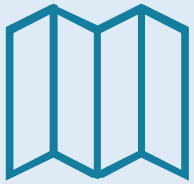


# 带你走入统计学

统计与数据科学学院 马俊玲

6

# 专题六
# 探索数据中的统计规律

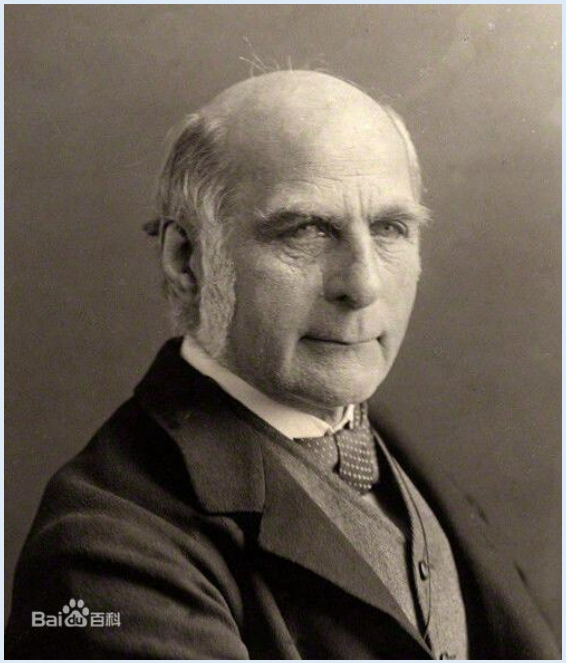
# 一、父代和子代身高的关系

## （一）背景介绍

1. 高尔顿和正态分布

高尔顿是一个“凯特莱主义者”，对正态分布怀有特殊的兴趣。

他在1908年发表的回忆录（《Memories of My Life》）中说，他最初接触凯特勒拟合正态曲线的方法是在1863年。在其后几年间，他使用各种数据，包括身高、胸围以至考试成绩等，结果都符合得很好。




弗朗西斯·高尔顿（Francis Galton）
英国科学家和探险家。

因此，他在1869年出版的一部著作中发表了与凯特勒一样的观点：与正态曲线拟合得好是数据同质性的可靠标志。

同质性：个体趋向于与其他相似的人交往和发展关系的一种倾向，就像谚语“物以类聚，人以群分”。

# （一）背景介绍

2. 高尔顿对遗传问题的思考

从19世纪80年代高尔顿就开始思考父代和子代是否相似，是否有同质性，如身高、性格及其它种种特制的相似性问题。

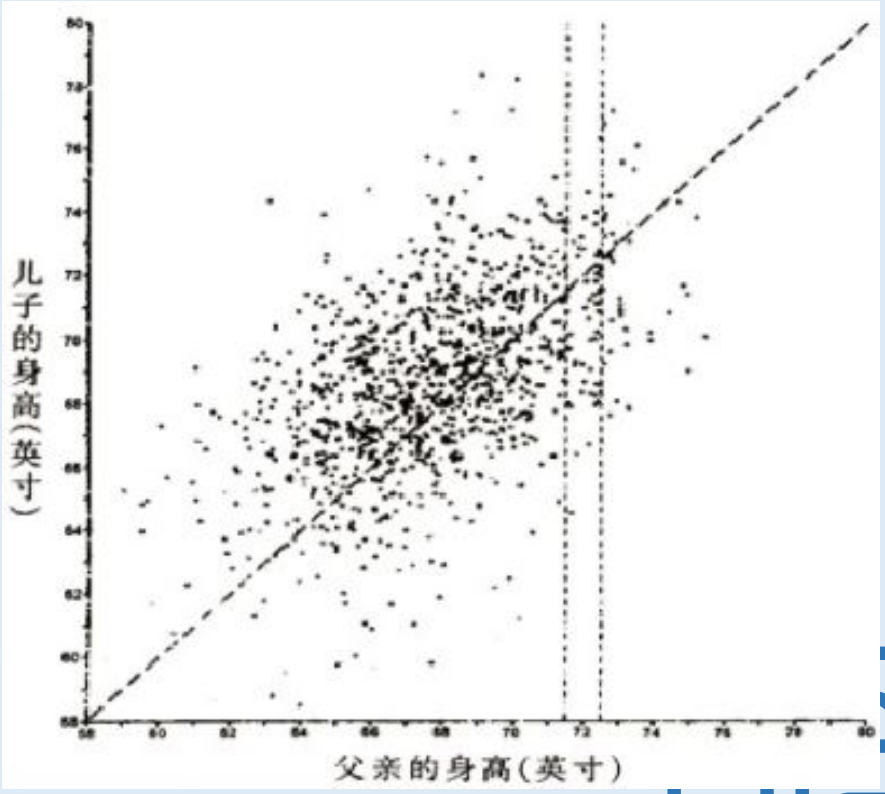
于是他选择了父母身高X与其子身高Y的关系作为研究对象。

## （二）案例问题

父母和子代身高是否存在规律？

## （三）案例分析

高尔顿观察了1074对父母及每对父母的一个儿子，将结果描成散点图，发现趋势近乎一条直线。
(如右图所示)



## （三）案例分析

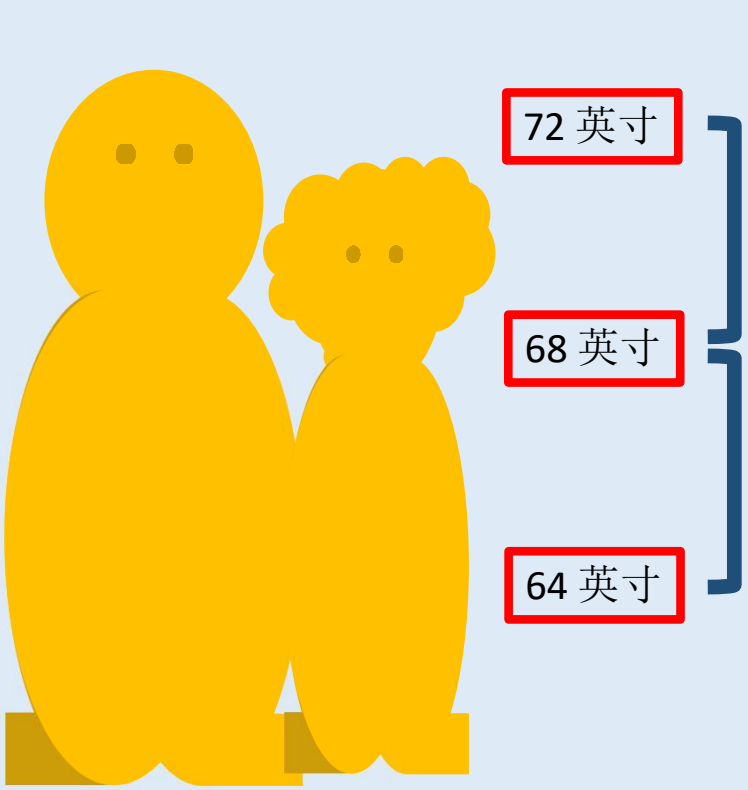
172.7cm

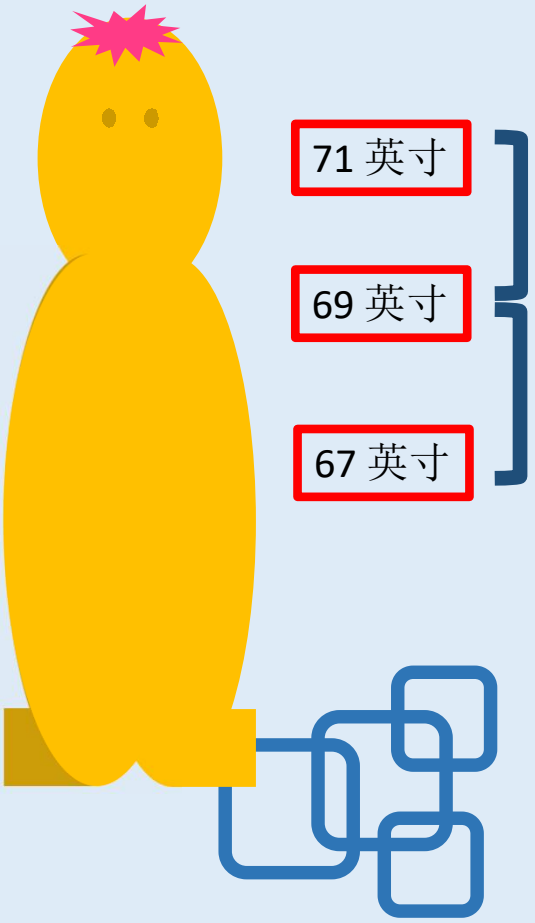
高尔顿发现这1074对父母平均身高的平均值为68 英寸（英国计量单位，1 英寸=2.54cm）时，1074个儿子的平均身高为69 英寸，比父母平均身高大1 英寸。

他推想，当父母平均身高为64 英寸时，1074个儿子的平均身高应为64+1=65 英寸；若父母的身高为72 英寸时，他们儿子的平均身高应为72+1=73 英寸，但观察结果却与此不符。

## （三）案例分析

高尔顿发现前一种情况是儿子的平均身高为67 英寸，高于父母平均值达3英寸，后者儿子的平均身高为71 英寸，比父母的平均身高低1 英寸。




回归模型

# （三）案例分析

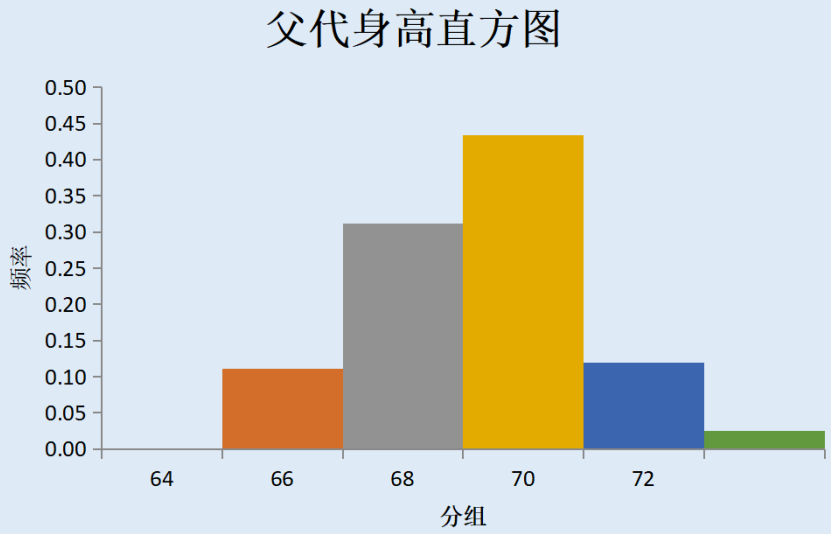
高尔顿数据部分展示

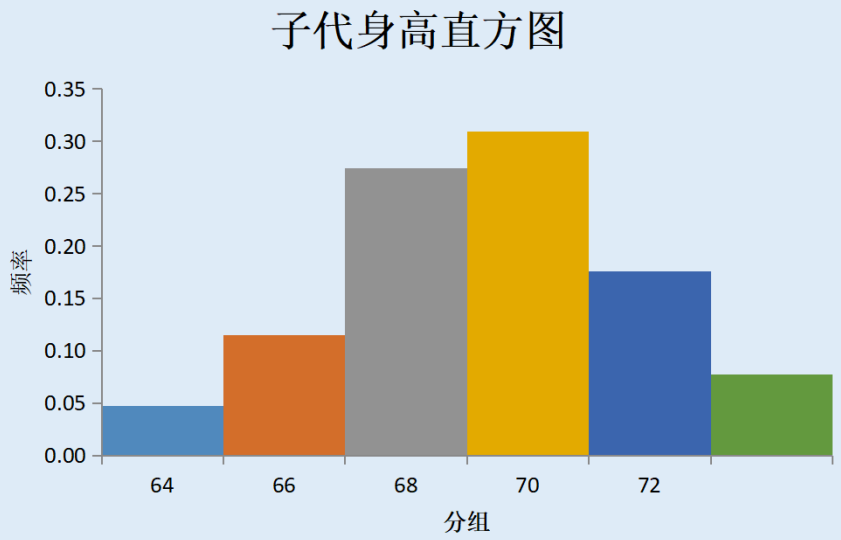
| Galton data | | |
|---|---|---|
| | child | parent |
| 1 | 61.7 | 70.5 |
| 2 | 61.7 | 68.5 |
| 3 | 61.7 | 65.5 |
| 4 | 61.7 | 64.5 |
| 5 | 61.7 | 64 |
| 6 | 62.2 | 67.5 |
| 7 | 62.2 | 67.5 |
| 8 | 62.2 | 67.5 |
| 9 | 62.2 | 66.5 |
| 10 | 62.2 | 66.5 |
| 11 | 62.2 | 66.5 |
| 12 | 62.2 | 64.5 |
| 13 | 63.2 | 70.5 |
| 14 | 63.2 | 69.5 |
| 15 | 63.2 | 68.5 |
| 16 | 63.2 | 68.5 |

高尔顿身高数据描述性分析

| child | | parent | |
|---|---|---|---|
| | | | |
| 平均 | 68.08847 | 平均 | 68.30819 |
| 标准误差 | 0.082655 | 标准误差 | 0.058672 |
| 中位数 | 68.2 | 中位数 | 68.5 |
| 众数 | 69.2 | 众数 | 68.5 |
| 标准差 | 2.517941 | 标准差 | 1.787333 |
| 方差 | 6.340029 | 方差 | 3.194561 |
| 峰度 | -0.33969 | 峰度 | 0.064434 |
| 偏度 | -0.08791 | 偏度 | -0.03515 |
| 区域 | 12 | 区域 | 9 |
| 最小值 | 61.7 | 最小值 | 64 |
| 最大值 | 73.7 | 最大值 | 73 |
| 求和 | 63186.1 | 求和 | 63390 |
| 观测数 | 928 | 观测数 | 928 |
| 最大(1) | 73.7 | 最大(1) | 73 |
| 最小(1) | 61.7 | 最小(1) | 64 |

## （三）案例分析


父代身高直方图
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
频率
64
66
68
70
72
分组


子代身高直方图
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
频率
64
66
68
70
72
分组

## （三）案例分析

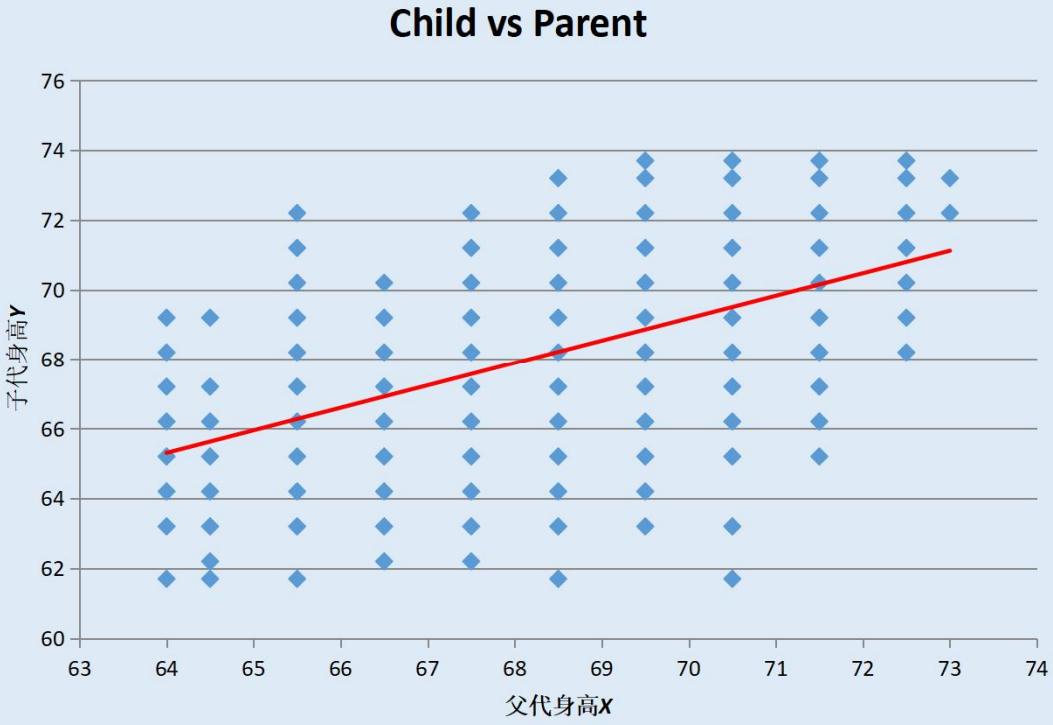
高尔顿对此研究后得出的解释是自然界有一种约束力，使人类身高在一定时期是相对稳定的。如果父 母身高高（或矮了），其子女比他们更高（矮），则人类身材将向高、矮两个极端分化。自然界不这样做，它让身高有一种回归到中心的作用。

例如，父母平均身高 72 英寸，这超过了平均值68英寸，表明这些父母属于高的一类，其儿子也倾向属于高的一类（其平均身高71 英寸大于子代69 英寸），但不像父母离子代那么远（71-69<72-68）。

反之，父母平均身高64 英寸，属于矮的一类，其儿子也倾向属于矮的一类（其平均67 英寸，小于子代的平均数69 英寸），但不像父母离中心那么远（69 -67< 68-64）。

## （三）案例分析

散点图




$y = 23.94 + 0.65x$

“回归”

# 二、豌豆父代和子代的关系

| | parent | child |
|---|---|---|
| 1 | 21 | 14.67 |
| 2 | 21 | 14.67 |
| 3 | 21 | 14.67 |
| 4 | 21 | 14.67 |
| 5 | 21 | 14.67 |
| 6 | 21 | 14.67 |
| 7 | 21 | 14.67 |
| 8 | 21 | 14.67 |
| 9 | 21 | 14.67 |
| 10 | 21 | 14.67 |
| 11 | 21 | 14.67 |
| . | . | |
| . | . | |
| . | . | |
| 699 | 15 | 19.77 |
| 700 | 15 | 19.77 |

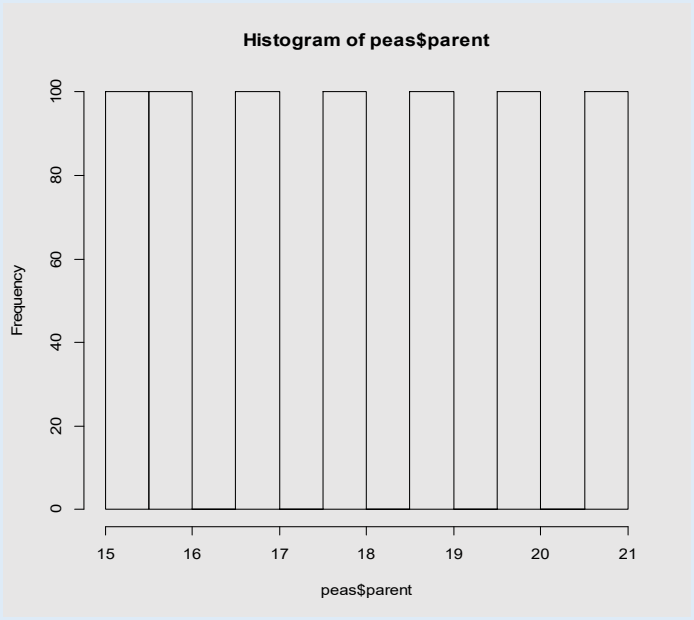
| vars | n | mean | sd | median | min | max | range | skew | kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| parent | 700 | 18.00 | 2.00 | 18.00 | 15.00 | 21.00 | 6.0 | 0.00 | -1.25 |
| child | 700 | 16.29 | 1.98 | 16.07 | 13.77 | 22.67 | 8.9 | 0.49 | -0.64 |

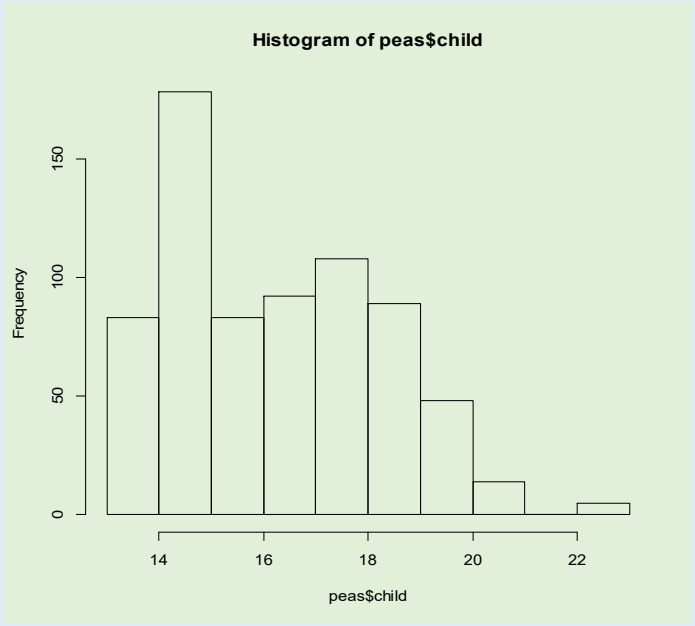
高尔顿豌豆数据(单位：0.01英寸)

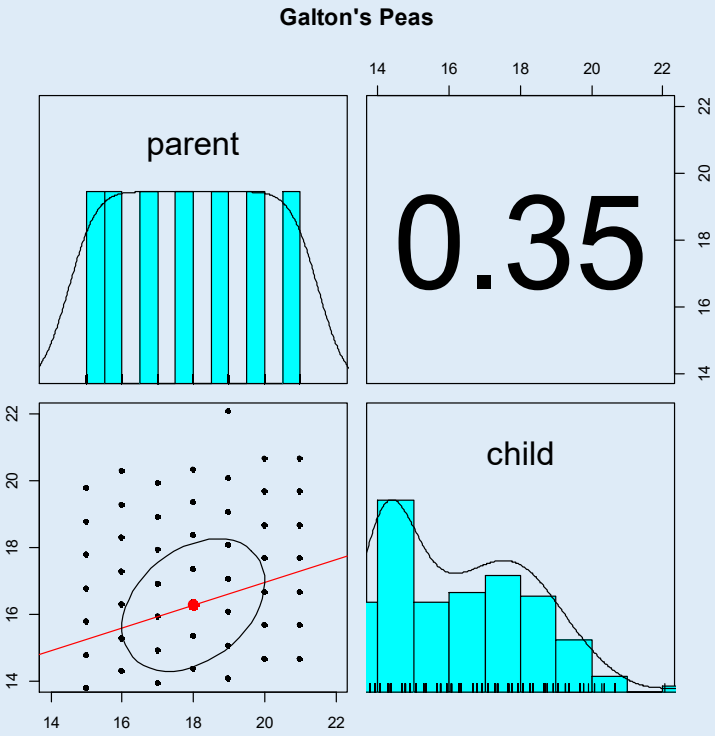
6

# 探索数据中的统计规律

## 二、豌豆父代和子代的关系



## 三、问题延伸

### 相关与回归

回归分析(regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

按照**涉及的变量的数量**，分为一元回归和多元回归分析；

按照**因变量的数量**，可分为简单回归分析和多重回归分析；

按照**自变量和因变量之间的关系类型**，可分为线性回归分析和非线性回归分析。

## 三、问题延伸

### 一元线性回归：

假设我们有一组人的身高和体重数据，可以通过简单线性回归分析，探索身高是否对体重有显著的线性影响。

建立线性模型: $Y = b_0 + b_1X + \varepsilon$，体重 = 截距 + 系数 × 身高 + 误差

回归系数：身高每增加一单位，体重平均增加的量。

## 三、问题延伸

**模型建立：** $Y = b_0 + b_1X + \varepsilon$

- 收集一组人的身高和体重数据，并确保数据的准确性。将数据整理成两列，一列为身高（自变量），另一列为体重（因变量）。
- 使用普通最小二乘法（OLS）估计模型中的$b_0$和 $b_1$：$\widehat{b_1} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$；$\widehat{b_0} = \overline{Y} - \widehat{b_1}\overline{X}$
- 将数据代入模型进行拟合。统计软件（如SPSS、R、Stata或Excel等）可以自动完成OLS回归，并输出系数（可用手动计算替代）
- 显著性检验：检验 $\widehat{b_1}$是否显著不为0，以确认身高对体重的影响

## 四、思考与练习

**老人的身心健康和其生活习惯有关吗？如何量化问题获取数据？如何建模分析？**

**根据2010年第六次全国人口普查详细汇总资料计算，我国人口平均预期寿命达到74.83岁[1]。据《2017年我国卫生健康事业发展统计公报》可知2017年我国居民人均预期寿命达76.7岁[2]**（2018年上海户籍人口人均期望寿命83.63岁，其中男性81.25岁，女性86.08岁[3]）

表2.4.1 平均预期寿命变化

单位:岁

| 年份 | 合计 | 男 | 女 | 男女之差 |
|---|---|---|---|---|
| 1981 | 67.77 | 66.28 | 69.27 | -2.99 |
| 1990 | 68.55 | 66.84 | 70.47 | -3.63 |
| 2000 | 71.40 | 69.63 | 73.33 | -3.70 |
| 2010 | **74.83** | 72.38 | 77.37 | -4.99 |

[1]来自国家统计局网页
http://www.stats.gov.cn/tjsj/tjgb/rkpcgb/qgrkpcgb/20120 9/t20120921_30330.html

[2]http://baijiahao.baidu.com/s?id=1603209126444 475473&wfr=spider&for=pc

[3]来自上海市人民政府网页
http://service.shanghai.gov.cn/SHVideo/ne wvideoshow.aspx?id=043B80172456C8A9

# THANK YOU!