描述性统计

统计模型是我们对**产生观察数据的概率试验**的先验知识的表达。

当然，根据之前的定义我们知道所谓的统计模型假设观察值 X 是由模型中的某个概率分布生成的。

那么我们如何找到一个好的模型？在某些情况下，模型可以从实验的设置方式中明确得出。例如，如果在民意调查中，样本是从一个定义明确的人群中随机且有放回地抽取的，那么二项分布是不可避免的。如果观察结果涉及发射的放射性粒子数量，根据放射性的物理理论，泊松分布是正确的选择。有时实验与过去的实验非常相似，经验会建议某个特定的模型。当然，统计模型的选择并非总是没有争议的。

至少，所选择的模型必须经过验证。在某些情况下，这在估计模型参数之前进行，而在其他情况下则在之后进行。这些方法不仅应用于数据本身，通常也应用于例如回归模型的残差。在描述性统计中我们将会讨论这一话题.

对于从未知分布中抽取的单变量样本 $X_1, \ldots, X_n$ 的定量度量：

- 样本均值提供了关于基础分布位置的信息：

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- 样本方差提供了关于基础分布离散度的信息：

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

样本标准差是 S_X，即 S_X^2 的平方根。

- 其他常用的度量包括样本中位数和样本四分位距。

对于从未知分布中抽取的单变量样本 $X_1, \ldots, X_n$ 的图形方法：

- 直方图提供了关于基础分布形态的一个直观印象。
- 箱线图展示了中位数、四分位区间以及样本中的异常值，它能够给出基础分布的位置信息、尺度信息以及分布的对称性和尾部厚度。
- QQ 图展示了样本分位数与选择的分布的分位数的对应关系。如果选择的分布与基础分布属于同一个位置尺度族，则点将近似落在一条直线上。

对于从未知分布中抽取的双变量样本 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 的图形方法与定量度量：

- 散点图是展示坐标相关性的图形表示。
- 样本相关系数是坐标线性相关性的定量度量：

$$r_{X,Y}=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{(n-1)\sqrt{S_X^2}\sqrt{S_Y^2}}$$

对于从未知分布中抽取的（可能依赖的）观测序列 $X_1,\ldots,X_n$ 的定量度量：

- 阶数为 h 的样本自相关系数用于寻找观测值之间可能的（时间）依赖性：

$$r_X(h)=\frac{\sum_{i=1}^{n-h}\left(X_{i+h}-\bar{x}\right)\left(X_i-\bar{x}\right)}{(n-h)S_X^2}$$

习题

1. 令 h_n 为从具有密度 f 的分布中抽取的样本 $X_1,\ldots,X_n$ 的缩放直方图。直方图的分割由 $a_0<a_1<\ldots<a_m$ 给出。证明对于 $a_{j-1}<x\leq a_j$，当 $n\to\infty$ 时，$h_n(x)\to\left(a_j-a_{j-1}\right)^{-1}\int_{a_{j-1}}^{a_j}f(s)ds$ 概率为 1。
2. 设 X 是具有分布函数 F 和分位数函数 Q 的随机变量。定义 x_α 为 F 的 α -分位数，y_α 为分布为 $Y=a+bX$ 的分布的 α -分位数。
   (i) 假设 F 严格递增且连续，因此 F 的反函数存在并等于 Q。利用 F 的可逆性，证明 $x_\alpha=F^{-1}(\alpha)$ 和 $y_\alpha=F_{a,b}^{-1}(\alpha)$ 之间存在线性相关性。
   (ii) 证明对于一般的分布函数 F，x_α 和 y_α 之间也存在线性相关性。使用 α -分位数的一般定义。
3. 标准指数分布在 $[0,\infty)$ 上的分布函数为 $x\mapsto 1-e^{-x}$。
   (i) 带有参数 λ 的指数分布是否属于与标准指数分布相关的位置尺度族？
   (ii) 在位置尺度族 $F_{a,b}$ 中，将参数 a 和 b 用带有分布 $F_{a,b}$ 的随机变量的期望值和方差表示。
4. 设 X 是在 $[-3,2]$ 上服从均匀分布的随机变量。
   (i) 确定 X 的分布函数 F。
   (ii) 确定 X 的分位数函数 F^{-1}。
5. 设 X 是具有概率密度

$$f(x)=\frac{2}{\theta^2}x1_{[0,\theta]}(x)$$

的随机变量，其中 $\theta>0$ 是常数。
(i) 确定 X 的分布函数 F。
(ii) 确定 X 的分位数函数 F^{-1}。
7. 设 $X_1,\ldots,X_n$ 是连续分布的样本，其分布函数为 F，密度为 f。证明第 k 个顺序统计量 $X_{(k)}$ 的概率密度为：

$$f_{(k)}(x)=\frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}(1-F(x))^{n-k}f(x)$$

通过首先确定 $X_{(k)}$ 的分布函数来证明。（提示：如果至少有 k 个观测值 X_i 小于或等于 x，则我们有 $X_{(k)}\leq x$。小于或等于 x 的 X_i 的数量服从参数为 n 和 $\mathrm{P}\left(X_i\leq x\right)$ 的二项分布。）
8. 设 $X_1,\ldots,X_n$ 是连续分布 F 的样本。在本练习中，我们要证明 $\mathrm{E}F\left(X_{(k)}\right)=k/(n+1)$。定义 $U_i=F\left(X_i\right)$ 对于 $i=1,\ldots,n$。
(i) 证明随机变量 $U_1,\ldots,U_n$ 组成了 $[0,1]$ 上均匀分布的样本。
(ii) 证明 $U_{(k)}$ 的分布函数 $F_{(k)}$ 为：

$$F_{(k)}(x)=\sum_{j=k}^{n}\binom{n}{j}x^{j}(1-x)^{n-j}$$

(iii) 证明 $U_{(k)}$ 的密度 $f_{(k)}$ 为：

$$f_{(k)}(x)=\frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}$$

(iv) 证明 $\mathrm{E}U_{(k)}=k/(n+1)$。

9. 绘制 $N\left(2,2^{2}\right)$ 分布的分位数相对于 $N\left(0,3^{2}\right)$ 分布的分位数的图。这条线是什么？

10. 设 X 是标准正态随机变量。计算随机变量 X 和 $Y=X^{2}$ 之间的相关系数。

11. 解释为什么当 n 很大时，样本相关系数 $r_{X,Y}$ 近似等于相关系数 ρ 是合理的。

12. 假设 X 和 Y 是独立的，并且都服从标准正态分布。计算 X 和 $Z=X+Y$ 之间的相关系数。