

统计学眼中的线性回归模型

假设检验与置信区间

小胖

目录

ONE 重新审视模型

从随机的角度理解线性回归模型

TWO 假设检验与置信区间

控制模型随机性的利器

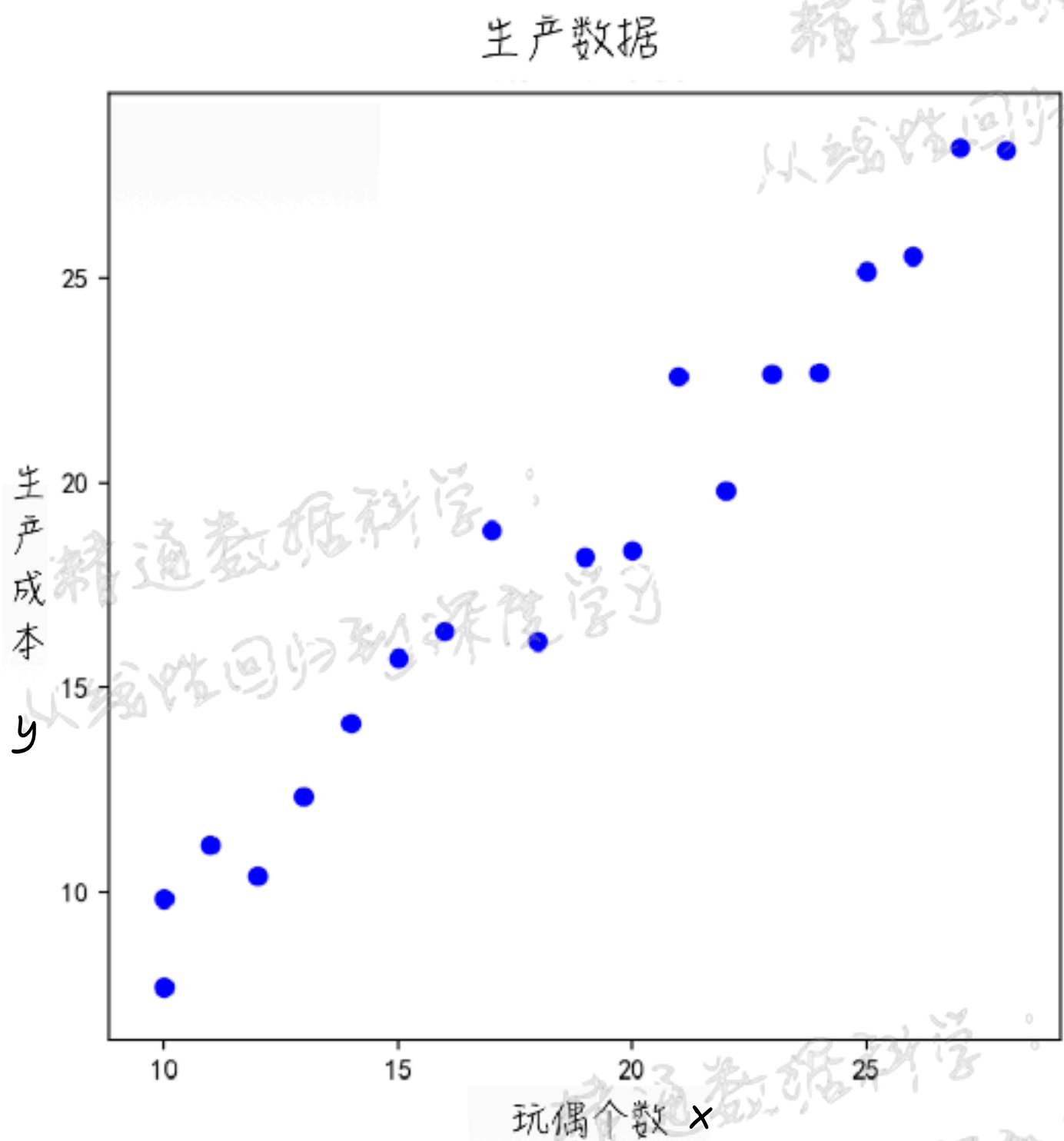
THREE 代码实现

废话少说，放码过来

重新审视模型

条件概率

站在统计学的角度出发，我们试图
弄清楚变量 y 和 x 之间的数学关系



x 与 y 之间可近
似为线性关系

生产记事本

| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|-------|-----|
| 04/01 | 10 | 7.7 | 1 |
| 04/02 | 10 | 9.87 | 2 |
| 04/03 | 11 | 10.87 | 3 |
| 04/04 | 12 | 12.18 | 4 |
| 04/05 | 13 | 11.43 | 5 |
| 04/06 | 14 | 13.36 | 6 |
| 04/07 | 15 | 15.15 | 7 |
| 04/08 | 16 | 16.73 | 8 |
| 04/09 | 17 | 17.4 | 9 |
| ... | ... | ... | ... |

生产记事本

| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|------|-----|
| 04/01 | 10 | 7.7 | 1 |
| 04/02 | 10 | 9.87 | 2 |

↓

$(x_1=10, y_1=7.7)$
 $(x_2=10, y_2=9.87)$

变量 y 似乎带有
某种随机性

于是假设： $y_i = ax_i + b + \varepsilon_i$
其中 ε_i 表示随时扰动项

重新审视模型

条件概率

$$y_i = ax_i + b + \varepsilon_i$$

在上面公式的基础上，进一步假设：

- 1

ε_i 服从正态分布

(期望等于0, 方差等于 σ^2)
- 2

ε_i 之间相互独立

?
- 3

ε_i 与 x_i 相互独立

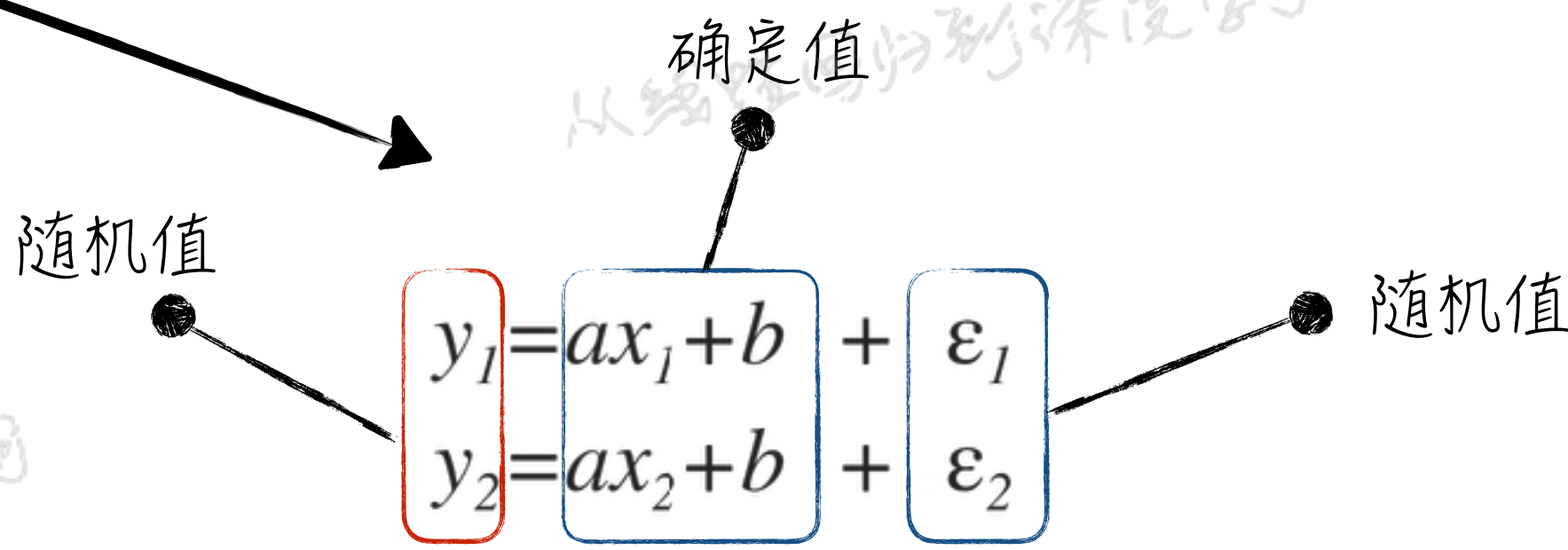
?

生产记事本

| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|------|-----|
| 04/01 | 10 | 7.7 | 1 |
| 04/02 | 10 | 9.87 | 2 |

$(x_1=10, y_1=7.7)$
 $(x_2=10, y_2=9.87)$

+ a, b, σ 已知



$y_1=7.7 \sim N(10a+b, \sigma^2)$
 $y_2=9.87 \sim N(10a+b, \sigma^2)$

相同的玩偶个数，不同的成本。
因为成本 y_1, y_2 是同一正态分布的两次独立观测值

重新审视模型

参数估计公式

根据上面的模型假设

$$y_i = ax_i + b + \varepsilon_i$$

- 变量y是随机变量
- y_i 之间相互独立, 而且都服从正态分布

由于y是随机变量, 定义参数的似然函数 (likelihood function)

$$L = P(Y|a, b, X, \sigma^2)$$

- 似然函数其实就是y的联合条件概率
- y_i 相互独立, 因此可以将似然函数改写如下

$$L = \prod P(y_i|a, b, x_i, \sigma^2)$$

既然y是随机变量, 那么模型参数估计的原则是y出现的概率达到最大

- 这就是最大似然估计法 (Maximum Likelihood Estimation, MLE)

$$(\hat{a}, \hat{b}) = \operatorname{argmax}_{a,b} L$$

$$(\hat{a}, \hat{b}) = \operatorname{argmax}_{a,b} \ln L$$

$$\ln L = -0.5n \ln(2\pi\sigma^2) - (1/2\sigma^2) \sum_i (y_i - ax_i - b)^2$$

与机器学习里面线性回归模型
(OLS) 的参数估计公式一致

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{a,b} \sum_i (y_i - ax_i - b)^2$$

目录

ONE

重新审视模型

从随机的角度理解线性回归模型

TWO

假设检验与置信区间

控制模型随机性的利器

THREE

代码实现

废话少说，放码过来

假设检验与置信区间

参数估计值的分布

$$(\hat{a}, \hat{b}) = \operatorname{argmin}_{a,b} \sum_i (y_i - ax_i - b)^2$$

根据模型参数的估计公式，可以得到参数a, b的估计值：

- 使用不同的数据训练模型，得到不同的参数估计值
- 数学上可以证明，参数估计值本身也是随机变量，而且服从正态分布

| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|-------|-----|
| 04/01 | 10 | 7.7 | 1 |
| 04/02 | 10 | 9.87 | 2 |
| 04/03 | 11 | 10.87 | 3 |

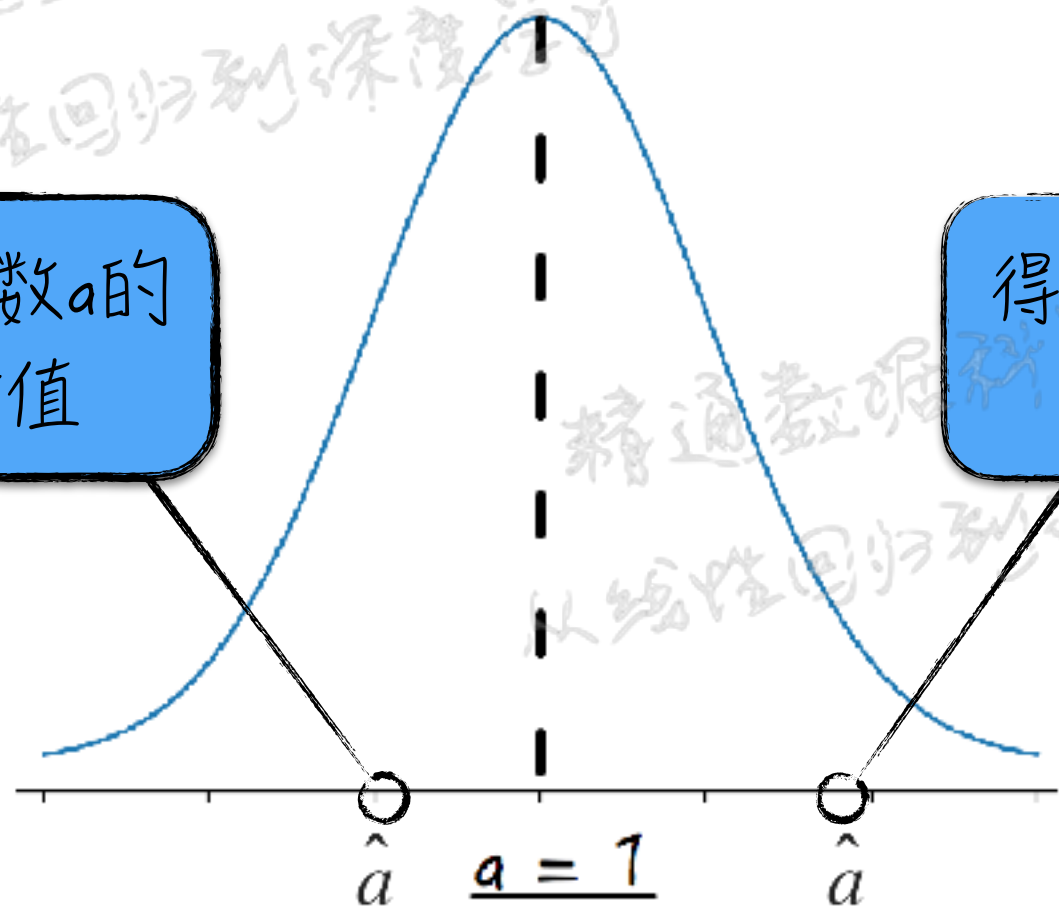
| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|-------|-----|
| 04/04 | 12 | 12.18 | 4 |
| 04/05 | 13 | 11.43 | 5 |
| 04/06 | 14 | 13.36 | 6 |

模型计算

得到参数a的估计值

模型计算

得到参数a的估计值

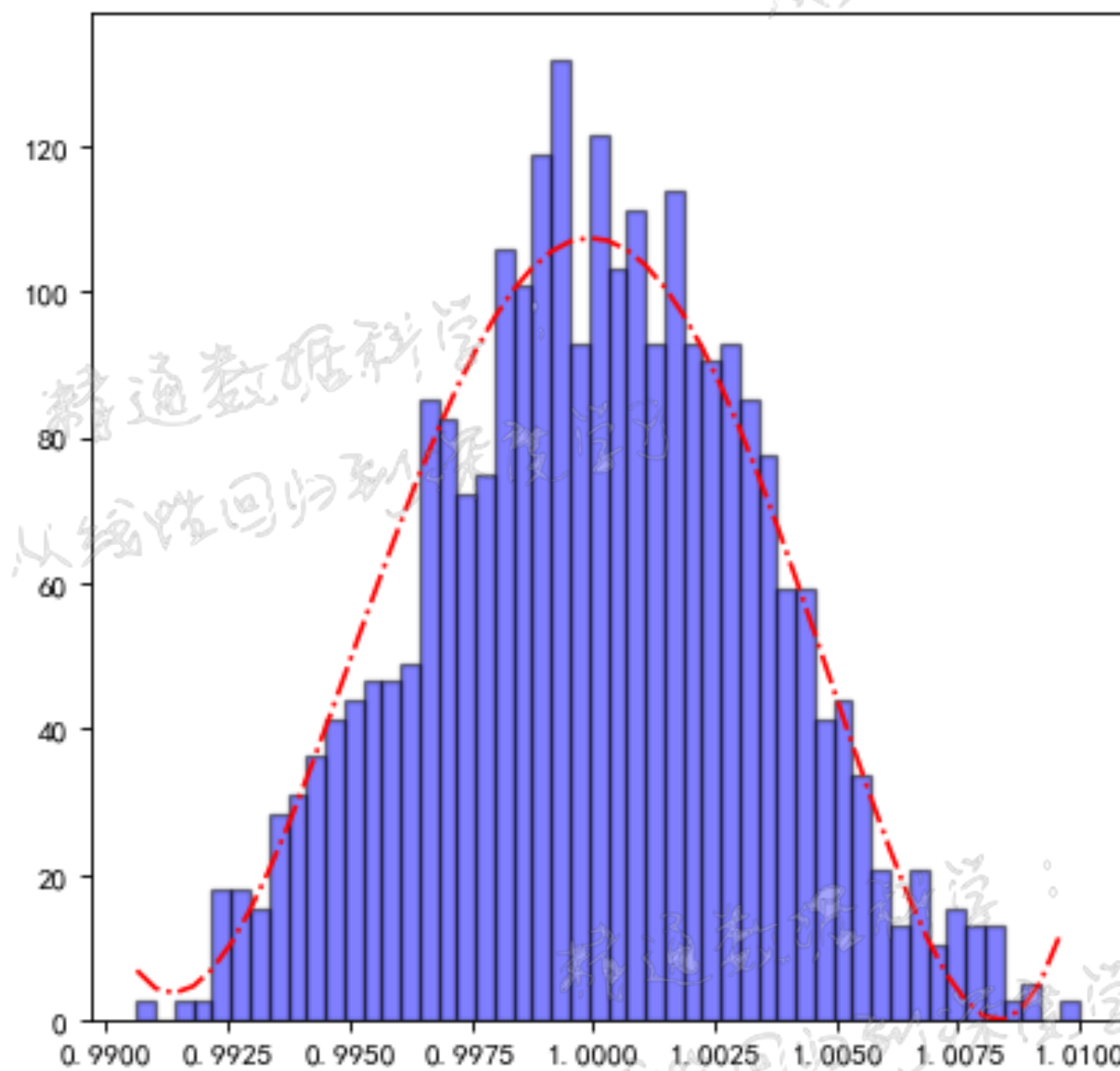


模型参数的估计值只是一个随机变量，具体数值依赖于使用的数据

假设检验与置信区间

实证例子

参数a的估计值分布



随机生成训练数据, x 和 y

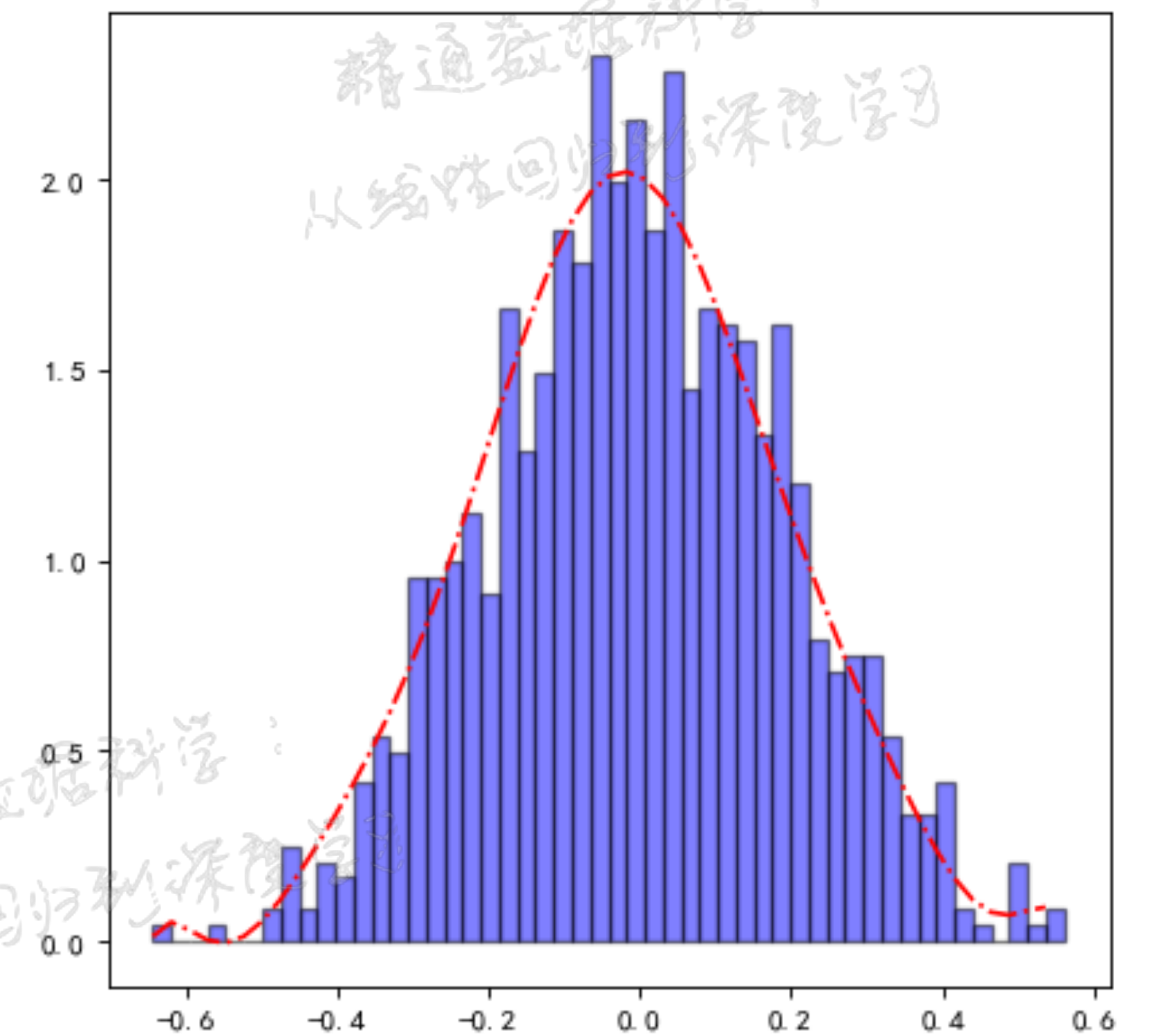
使用scikit-learn, 训练模型

循环1000次

记录参数 a 、 b 的估计值

使用matplotlib, 将模型结果可视化

参数b的估计值分布



假设检验与置信区间

置信区间

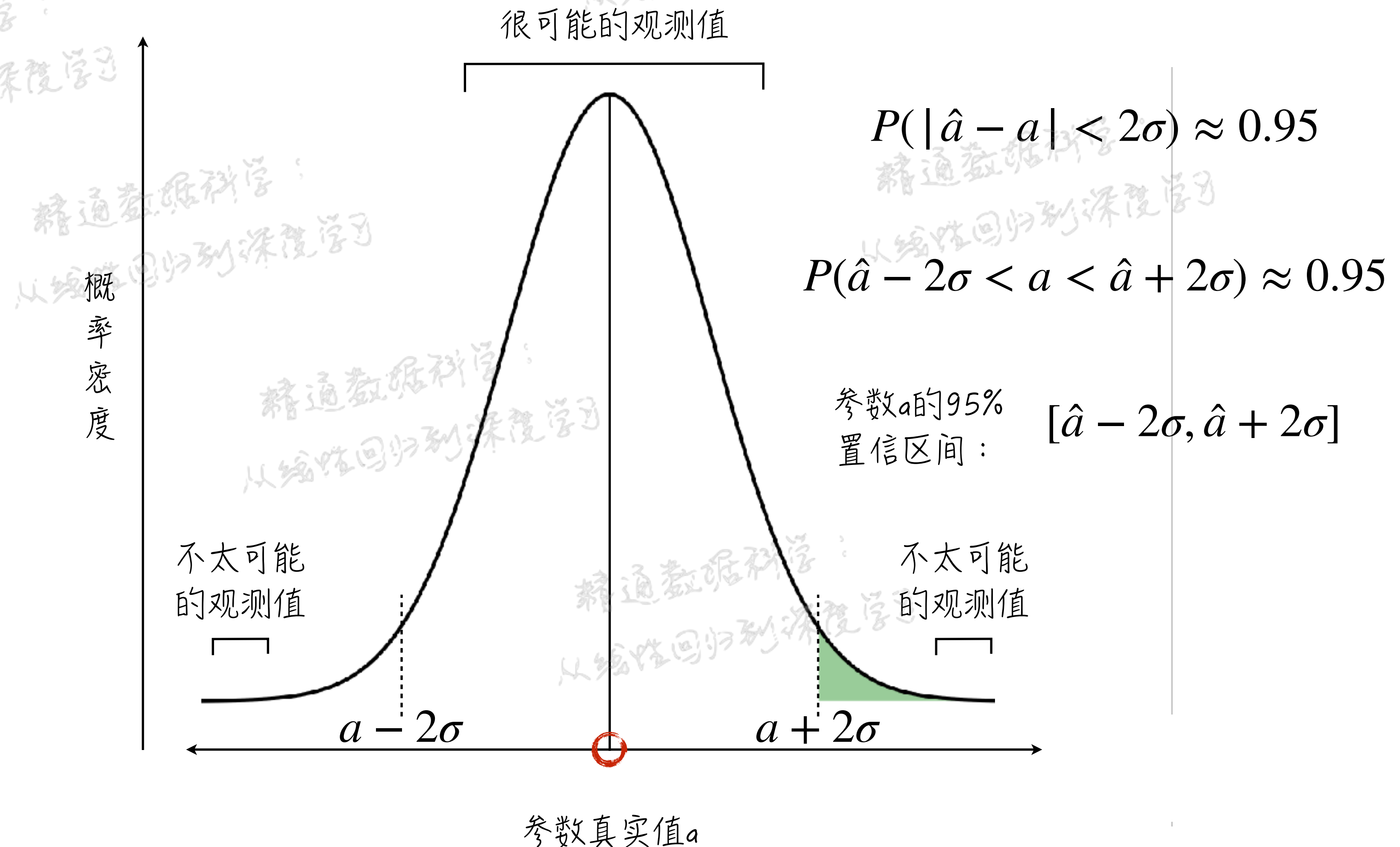
估计值 \hat{a} 服从以真实值为期望的正态分布

既然得到估计值只是随机变量的一次观测值

- 更关心估计值离真实值有多远?
- 解决方案: 定义参数真实值的置信区间

95%的置信区间表示

- 重复100次的模型训练, 并按公式得到置信区间, 那么有95次, 参数 a 的真实值将落在这个区间里
- 可以“通俗地”理解为参数 a 的大致取值范围

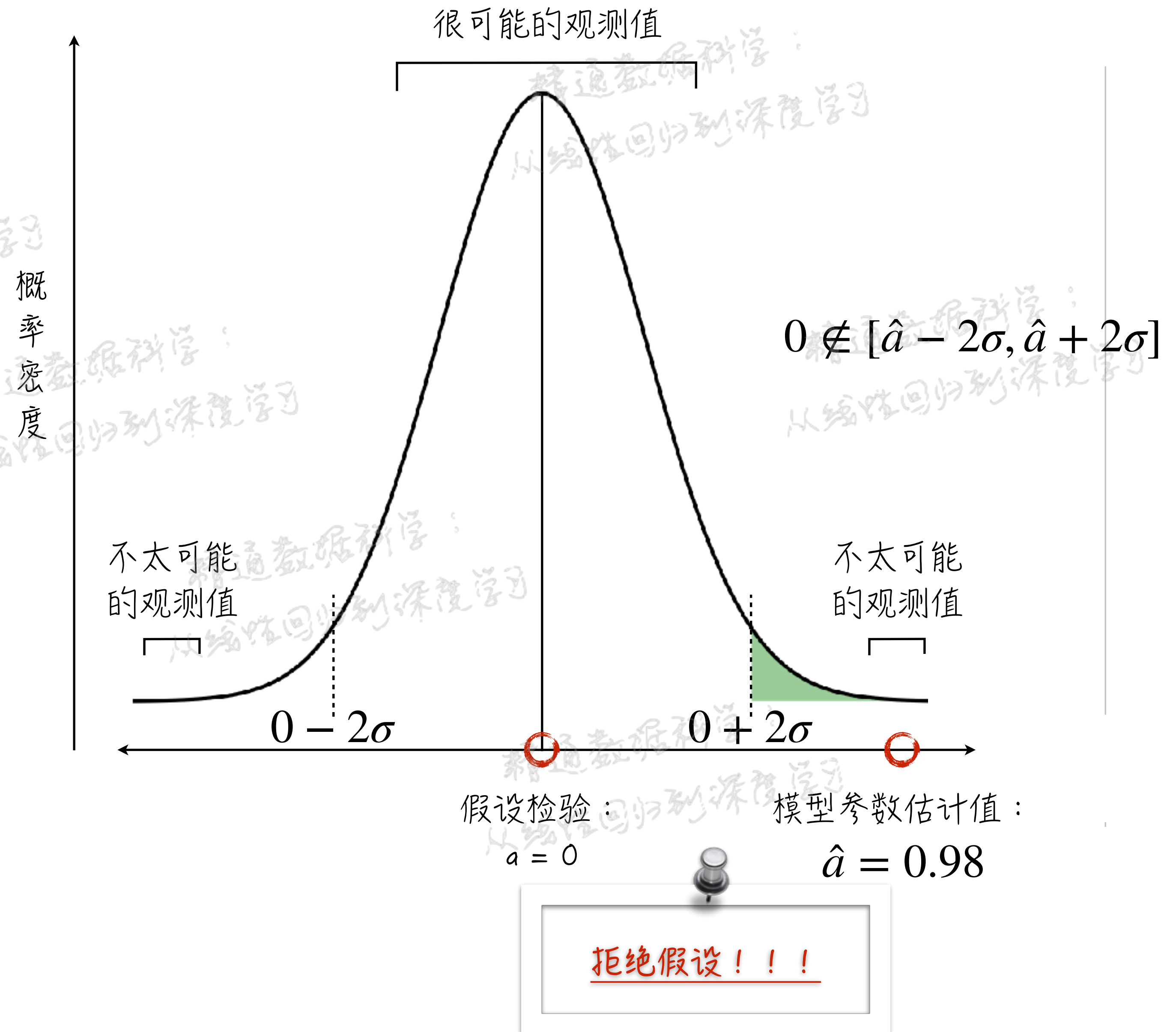


假设检验与置信区间

假设检验

除了置信区间外，还可以使用假设检验来得到更有把握的结果

- 对单个参数的假设检验与置信区间比较类似
- 也可以对多个参数做组合的假设检验



精通数据科学：
从线性回归到深度学习

THANK YOU