在实际操作中，测量误差通常（但并非总是）大致服从正态分布.如果测量误差可以看作是大量小的独立测量误差（具有有限方差）的总和，则可以使用中心极限定理来证明这种正态性.

我们试图精确定义模型的重要性之一在于它能帮助我们确定从观测值中如何“有意义”地估计 μ.显而易见的选择是取 $x_1, \ldots, x_n$ 的平均值.如果测量误差确实服从期望为 0 的正态分布，这确实是最佳选择（根据某个特定标准）.另一方面，如果测量误差服从柯西分布，则取平均值将是灾难性的.下图展示了这一点，它显示了从标准柯西分布中取样的前 n 次观测值 $x_1, \ldots, x_{1000}$ 的平均值 $n^{-1}\sum_{i=1}^{n} x_i$，其中 $n = 1, 2, \ldots, 1000$.这些平均值的行为非常混乱，并且它们不会收敛到 0.这可以通过一个显著的理论结果来解释：独立标准柯西分布的随机变量 $X_1, \ldots, X_n$ 的平均值 $n^{-1}\sum_{i=1}^{n} X_i$ 也服从标准柯西分布.因此，取平均值并没有改变任何结果!



例 3 库存 (泊松模型)

某种产品在不同零售商之间的销售数量不同，并且随着时间波动.为了估计所需的总产品数量，中央配送中心记录了几周内每周每个零售商的产品销售总量.他们观察到的数据是 $x = (x_{1,1}, x_{1,2}, \ldots, x_{I,J})$，其中 $x_{i,j}$ 是零售商 i 在第 j 周的销售数量.因此，观察值是一个长度为零售商数量与周数之积的向量，具有整数坐标.这些观测值可以视为随机向量 $X = (X_{1,1}, X_{1,2}, \ldots, X_{I,J})$ 的实现(数据的两性,可能是随机变量,也可能是一个确定值,即随机变量实现出的一个具体值).在特定情况下，许多不同的统计模型都是可能且有意义的.

一种常见的（因为它通常拟合效果比较好）模型如下：

- 每个 $X_{i,j}$ 服从未知参数 $\mu_{i,j}$ 的泊松分布.
- $X_{1,1}, \ldots, X_{I,J}$ 是独立的.

这确定了 X 的概率分布，当然期望值 $\mu_{i,j}=\mathrm{E}X_{i,j}$ 是未知的, 其和参数本身有关.不过配送中心关心的就是这些期望值.例如，某一周的总需求的期望值为 $\sum_i \mu_{i,j}$.利用需求 $\sum_i X_{i,j}$ 的泊松特性，配送中心可以选择一个库存量，使得有一定的（较高的）概率库存量充足.

统计分析的目标是从数据中推导出 $\mu_{i,j}$.到目前为止， $\mu_{i,j}$ 是完全“自由”的.由于对于每个 $\mu_{i,j}$ 只有一个观测值 $x_{i,j}$，因此很难从数据中估计它们.合理的做法是通过在 $\mu_{i,j}$ 上添加先验假设来减少统计模型的复杂度.我们可以假设 $\mu_{i,j}=\mu_i$，即它不依赖于 j.这样，所售产品的期望数量取决于零售商，但在时间尺度上保持不变.这时我们剩下 I 个未知数，只要周数 J 足够大，就可以“合理地”从数据中估计这些未知数.更灵活的替代模型包括 $\mu_{i,j}=\mu_i+\beta_i j$ 和 $\mu_{i,j}=\mu_i+\beta\mu_i j$，其中分别有 $2I$ 和 $I+1$ 个参数.两个模型都对应了期望需求随时间线性变化.

例 4 回归模型

高个子的父母通常生出高个子的孩子，而矮个子的父母则生出矮个子的孩子.父母的身高对孩子最终（成人后）的身高有很高的预测价值，但这并不是唯一的影响因素.孩子的性别当然起着重要作用，环境因素如健康的饮食习惯和卫生条件也非常重要.过去 150 年里，营养改善和卫生条件提高，使得阻碍生长的因素如传染病和营养不良在我国减少，因此平均身高有所增加，每一代的孩子都比上一代更高.

孩子的目标身高是基于父母的身高、孩子的性别以及世代之间的身高增加趋势来预期的身高.问题在于，目标身高如何依赖于这些因素.

设 Y 为孩子将达到的身高，设 x_1 和 x_2 分别为生父和生母的身高，x_3 是一个表示性别的指标（女孩为 $x_3=-1$，男孩为 $x_3=1$）.目标身高 $\mathrm{E}Y$ 可以通过所谓的线性回归模型进行建模：

$$\mathrm{E}Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$$

其中 β_0 表示每一代平均身高的增长，β_1 和 β_2 表示父母身高对后代目标身高的影响程度，β_3 表示由于孩子性别导致的目标身高的偏差.由于男性平均比女性高，因此 β_3 会是正值.

上述模型没有对个人身高做出任何说明，只是描述了特定身高父母的后代身高.在这样的假设之下，两个兄弟的目标身高是相同的，因为他们有相同的生物学父母，相同的性别，并且属于同一代.实际的最终身高 Y 可以描述为：

$$Y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+e$$

其中 $e=Y-\mathrm{E}Y$ 是实际最终身高 Y 与目标身高 $\mathrm{E}Y$ 之间的偏差, 这是一个随机偏差.观测值 Y 被称为因变量，而 x_1,x_2 和 x_3 被称为自变量或预测变量.常假设偏差 e 服从期望为 0 且方差为 σ^2 的正态分布.最终身高 Y 服从期望为 $\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$ 且方差为 σ^2 的正态分布.

在我国，青少年的身高增长会定期记录.1997 年，进行了第四次全国身高研究，其中一部分是确定孩子的最终身高与父母身高之间的相关性.为此，收集了青少年及其父母的数据，得到了一些观测值：$(y_1,x_{1,1},x_{1,2},x_{1,3}),\ldots,(y_n,x_{n,1},x_{n,2},x_{n,3})$，其中 y_i 是第 i 个青少年的身高，$x_{i,1}$ 和 $x_{i,2}$ 是其父母的身高，$x_{i,3}$ 是表示第 i 个青少年性别的指标.假设这些观测值是上述线性回归模型的独立重复实验的结果；换句话说，给定 $x_{i,1},x_{i,2}$ 和 $x_{i,3}$，变量 Y_i 的期望值为 $\beta_0+\beta_1x_{i,1}+\beta_2x_{i,2}+\beta_3x_{i,3}$，方差为 σ^2.

这些参数 $(\beta_0,\beta_1,\beta_2,\beta_3)$ 是未知的，可以从观测值中估计出来.为了简化模型的解释，我们选择 $\beta_1=\beta_2=1/2$，因此目标身高等于父母平均身高，经过对孩子性别和时间影响的修正.参数 β_0 和

β_3 分别表示相对上一代身高的增长以及男性和女性平均身高差的一半.这些参数通过最小二乘法估计出来（见示例 3.44）.估计的 β_0 为 4.5 厘米，β_3 为 6.5 厘米.

估计的回归模型如下：

$$Y = 4.5 + \frac{1}{2}(x_1 + x_2) + 6.5x_3 + e$$

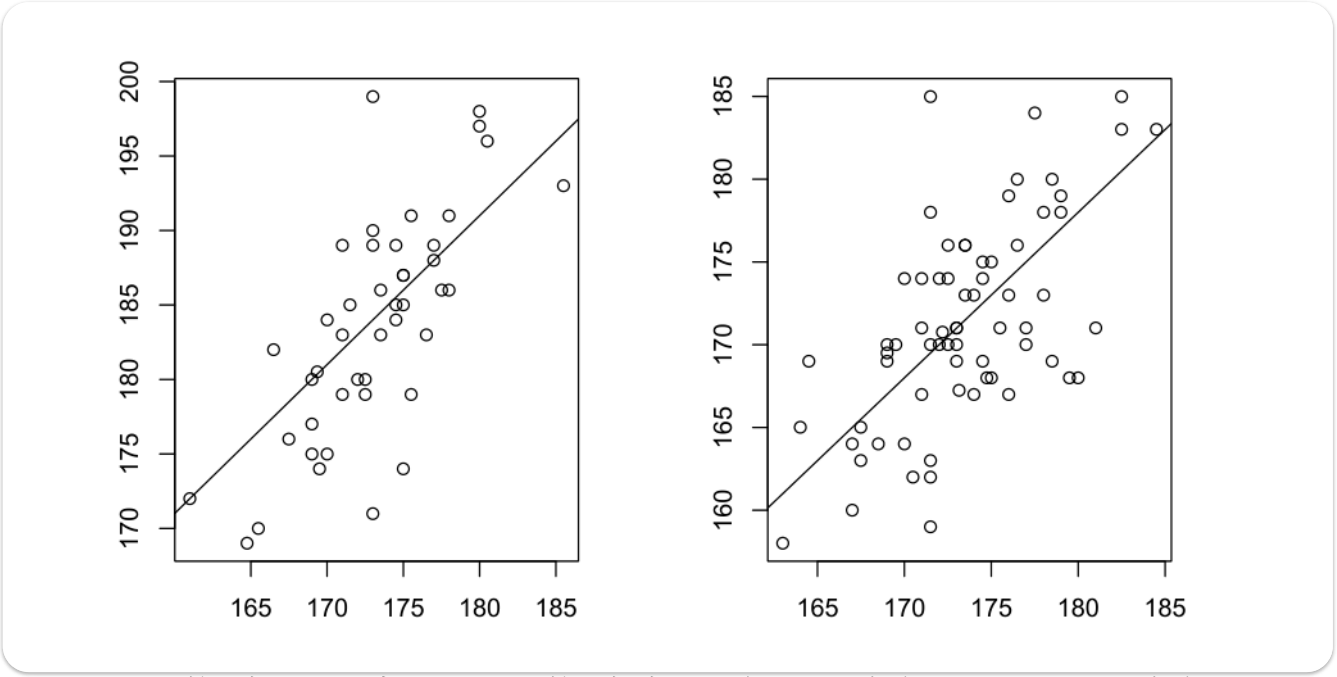


图 显示了 44 位年轻男性（左侧）和 67 位年轻女性（右侧）的身高与其父母的平均身高之间的关系.实线为第四次全国生长研究中的回归线.

我们可以使用第四次全国生长研究中估计出的回归模型预测现在出生的孩子的最终身高.此时，我们必须假设下一代的身高增长仍然为 4.5 厘米，并且男性和女性的平均身高差仍然为 13 厘米.根据上述模型，一个身高 180 厘米的男性和一个身高 172 厘米的女性的儿子和女儿的目标身高分别为 187 厘米和 174 厘米.

在欧洲国家就得使用不同的模型.例如，在瑞士，目标身高为：

$$\mathrm{E}Y = 51.1 + 0.718\frac{x_1 + x_2}{2} + 6.5x_3$$

在这种情况下，同样身高的父母的儿子和女儿的目标身高分别为 184 厘米和 171 厘米.

在上述例子中，响应变量 Y 和未知参数 $\beta_0, \ldots, \beta_3$ 之间有线性相关性, 注意这里重要的是相应变量和参数之间的线性关系而不是和自变量 X.这种情况下，我们称之为线性回归模型.最简单的线性回归模型是仅有一个预测变量的情况：

$$Y = \beta_0 + \beta_1 x + e$$

这被称为简单线性回归模型（与有多个预测变量的多元线性回归模型对比）.

一般来说，当响应变量 Y 和观测值 $x_1, \ldots, x_p$ 之间存在特定的相关性时，我们称之为回归模型：

$$Y = f_\theta(x_1, \ldots, x_p) + e$$

其中 f_θ 描述了观测值 $x_1, \ldots, x_p$ 与响应变量 Y 之间的关系，随机变量 e 是不可观测的测量误差，期望为 0 且方差为 σ^2.如果函数 f_θ 由有限维参数 θ 确定，则我们称之为参数化模型.线性回归模型

就是一个例子；在该模型中，$\theta=(\beta_0,\ldots,\beta_p)\in\mathbb{R}^{p+1}$，并且 $f_\theta(x_1,\ldots,x_p)=\beta_0+\beta_1x_1+\ldots+\beta_px_p$.如果我们知道 θ 和 σ^2 的值，那么这个回归模型就确定了.当然有的时候这个 f_θ 可能有一个有限维的参数 θ 和一个无穷维的参数, 那么这时候, 我们就说这个是一个半参数模型, 其中一个例子就是 Cox 回归, 我们将在之后讨论.

例 5 水位 (极值模型)

在 20 世纪（1910 年到 2000 年期间），在大水镇附近的波里个浪河中测量到了 70 次极端水位."极端"由国家水管理局定义为"超过 1250 m^3/s".这些 70 次极端水流的最大值按时间顺序显示在图中.问题是如何预测未来的情况.国家水管理局特别关心的是：为了确保每 10000 年最多经历一次洪水(也就是说面临万年一遇洪水不奢求挡住, 使用别的赈灾方法, 或者说经历洪水的概率要降低到 $\frac{1}{10000}$)，堤坝高度应该建多高？我们可以使用一个模型，根据水流量计算水位高度.

由于这些极端水流量 $x_1,\ldots,x_{70}$ 是在（大部分）不同年份测量的，并且波里个浪河的水位主要取决于上游山脉及更上游的气候条件，因此将这些数值视为独立随机变量 $X_1,\ldots,X_{70}$ 的实现是不无道理的.假设这些参数也是同分布的, 当然这一点上略有争议，因为在几个世纪以来，波里个浪河的河道（以及气候）逐渐发生了变化，但这一同分布假设通常仍然被采用.然后我们可以将 $X_1,\ldots,X_{70}$ 视为某一变量 X 的独立副本(所谓 $i.i.d.$)，并使用测量的数值 $x_1,\ldots,x_{70}$ 来回答问题.

设 E 表示在某一年发生洪水的事件.事件 E 的概率大约等于每年极端期数的期望值 $\mathrm{E}N$ 乘以在极端期发生洪水的概率，即 $\mathrm{P}(E)\approx\mathrm{E}N\mathrm{P}(X>h)$，这里 $P(E)$ 根据我们之前的假设应该是 $\frac{1}{10000}$ 其中 X 是极端水流期的最大水流量，h 是不发生洪水的最大水流量，N 是任意一年内出现极端水位的次数.此计算利用了极端期内发生洪水的概率 $\mathrm{P}(X>h)$ 很小.

这里 N 的概率分布是未知的，不过一个合理的假设是，N 的期望值大约等于过去 90 年每年极端水流期的平均次数，因此 $\mathrm{E}N\approx 70/90$.现在问题是：我们应取何数值 h 以满足 $\mathrm{P}(X>h)=1/10000\cdot 90/70=0.00013$?

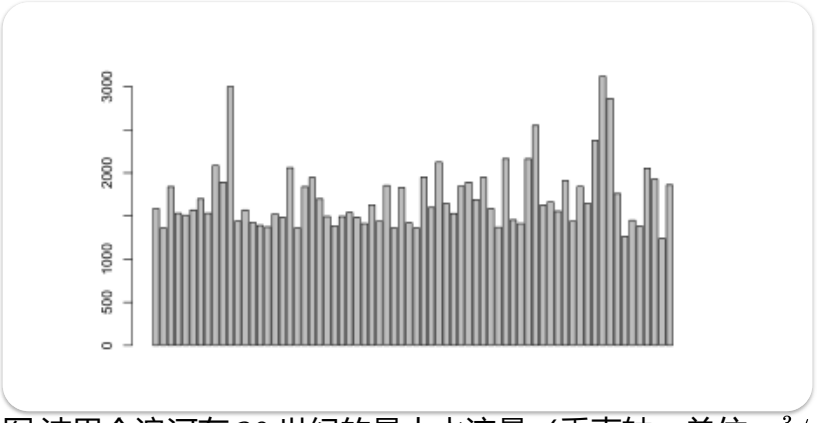


图 波里个浪河在 20 世纪的最大水流量（垂直轴，单位 m^3/s）按时间顺序（水平轴）排列.

这个问题并不容易回答.如果我们可以获得一个 100000 年（或更长时间）周期的观测极值，我们可以通过确定第 10%最高的观测水位（即 $10000/100000$）来相对准确地确定 h.不幸的是，我们只有 70 个观测值，必须进行合理地外推以尽可能长远地预测可能比以往任何测量值都更极端的情况.

还是回到刚刚之前的难题:如果我们能确定一个合适的 X 分布模型，那么这都不会成为问题.例如，如果我们知道 X 服从标准指数分布，那么我们可以通过方程 $0.00013=\mathrm{P}(X>h)=e^{-h}$ 来确定 h，但这是不现实的假设.

另一种方法是将数据用一种"极值分布"去拟合.这些分布通常用于建模看作是最大值的变量 $X = \max(Y_1, \ldots, Y_m)$，其中 $Y_1, \ldots, Y_m$ 是大量独立变量.考虑到 X 作为某一时期的最大水流量的解释，这些分布是合理的.在三种极值分布中，有一种分布能够很好地拟合数据，即 Fréchet 分布，其分布函数为：

$$F(x) = \begin{cases} e^{-((x-a)/b)^{-\alpha}} & \text{if } x \geq a \\ 0 & \text{if } x < a \end{cases}$$

Fréchet 分布有三个参数：$a \in \mathbb{R}$，$b > 0$，以及 $\alpha > 0$.如果我们确信所得到的模型是有用的，我们可以根据这 70 个数据点估计这些参数，并通过简单的计算回答问题.

例 5 生存分析

在生存分析中，我们研究**时间跨度的**概率分布.你可以设想一些常见的例子,比如灯泡的寿命、计算机程序中下一次故障发生的时间（这种一般也叫“可靠性分析”），特别是在医学统计中，直到死亡或疾病发生前剩余的时间.

下面是一个例子.

在心脏瓣膜有泄漏的患者中，通常会将心脏瓣膜替换为生物瓣膜或机械瓣膜.生物瓣膜相对于机械瓣膜的一个缺点是其寿命较短（10 至 15 年）.为了研究生物瓣膜寿命的分布函数 F，我们跟踪了接受此类瓣膜手术的 n 名患者，直到瓣膜必须更换的时刻.我们固定一段试验的年限,例如一项持续了 20 年的研究.

研究结束时，我们有了这 n 个心脏瓣膜的寿命 $t_1, \ldots, t_n$ 的记录.我们将这些数字视为独立随机变量 $T_1, \ldots, T_n$ 的实现，它们的分布函数为 F.瓣膜在 t 年内必须更换的概率 $F(t)$ 可以通过样本中在 t 年内更换的瓣膜比例来估计.

生存分析的一个特殊的方面是，通常并非所有的"寿命"都能观察到.当我们试图分析数据时候会发现，例如，有时候并非所有的瓣膜在我们试验结束前都需要更换，或者患者可能已经死亡，但瓣膜仍然完好.在这些情况下，我们只能观察到寿命的下限，即直到研究结束或患者死亡的时间.我们知道在研究结束或患者死亡时瓣膜仍在工作.我们称之为“删失数据”.

值得注意的要点是,长寿命的数据比短寿命的数据更常被删失，因为患者在长时间内死亡的概率比在短时间内更大（在研究结束前也没有损坏的情况也类似）.因此，忽略删失数据并仅基于未删失数据来估计分布函数 F 是错误的.这将导致过高估计寿命分布函数，低估预期寿命，因为相对较长的寿命这一整个部分都将被忽略.正确的方法是对所有观测值（包括删失和未删失的）使用统计模型.

如果我们怀疑某些因素可能影响心脏瓣膜的寿命，例如患者的年龄、体重或性别，那么统计模型会更加复杂.在这种情况下，可以使用 Cox 回归模型对寿命进行建模.

例 6 选择偏差

为了正确回答一个研究问题，确保问题、收集的数据和统计模型的正确对齐和匹配非常重要.以下例子对此进行了说明.

福克斯(FOX)铁路公司经常收到关于高峰时段火车拥挤的投诉.为此进行了一项研究，以调查这些投诉是否合理.

研究关心两个问题.第一个问题是高峰时段有多少乘客没有座位.第二个问题是高峰时段有多少列车过于拥挤.

这两个问题从根本上是不同的.第一个问题关心的是人，即乘客的百分比，而第二个问题关心的是火车.一个乘客可能只关心第一个问题，而公司则对第二个问题的答案更为关心，因为他们需要确定在哪些列车上存在问题，以便采取措施.

为了解答第一个问题，从刚下火车的乘客中随机选择了 50 名乘客，询问他们是否有座位.我们观察到序列 $x_1, \ldots, x_{50}$，其中，如果第 i 个乘客没有座位，则 $x_i = 1$，如果有座位，则 $x_i = 0$.于是，$x_1, \ldots, x_{50}$ 是独立随机变量 $X_1, \ldots, X_{50}$ 的实现，X_i 服从参数为 p 的伯努利分布，其中 $p = \mathrm{P}(X_i = 1)$ 是无法找到座位的乘客的比例.与例 1 类似，我们可以使用样本均值 $50^{-1} \sum_{i=1}^{50} x_i$ 来估计比例 p.这是回答该研究问题的正确方法.

回答第二个问题则更加困难，因为这个问题关心的是列车而不是乘客.为进行这项研究，在高峰时段随机选择了 50 名列车长，并询问他们所乘的列车是否过于拥挤.我们观察到序列 $y_1, \ldots, y_{50}$，其中，如果第 i 个列车长表示列车过于拥挤，则 $y_i = 1$，否则 $y_i = 0$.同样，我们可以将 $y_1, \ldots, y_{50}$ 视为独立伯努利变量 $Y_1, \ldots, Y_{50}$ 的实现，拥挤列车的比例 $q = \mathrm{P}(Y_i = 1)$.如果假设每列车上只有一名列车长，则 q 等于高峰时段拥挤列车的比例.我们可以将 $Y_1, \ldots, Y_{50}$ 视为从刚进站的列车中抽取的样本，通过样本均值 $50^{-1} \sum_{i=1}^{50} y_i$ 来估计 q.

那为什么我们不拿回答第一个问题所抽取的乘客样本, 来回答列车是否拥挤这第二个问题呢，这似乎更加简单.让我们来仔细研究一下这种情况，我们观察到独立伯努利变量 $Z_1, \ldots, Z_{50}$ 的实现，$r = \mathrm{P}(Z_i = 1)$.这里 Z_i 的定义类似于 Y_i.不过这里有一个严重的问题,由于一列列车搭载的乘客不止一人，因此不是每个乘客都会对应唯一的列车.由于拥挤的列车上乘客更多，"来自拥挤列车的乘客" 在乘客总体中所占的比例将远高于"拥挤列车"在列车总体中所占的比例.换句话说，r 会大于 q.在不做额外假设的情况下，很难给出 r 和 q 之间的关系.因此，基于乘客样本回答第二个研究问题是很困难的，而回答第一个问题则较为简单.

总结

在上述大多数示例中，统计模型通过一个或几个参数进行参数化，例如 p, (μ, σ^2), $(\beta_0, \beta_1, \beta_2, \beta_3)$，或者 (a, b, α).许多统计模型都是部分未知的, 也就是说除了某几个参数以外其他部分都是已知.在本书中，我们通常用 θ 表示该参数.统计模型可以表示为 $\{P_\theta : \theta \in \Theta\}$，其中 P_θ 是观察值 X 的概率分布，Θ 是可能参数的集合.隐含的假设是，有且仅有一个参数值（或模型中的一个元素）给出了 X 的"真实"分布.统计学的目的就是找到这个值.为什么需要统计学的动机和让统计学变得困难的问题是，我们永远无法完全成功找到真实分布，关于真实参数值的数据总是包含一定的不确定性(随机性).

习题

1. 假设我们在美国街头从某一人群中随机选择了 n 人，并询问他们的政治倾向.设 X 表示样本中属于 A 政党的人数. 需要在此留意的是该人群中属于 A 政党的比例是未知的概率 p.描述一个对应的统计模型.给出一个直观上合理的 p 的估计值.
2. 假设随机选择了 $m + n$ 名高血压患者，并任意分为两组，组别大小分别为 m 和 n.第一组为“治疗组”，给予一种特定的降压药物；第二组为“对照组”，给予安慰剂.每个患者的血压在用药或服用安慰剂前后都被测量，并得出血压的差值，得到观测值 $x_1, \ldots, x_m$ 和 $y_1, \ldots, y_n$.

(i) 制定一个合适的统计模型.

(ii) 基于这些观测值，给出药物对血压影响的直观合理估计（多种答案可能都合理）.

3. 我们希望估计池塘中的鱼的数量 N.我们采取如下步骤：捕捉 r 条鱼并将它们标记.然后将它们放回池塘.过一段时间，我们再次捕捉 n 条鱼（不放回），其中 X 条是标记过的.将 r 和 n 视为我们自行选择的常数，并将 X 作为观测值.

   (i) 制定一个合适的统计模型.

   (ii) 基于观测值给出 N 的直观合理估计.

   (iii) 如果我们在第二次捕捉鱼时直接将它们放回（有放回的抽样），重新回答上述问题.

4. 在评估一批商品时，我们一直进行检验，直到发现 3 个次品为止.

   (i) 制定一个合适的统计模型.

   (ii) 第 3 个次品是在我们检查的第 50 个商品时发现的.给出该批次商品中次品比例的估计，并说明理由.

5. 邮局的客户人数似乎取决于星期几（工作日或星期六）和时段（上午或下午）.在工作日，邮局上午和下午都开放；在星期六，它只在上午开放.为了确定需要多少员工提供及时的服务，记录了十周内每天上午（工作日和星期六）和下午（仅工作日）邮局的客户数量.

   (i) 制定一个合适的统计模型.

   (ii) 给出星期一下午客户数量的直观合理估计，并说明理由.

   (iii) 最大的客户数量差异出现在工作日的上午和下午时段以及星期六的上午时段.因此，决定只在员工安排中考虑这一差异.重新制定统计模型并给出新的估计值.

6. 非洲城市 Masvinguigui 的年度用水需求大于一年中从降水中回收的水量.因此，根据需要从附近的湖泊供应水.每年的供水量取决于该年降水量和 Masvinguigui 的总人口.此外，富人用水量比穷人更多.描述一个以“供水量”为因变量，“人口数量”“降水量”和“平均收入”为自变量的线性回归模型.分别指出每个参数是正相关还是负相关.

7. 怀疑个人的收入与年龄和教育水平（低、中、高）之间存在线性相关性.

   (i) 描述一个以“收入”为因变量，以“年龄”和“教育水平”为自变量的线性回归模型.仔细考虑如何将“教育水平”变量包含在模型中.

   (ii) 我们想研究性别是否对收入有影响.调整线性回归模型以便研究这一点.

8. 我们希望估计一个大箱子中羊毛纤维的平均长度.箱子首先被充分摇匀，然后我们闭着眼睛从中取出预定数量的羊毛纤维.我们将箱子中羊毛纤维的平均长度估计为样本中羊毛纤维的平均长度.估计的长度是系统性偏长、系统性偏短还是刚好合适？

9. 在一个呼叫中心，我们希望估计客户等待服务的时间.我们记录下某天每位客户的等待时间.如果客户失去耐心并挂断电话，我们记录他们到挂断时的等待时间.之后，我们通过计算所有记录时间的平均值来估计新客户的等待时间.你对这种方法有何看法？