

故事中的统计学

天津医科大学
生物医学工程与技术学院

2017-2018 学年下学期 (春)
公共选修课

第二章 精心挑选的平均数

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程与技术学院

2018 年 3 月



教学提纲

- 1 平均数
- 2 花言巧语的收入
- 3 其他案例分析

- 4 安全的旅行方式
- 5 知识拓展
- 6 图说天下
- 7 统计知识

1 平均数

2 花言巧语的收入

3 其他案例分析

4

安全的旅行方式

5

知识拓展

6

图说天下

7

统计知识



平均数

把大量的数据压缩成一个单独的数字，这种做法在实践过程中轻而易举就可以实现。可惜这对于许多事情来说太过分了，模糊了事实上存在的巨大差异，掩盖了一些事情的真相。

笑话

两个男人坐在一间酒馆里，其中一个人吃了一条牛腿，另一个人喝了两大桶啤酒。从统计学角度来看，每个人都喝了一桶啤酒，吃了半条牛腿，但实际的结果是，一个人吃得太多了，而另一个人却喝得烂醉。



平均数

把大量的数据压缩成一个单独的数字，这种做法在实践过程中轻而易举就可以实现。可惜这对于许多事情来说太过分了，模糊了事实上存在的巨大差异，掩盖了一些事情的真相。

笑话

两个男人坐在一间酒馆里，其中一个人吃了一条牛腿，另一个人喝了两大桶啤酒。从统计学角度来看，每个人都喝了一桶啤酒，吃了半条牛腿，但实际的结果是，一个人吃得太多了，而另一个人却喝得烂醉。



实例

如果在一个村庄上有 10 个农民，其中 1 个农民拥有 40 头牛，其他 9 个农民一头牛也没有。如果计算他们的平均数，则每个农民拥有 4 头牛。对于 9 个一无所有的农民来说，这种平均的结果只是一个苍白无力的安慰、水中捞月的梦想。

实例

英国的普利茅斯市和美国的明尼阿波利斯市拥有同样的年平均气温—— 13°C 。而真实的气候：

- 普利茅斯市：最冷的 2 月份最低气温也永远是 8°C ，最热的 7 月份最高气温也绝对不会超过 21°C 。
- 明尼阿波利斯市：在 1 月份平均气温通常都在 -15°C ，到了夏天平均气温在 30°C 以上，有时甚至会超过 40°C 。

平均数 | 简介

实例

如果在一个村庄上有 10 个农民，其中 1 个农民拥有 40 头牛，其他 9 个农民一头牛也没有。如果计算他们的平均数，则每个农民拥有 4 头牛。对于 9 个一无所有的农民来说，这种平均的结果只是一个苍白无力的安慰、水中捞月的梦想。

实例

英国的普利茅斯市和美国的明尼阿波利斯市拥有同样的年平均气温—— 13°C 。而真实的气候：

- 普利茅斯市：最冷的 2 月份最低气温也永远是 8°C ，最热的 7 月份最高气温也绝对不会超过 21°C 。
- 明尼阿波利斯市：在 1 月份平均气温通常都在 -15°C ，到了夏天平均气温在 30°C 以上，有时甚至会超过 40°C 。

实例

如果在一个村庄上有 10 个农民，其中 1 个农民拥有 40 头牛，其他 9 个农民一头牛也没有。如果计算他们的平均数，则每个农民拥有 4 头牛。对于 9 个一无所有的农民来说，这种平均的结果只是一个苍白无力的安慰、水中捞月的梦想。

实例

英国的普利茅斯市和美国的明尼阿波利斯市拥有同样的年平均气温—— 13°C 。而真实的气候：

- 普利茅斯市：最冷的 2 月份最低气温也永远是 8°C ，最热的 7 月份最高气温也绝对不会超过 21°C 。
- 明尼阿波利斯市：在 1 月份平均气温通常都在 -15°C ，到了夏天平均气温在 30°C 以上，有时甚至会超过 40°C 。

现象

当一个家伙希望用数据影响公众观点，或者向其他人推销广告版面，平均数便是一个经常被使用的伎俩，虽然偶尔是出于无心，但更多的时候是明知故犯。

原因

- “平均数”这个词具有宽泛的涵义，不同情境下使用不同的平均数。
- 某种条件下，各种类型平均数的数值十分接近，如果出于一般的目的，便没有必要区分它们。

结论

当你被告知某个数是平均数时，除非能说出它的具体种类——均值、中位数还是众数，否则你对它的具体涵义仍知之甚少。

现象

当一个家伙希望用数据影响公众观点，或者向其他人推销广告版面，平均数便是一个经常被使用的伎俩，虽然偶尔是出于无心，但更多的时候是明知故犯。

原因

- “平均数”这个词具有宽泛的涵义，不同情境下使用不同的平均数。
- 某种条件下，各种类型平均数的数值十分接近，如果出于一般的目的，便没有必要区分它们。

结论

当你被告知某个数是平均数时，除非能说出它的具体种类——均值、中位数还是众数，否则你对它的具体涵义仍知之甚少。

现象

当一个家伙希望用数据影响公众观点，或者向其他人推销广告版面，平均数便是一个经常被使用的伎俩，虽然偶尔是出于无心，但更多的时候是明知故犯。

原因

- “平均数”这个词具有宽泛的涵义，不同情境下使用不同的平均数。
- 某种条件下，各种类型平均数的数值十分接近，如果出于一般的目的，便没有必要区分它们。

结论

当你被告知某个数是平均数时，除非能说出它的具体种类——均值、中位数还是众数，否则你对它的具体涵义仍知之甚少。

Mean • Median • Mode

mean

The mean is the average of a set of numbers.

To find the mean:

- Add together all of the numbers in the set.
- Divide the total by how many numbers were added.

$$1 + 3 + 6 + 9 + 11 + 12 + 14 = 56 \rightarrow 56 \div 7 = 8$$

The mean is 8.

median

The median is the middle number in a sequence.

To find the median:

- Organize the set of numbers from smallest to largest.
- Locate the middle number.

2, 3, 5, 6, 7, 8, 10, 13, 14

The median is 7.

mode

The mode is the number that occurs most frequently.

To find the mode:

- Organize the list of numbers from smallest to largest.
- Locate the number that appears in the list most often.

1, 2, 2, 3, 4, 4, 5, 7, 7, 8

The mode is 4.

Math Matters!

If there are two values located in the middle, find the mean of those numbers to solve for the median. (Example: The median for 2, 3, 4, 6, 11, 15 is 5.)

It is possible to have no mode when there are no repeated numbers or when there is more than one mode within a set of numbers.



平均数 | 种类

MEAN

Commonly used in sport to find out a score in sports like Football, Basketball and Cricket

Is also known as the "average"

1. Add up all the values to get the total
2. Then divide the total by the number of values you added together

$$\begin{array}{r} 3 + 4 + 8 + 7 + 5 + 3 = 30 \\ \hline 30 \div 6 = 5 \end{array}$$

The average for these values is 5



MEDIAN

Used when comparing house prices.

The "middle" number in a set of values

1. First put all the values in order
2. Find the middle number in the set of data
3. If there are two values in the middle, find the mean of these two.

$$1, 2, 4, \star 5 \underline{6}, 8, 9$$



The median is 5.

Mode

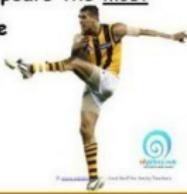
Eg. What is the mode of goals kicked by a footballer after each round?

The number which occurs the most

1. Count how many of each value appears
2. The mode is the value which appears the most
3. There can be more than 1 mode

$$1, 2, 2, 5, 6, 6, 9$$

2 and 6 are the mode for these values.



range

Measures difference between all the values.
Used in weather.

The range is the difference between the highest and lowest value

1. Find the highest and lowest values
2. Subtract the lowest value from the highest

$$1, 2, 2, 5, 6, 6, 9$$

$$9 - 1 = 8 \quad \text{The range is 8}$$



Mean, Median, Mode, and Range

First, arrange the numbers in order by size.

Example: 3, 5, 5, 6, 8, 10, 12

Mean

the average
of the numbers

1. Add the numbers together.
2. Divide by how many numbers were added.

$$3+5+5+6+8+10+12=49$$

$$49 \div 7 = 7$$

The mean is 7.

Median

the middle
number of
a sequence

The median is the middle number when numbers are arranged in order by size.

For an even number of numbers, the median is the average of the two numbers in the middle.

The middle number is 6.

The median is 6.

Mode

the number
that occurs
most often

Find the number(s) that occurs most often in the sequence (there may be more than one).

There are two 5s and one of each of the other numbers.

The mode is 5.

Range

the difference
between the
lowest and
highest values

Subtract the smallest number from the largest number.

$$12 - 3 = 9$$

The range is 9.

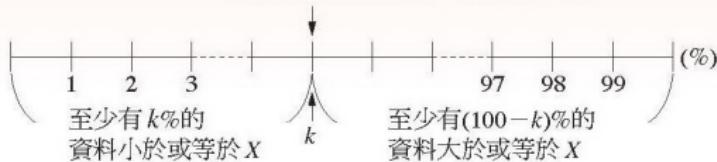


平均數 | 百分位數

● 百分位數(P_k)

將蒐集來的資料由小到大排序，若至少有 $k\%$ 的資料小於或等於 X ，且至少有 $(100-k)\%$ 的資料大於或等於 X ，則 X 稱為這一群資料的第 k 百分位數，其中 k 是正整數，且 $1 \leq k \leq 99$ 。

第 k 百分位數是 X ，記作 $P_k = X$



平均数 | 百分位数

● 百分位數(P_k)

將蒐集來的資料由小到大排序，若至少有 $k\%$ 的資料小於或等於 X ，且至少有 $(100-k)\%$ 的資料大於或等於 X ，則 X 稱為這一群資料的第 k 百分位數，其中 k 是正整數，且 $1 \leq k \leq 99$ 。

第 k 百分位數是 X ，記作 $P_k = X$



Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

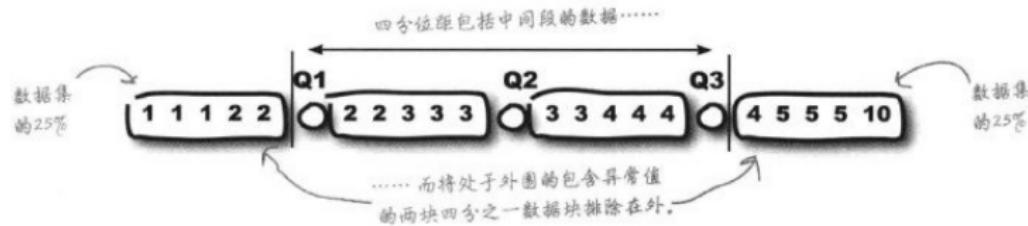


That means you are at the **80th percentile**.

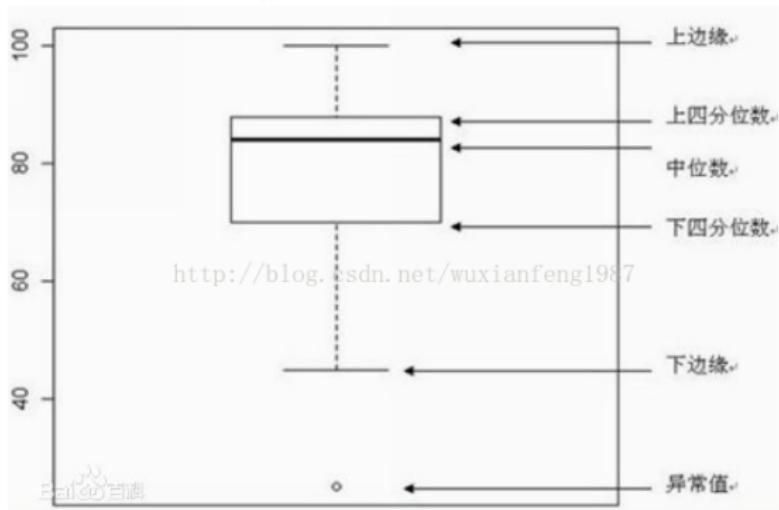
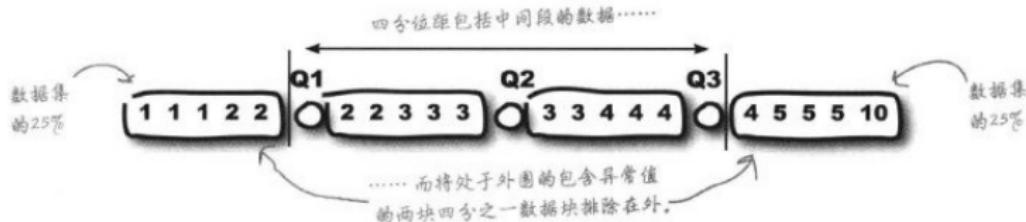
If your height is 1.85m then "1.85m" is the 80th percentile height in that group.



平均数 | 四分位数



平均数 | 四分位数



- 1 平均数
- 2 花言巧语的收入
- 3 其他案例分析

- 4 安全的旅行方式
- 5 知识拓展
- 6 图说天下
- 7 统计知识



案例 | 小区居民年收入

谁在说谎？

房产中介 小区居民平均年收入 10000 英镑

小区居民（纳税人） 小区居民平均年收入只有 2000 英镑

吃瓜群众 小区居民平均年收入 3000 英镑

- 三个数字都是基于相同的数据（相同的居民、相同的收入），都是正规的平均数，计算方法也完全正确
- 三次分别使用了不同的平均数：均值（算术平均数）、众数、中位数
- 关于收入的数据，不合适的“平均数”实际上是毫无意义的



案例 | 小区居民年收入

谁在说谎？

房产中介 小区居民平均年收入 10000 英镑

小区居民（纳税人） 小区居民平均年收入只有 2000 英镑

吃瓜群众 小区居民平均年收入 3000 英镑

统计在撒谎！

- 三个数字都是基于相同的数据（相同的居民、相同的收入），都是正规的平均数，计算方法也完全正确
- 三次分别使用了不同的平均数：均值（算术平均数）、众数、中位数
- 关于收入的数据，不合适的“平均数”实际上是毫无意义的



案例 | 小区居民年收入

谁在说谎？

房产中介 小区居民平均年收入 10000 英镑

小区居民（纳税人） 小区居民平均年收入只有 2000 英镑

吃瓜群众 小区居民平均年收入 3000 英镑

统计在撒谎！

- 三个数字都是基于相同的数据（相同的居民、相同的收入），都是正规的平均数，计算方法也完全正确
- 三次分别使用了不同的平均数：均值（算术平均数）、众数、中位数
- 关于收入的数据，不合适的“平均数”实际上是毫无意义的



案例 | 小区居民年收入

谁在说谎？

房产中介 小区居民平均年收入 10000 英镑

小区居民（纳税人） 小区居民平均年收入只有 2000 英镑

吃瓜群众 小区居民平均年收入 3000 英镑

统计在撒谎！

- 三个数字都是基于相同的数据（相同的居民、相同的收入），都是正规的平均数，计算方法也完全正确
- 三次分别使用了不同的平均数：均值（算术平均数）、众数、中位数
- 关于收入的数据，不合适的“平均数”实际上是毫无意义的



案例 | 小区居民年收入

谁在说谎？

房产中介 小区居民平均年收入 10000 英镑

小区居民（纳税人） 小区居民平均年收入只有 2000 英镑

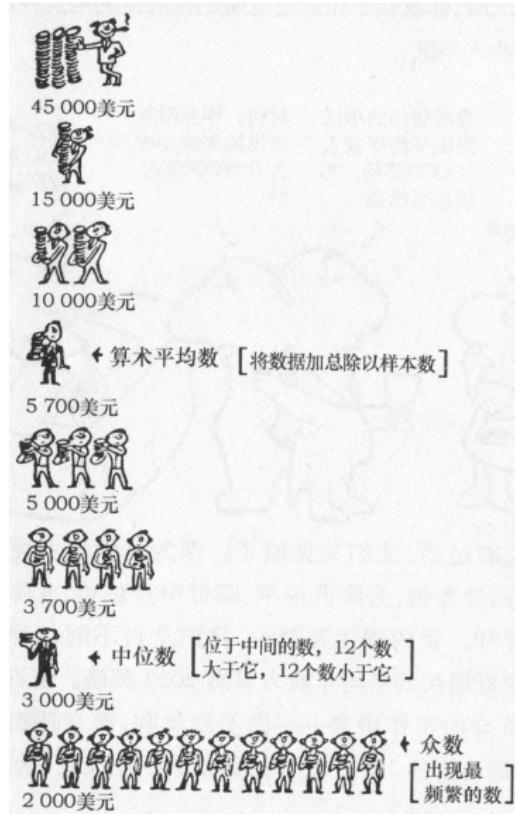
吃瓜群众 小区居民平均年收入 3000 英镑

统计在撒谎！

- 三个数字都是基于相同的数据（相同的居民、相同的收入），都是正规的平均数，计算方法也完全正确
- 三次分别使用了不同的平均数：均值（算术平均数）、众数、中位数
- 关于收入的数据，不合适的“平均数”实际上是毫无意义的



案例 | 小区居民年收入



正态分布

在处理诸如人类特征的数据时，各种平均数的数值十分接近。这些数据具有我们常说的正态分布的形态特点，在用曲线绘制正态分布时，将看到一条钟形的曲线，均值、中位数和众数都落在相同的点上。

非正态分布

收入的分布不再像钟形一样对称，而是有偏的，它的形状类似于孩子玩的滑梯，梯子一侧是陡斜地升到顶部，而滑道一侧则缓慢向下倾斜。此时，均值与中位数、众数相差甚远。



案例 | 小区居民年收入 | 总结

正态分布

在处理诸如人类特征的数据时，各种平均数的数值十分接近。这些数据具有我们常说的正态分布的形态特点，在用曲线绘制正态分布时，将看到一条钟形的曲线，均值、中位数和众数都落在相同的点上。

非正态分布

收入的分布不再像钟形一样对称，而是有偏的，它的形状类似于孩子玩的滑梯，梯子一侧是陡斜地升到顶部，而滑道一侧则缓慢向下倾斜。此时，均值与中位数、众数相差甚远。

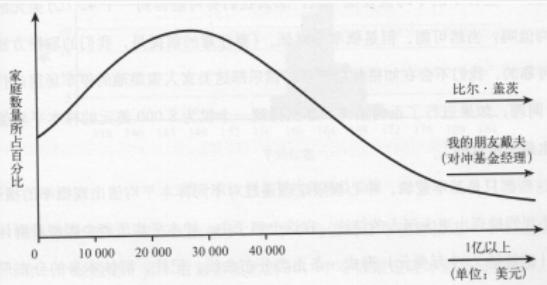
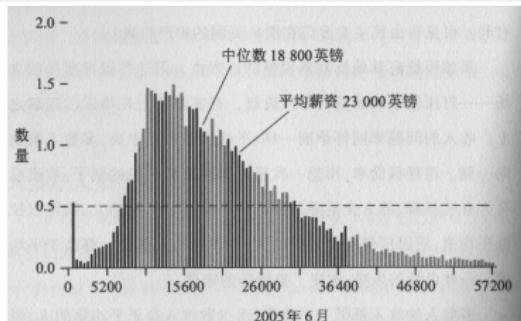
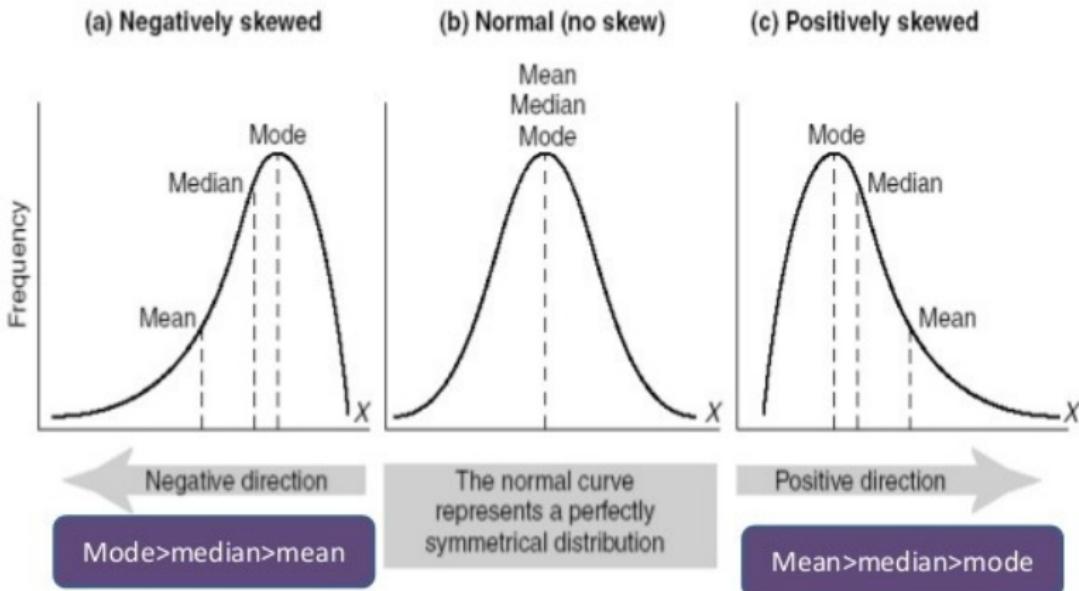


图 9-1 美国家庭年收入分布

Position of mean median mode



公司员工的平均收入

当你听到公司执行总裁或企业所有者宣称，在他的企业中员工的平均收入是多少时，你应该好好思考一下其中的原因。

- 如果这个数是中位数，你可以获得一些显而易见的信息：一半员工赚得比它多，一半比它少。
- 如果是均值（当没有确切指出它的种类时，多半是均值），它仅仅是所有者的高收入与全体工人低水平收入的平均数，根本没有什么意义。“平均年收入为 3800 英镑”既隐瞒了 1400 英镑的低收入，又隐瞒了所有者以巨额薪金形式抽取的高额利润。



案例 | 公司员工收入



虹宇公司员工的月薪如下：

经

应聘者

阿冲

员工	经理	副经理	职员 A	职员 B	职员 C	职员 D	职员 E	职员 F	职员 G
月薪(元)	6000	4000	1700	1300	1200	1100	1100	1100	500

1. 经理说平均工资有2000元是否欺骗了阿冲？
2. 平均工资2000元能否客观地反映公司员工的平均收入？
3. 若不能，你认为用哪个数据表示该公司员工收入的“平均水平”更合适？



背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

解决方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 $(99000/90 = 1100)$

所有者的平均工资及利润 12500 英镑 $((5500 \times 3) + 21000)/3 = 12500$

谭武的思考

这看上去太不公平了，自己都看不过去！

背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

初级方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

高级的思考

这看上去太不公平了，自己都看不过去！

背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

初级方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

谭武的思考

这看上去太不公平了，自己都看不过去！

背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

初级方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

谭武的思考

这看上去太不公平了，自己都看不过去！

背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

初级方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

谭武的思考

这看上去太不公平了，自己都看不过去！

背景介绍

谭武是某个小型制造企业的 3 个合伙人之一。到了年底，谭武给企业的 90 个职工共发了 99000 英镑；谭武和其他合伙人每人各获得 5500 英镑的工资；最后还余下 21000 英镑，作为利润可供 3 个合伙人平分。谭武将如何说明这种情况呢？

初级方案

为了便于理解，谭武打算采用平均数的形式：

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

谭武的思考

这看上去太不公平了，自己都看不过去！

换一种形式

- ① 从 21000 英镑的利润中拿出 15000 英镑以奖金的形式平分给 3 位合伙人
- ② 将包括了所有者和职工的工资进行平均

改进方案

所有人员的平均工资或薪金 1403 英镑

$$((99000 + 5500 \times 3 + 15000) / 93 = 1403.226)$$

所有者平均利润 2000 英镑 $((21000 - 15000) / 3 = 2000)$

谭武的思考

现在看上去好多了，足以作为公布的内容张贴在公告栏中，或者作为与职工谈判的依据。

换一种形式

- ① 从 21000 英镑的利润中拿出 15000 英镑以奖金的形式平分给 3 位合伙人
- ② 将包括了所有者和职工的工资进行平均

改进方案

所有人员的平均工资或薪金 1403 英镑

$$((99000 + 5500 \times 3 + 15000) / 93 = 1403.226)$$

所有者平均利润 2000 英镑 $((21000 - 15000) / 3 = 2000)$

谭武的思考

现在看上去好多了，足以作为公布的内容张贴在公告栏中，或者作为与职工谈判的依据。

换一种形式

- ① 从 21000 英镑的利润中拿出 15000 英镑以奖金的形式平分给 3 位合伙人
- ② 将包括了所有者和职工的工资进行平均

改进方案

所有人员的平均工资或薪金 1403 英镑

$$((99000 + 5500 \times 3 + 15000) / 93 = 1403.226)$$

所有者平均利润 2000 英镑 $((21000 - 15000) / 3 = 2000)$

谭武的思考

现在看上去好多了，足以作为公布的内容张贴在公告栏中，或者作为与职工谈判的依据。

案例 | 公司员工收入 | 公司的声明

总结

- 本质不变：总的钱数没变，每个职工拿到的钱数没变，每个合伙人拿到的钱数没变！
- 表象反转：最终的结论/声明/公告竟有天壤之别！

初级方案

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

改进方案

所有人员的平均工资或薪金 1403 英镑

$((99000 + 5500 \times 3 + 15000)/93 = 1403.226)$

所有者平均利润 2000 英镑 ($(21000 - 15000)/3 = 2000$)

案例 | 公司员工收入 | 公司的声明

总结

- 本质不变：总的钱数没变，每个职工拿到的钱数没变，每个合伙人拿到的钱数没变！
- 表象反转：最终的结论/声明/公告竟有天壤之别！

初级方案

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

改进方案

所有人员的平均工资或薪金 1403 英镑

$((99000 + 5500 \times 3 + 15000)/93 = 1403.226)$

所有者平均利润 2000 英镑 ($(21000 - 15000)/3 = 2000$)

总结

- 本质不变：总的钱数没变，每个职工拿到的钱数没变，每个合伙人拿到的钱数没变！
- 表象反转：最终的结论/声明/公告竟有天壤之别！

初级方案

职工的平均工资 1100 英镑 ($99000/90 = 1100$)

所有者的平均工资及利润 12500 英镑 ($((5500 \times 3) + 21000)/3 = 12500$)

改进方案

所有人员的平均工资或薪金 1403 英镑

$((99000 + 5500 \times 3 + 15000)/93 = 1403.226)$

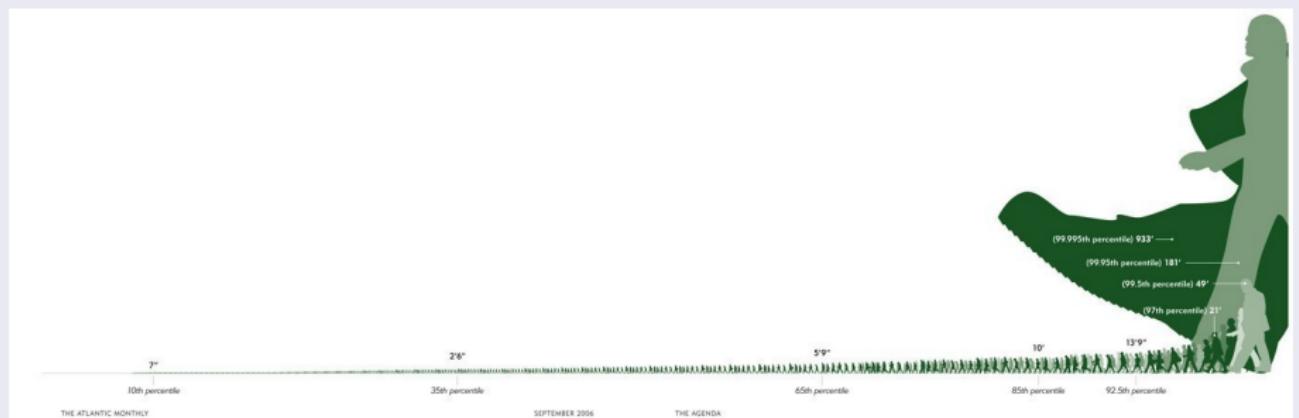
所有者平均利润 2000 英镑 ($(21000 - 15000)/3 = 2000$)

收入

- 收入的问题如同怀孕期一样：平均值不在正中央，多数人都偏向一侧。
- 多数人的收入都低于平均值，而少数收入高于平均值的人，将平均值远远拉离中间数。
- 因为平均收入被拉高了，以致多数人的收入都低于平均值。
- 领平均薪资的人已经算是相当富有了，因为平均值已经被少数高收入者远远拉离中间数。
- 从经济的角度而言，很多有关“英国中产阶级”的政治媒体言论根本不晓得“中产阶级”的定义是什么。



有钱的巨人



一个百万富翁对于平均值的影响，远大于成千上万个贫民，而亿万富翁的影响力则又大了 100 倍。他们的影响力，大到让全世界 80% 人口的财富低于平均值。

总结

当你看到某个平均收入时，首先问问：是什么的平均？包括了哪些人？

包括了哪些人？

美国钢铁公司曾经指出：10 年间，该公司职工的平均周收入攀升了 107%！然并卵……

- 早期的数据包括了兼职员工
- 如果你某年只工作了半年，而第二年全年都在工作，你的收入毫无疑问会翻番，但这并不意味着工资率发生了变动。



总结

当你看到某个平均收入时，首先问问：是什么的平均？包括了哪些人？

包括了哪些人？

美国钢铁公司曾经指出：10 年间，该公司职工的平均周收入攀升了 107%！然并卵……

- 早期的数据包括了兼职员工
- 如果你某年只工作了半年，而第二年全年都在工作，你的收入毫无疑问会翻番，但这并不意味着工资率发生了变动。



总结

当你看到某个平均收入时，首先问问：是什么的平均？包括了哪些人？

包括了哪些人？

美国钢铁公司曾经指出：10 年间，该公司职工的平均周收入攀升了 107%！然并卵……

- 早期的数据包括了兼职员工
- 如果你某年只工作了半年，而第二年全年都在工作，你的收入毫无疑问会翻番，但这并不意味着工资率发生了变动。



报道

你也许曾在报纸上看到过，某年美国的家庭平均收入是 6940 美元。别太在意这个数字，除非知道：

- 这个数字包括了哪些家庭？
- 使用了哪种平均数？
- 谁说的？他是如何获得该信息的？这个数的准确性如何？

普查局

- “家庭”是指两个或更多具有亲属关系的人住在一起所形成的“家庭”。
- 这是个中位数。
- 数据建立在抽样基础上，该调查以 19/20 的概率保证真实的数值会落在估计值加减 71 美元的范围之内。

报道

你也许曾在报纸上看到过，某年美国的家庭平均收入是 6940 美元。别太在意这个数字，除非知道：

- 这个数字包括了哪些家庭？
- 使用了哪种平均数？
- 谁说的？他是如何获得该信息的？这个数的准确性如何？

普查局

- “家庭”是指两个或更多具有亲属关系的人住在一起所形成的“家庭”。
- 这是个中位数。
- 数据建立在抽样基础上，该调查以 $19/20$ 的概率保证真实的数值会落在估计值加减 71 美元的范围之内。

问题

美国中产阶级的经济健康状况如何？我们称之为“中产阶级”的人到底是更富了、更穷了，还是在原地踏步？

答案

一个合理的答案——肯定不会有“正确”的答案——就是，计算一代美国人（大约为 30 年）的人均收入，观察其变化趋势。

存在的问题

- 没有考虑通货膨胀因素
- 我们需要知道的是普通美国人的收入，而不是泛泛的人均收入，这两者有本质上的区别

案例 | 中产阶级

问题

美国中产阶级的经济健康状况如何？我们称之为“中产阶级”的人到底是更富了、更穷了，还是在原地踏步？

答案

一个合理的答案——肯定不会有“正确”的答案——就是，计算一代美国人（大约为 30 年）的人均收入，观察其变化趋势。

存在的问题

- 没有考虑通货膨胀因素
- 我们需要知道的是普通美国人的收入，而不是泛泛的人均收入，这两者有本质上的区别

案例 | 中产阶级

问题

美国中产阶级的经济健康状况如何？我们称之为“中产阶级”的人到底是更富了、更穷了，还是在原地踏步？

答案

一个合理的答案——肯定不会有“正确”的答案——就是，计算一代美国人（大约为 30 年）的人均收入，观察其变化趋势。

存在的问题

- 没有考虑通货膨胀因素
- 我们需要知道的是普通美国人的收入，而不是泛泛的人均收入，这两者有本质上的区别

案例 | 中产阶级 | 通货膨胀

排名	电影片名	电影原名	发行商	全球票房	发行年
1	阿凡达	Avatar	二十世纪福斯	\$2,787,965,087	2009
2	泰坦尼克号	Titanic	派拉蒙影业 / 二十世纪福斯	\$2,187,463,944	1997
3	星球大战：原力觉醒	Star Wars: The Force Awakens	华特迪士尼影业	\$2,068,223,624	2015
4	侏罗纪世界	Jurassic World	环球影业	\$1,671,713,208	2015
5	复仇者联盟	Marvel's The Avengers	华特迪士尼影业	\$1,518,812,988	2012
6	速度与激情7	Fast And Furious 7	环球影业	\$1,516,045,911	2015
7	复仇者联盟2：奥创纪元	Avengers: Age of Ultron	华特迪士尼影业	\$1,405,403,694	2015
8	哈利·波特与死亡圣器（下）	Harry Potter and the Deathly Hallows Part 2	华纳兄弟	\$1,341,511,219	2011
9	星球大战：最后的绝地武士	Star Wars: The Last Jedi	华特迪士尼影业	\$1,311,425,821	2017
10	冰雪奇缘	Frozen	华特迪士尼影业	\$1,276,480,335	2013



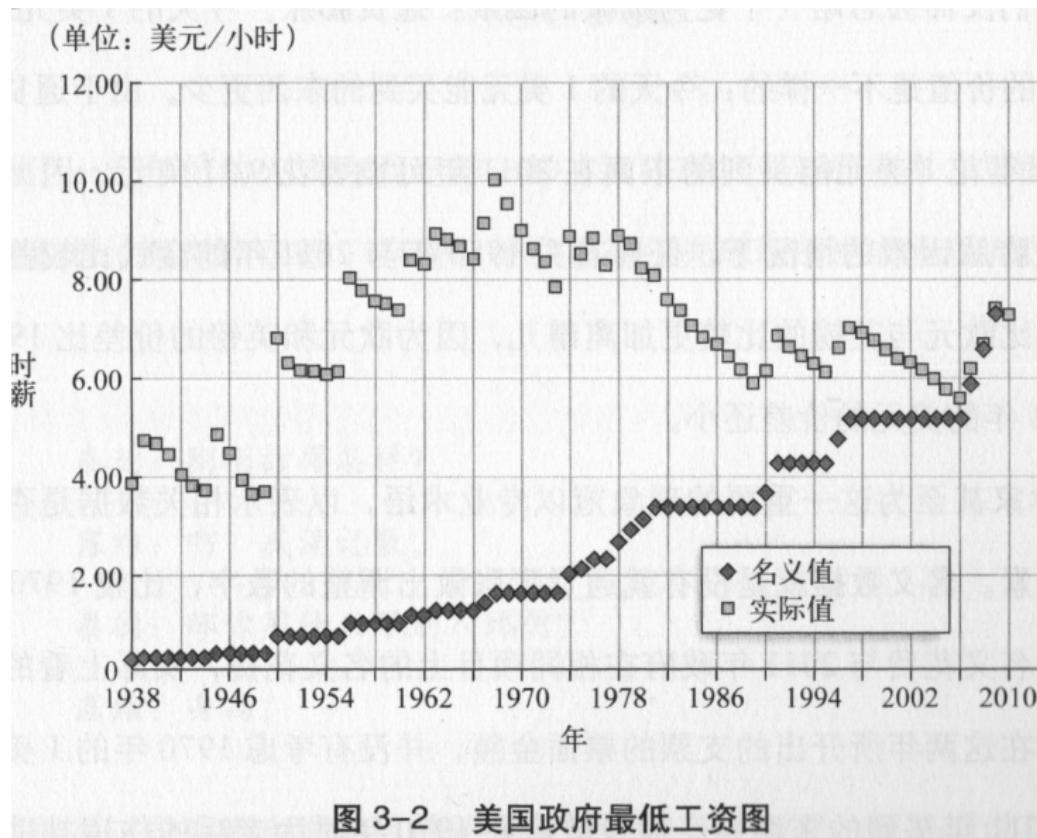
案例 | 中产阶级 | 通货膨胀

排名	电影片名	电影原名	发行商	全球票房	发行年
1	阿凡达	Avatar	二十世纪福斯	\$2,787,965,087	2009
2	泰坦尼克号	Titanic	派拉蒙影业 / 二十世纪福斯	\$2,187,463,944	1997
3	星球大战：原力觉醒	Star Wars: The Force Awakens	华特迪士尼影业	\$2,068,223,624	2015
4	侏罗纪世界	Jurassic World	环球影业	\$1,671,713,208	2015
5	复仇者联盟	Marvel's The Avengers	华特迪士尼影业	\$1,518,812,988	2012
6	速度与激情7	Fast And Furious 7	环球影业	\$1,516,045,911	2015
7	复仇者联盟2：奥创纪元	Avengers: Age of Ultron	华特迪士尼影业	\$1,405,403,694	2015
8	哈利·波特与死亡圣器（下）	Harry Potter and the Deathly Hallows Part 2	华纳兄弟	\$1,341,511,219	2011
9	星球大战：最后的绝地武士	Star Wars: The Last Jedi	华特迪士尼影业	\$1,311,425,821	2017
10	冰雪奇缘	Frozen	华特迪士尼影业	\$1,276,480,335	2013

排名	电影片名	电影原名	发行年	观影人次估计 ^[1]	经通胀计算后之票房收入
1	乱世佳人	Gone With The Wind	1939	202,044,600	\$1,687,072,600
2	星球大战四部曲：曙光乍现	Star Wars Episode IV: A New Hope	1977	178,119,600	\$1,487,298,600
3	音乐之声	The Sound of Music	1965	142,415,400	\$1,189,168,400
4	E.T.外星人	E.T. the Extra-Terrestrial	1982	141,854,300	\$1,184,483,700
5	泰坦尼克号	Titanic	1997	135,654,500	\$1,131,211,800
6	十诫	The Ten Commandments	1956	131,000,000	\$1,093,850,000
7	大白鲨	Jaws	1975	128,078,800	\$1,069,458,100
8	日瓦戈医生	Doctor Zhivago	1965	124,135,500	\$1,036,531,100
9	驱魔人	The Exorcist	1973	110,599,200	\$923,503,000
10	白雪公主	Snow White and the Seven Dwarfs	1937	109,000,000	\$910,150,000



案例 | 中产阶级 | 通货膨胀



专家的答案

要评价美国“中产阶级”的经济状况，我们需要了解（通货膨胀调整后的）工资中位数在过去几十年中的变化，还建议留意一下处于第 25 百分位数和第 75 百分位数人群的工资变化，因为这两拨人通常被认为是中产阶级中的高收入和低收入人群。

注意

在评价经济状况的过程中，不能将收入和工资等同起来。这两者是不同的，工资是我们付出的固定份额的劳动所得，如时薪或周薪；收入是全部所得的总和，来源有多种。相比于收入来说，工资是评价美国人劳动收益的一个更加直观的指标，工资越高，工人们每工作 1 小时能领到的钱也就越多。



专家的答案

要评价美国“中产阶级”的经济状况，我们需要了解（通货膨胀调整后的）工资中位数在过去几十年中的变化，还建议留意一下处于第 25 百分位数和第 75 百分位数人群的工资变化，因为这两拨人通常被认为是中产阶级中的高收入和低收入人群。

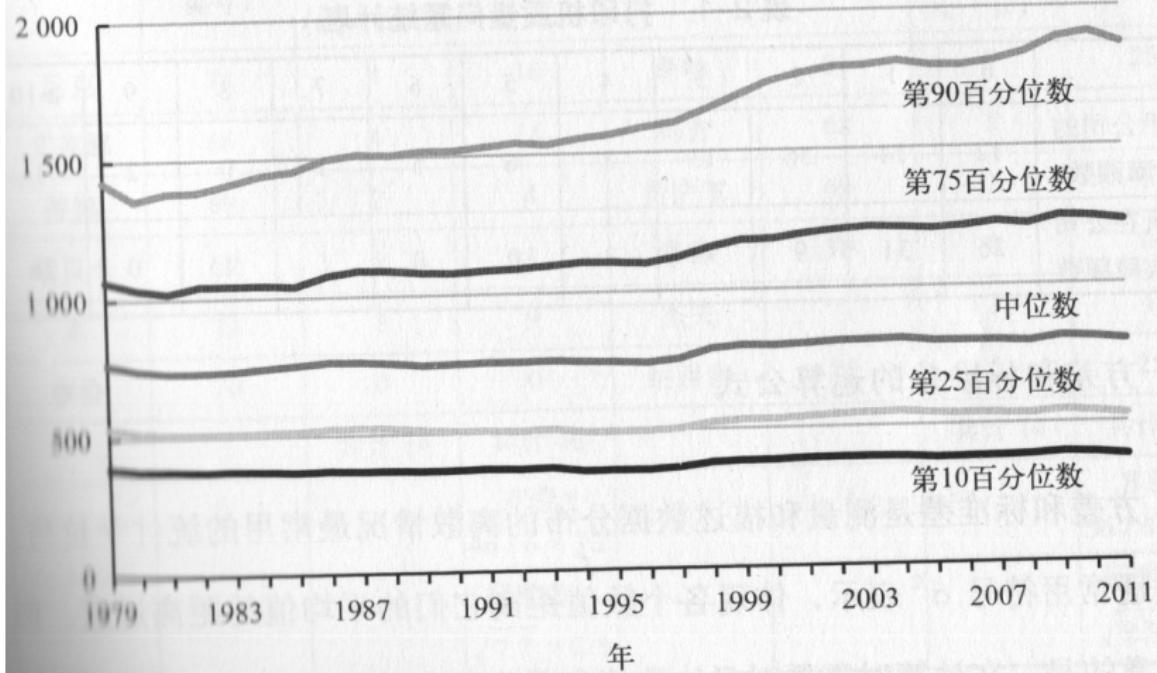
注意

在评价经济状况的过程中，不能将收入和工资等同起来。这两者是不同的，工资是我们付出的固定份额的劳动所得，如时薪或周薪；收入是全部所得的总和，来源有多种。相比于收入来说，工资是评价美国人劳动收益的一个更加直观的指标，工资越高，工人们每工作 1 小时能领到的钱也就越多。



案例 | 中产阶级

(单位: 美元)



减税政策

美国前总统小布什政府的说法，减税政策将惠及绝大多数的美国家庭。在这项政策推行后，将会有 9200 万美国人享受减税待遇，人均减税额超过 1000 美元（具体数字应该是 1083 美元）。这个关于减税政策的概括准确吗？

《纽约时报》评价说：“数据本身并没有撒谎，只不过有些数据没有发出声音罢了。”是不是会有 9200 万美国人将享受减税待遇？答案是肯定的。那么，这些人中的大部分人都可以少缴纳约 1000 美元的税款吗？不是的。因为减税额的中位数还不足 100 美元。只有数量相对少的巨富们才有资格享受大额减税，而正是这些人拉高了平均值，让人均减税额看起来比绝大多数美国人真正享受到的要高。

减税政策

美国前总统小布什政府的说法，减税政策将惠及绝大多数的美国家庭。在这项政策推行后，将会有 9200 万美国人享受减税待遇，人均减税额超过 1000 美元（具体数字应该是 1083 美元）。这个关于减税政策的概括准确吗？

解析

《纽约时报》评价说：“数据本身并没有撒谎，只不过有些数据没有发出声音罢了。”是不是会有 9200 万美国人将享受减税待遇？答案是肯定的。那么，这些人中的大部分人都可以少缴纳约 1000 美元的税款吗？不是的。因为减税额的中位数还不足 100 美元。只有数量相对少的巨富们才有资格享受大额减税，而正是这些人拉高了平均值，让人均减税额看起来比绝大多数美国人真正享受到的要高。



减税政策

美国前总统小布什政府的说法，减税政策将惠及绝大多数的美国家庭。在这项政策推行后，将会有 9200 万美国人享受减税待遇，人均减税额超过 1000 美元（具体数字应该是 1083 美元）。这个关于减税政策的概括准确吗？

解析

《纽约时报》评价说：“数据本身并没有撒谎，只不过有些数据没有发出声音罢了。”是不是会有 9200 万美国人将享受减税待遇？答案是肯定的。那么，这些人中的大部分人都可以少缴纳约 1000 美元的税款吗？不是的。因为减税额的中位数还不足 100 美元。只有数量相对少的巨富们才有资格享受大额减税，而正是这些人拉高了平均值，让人均减税额看起来比绝大多数美国人真正享受到的要高。

减税政策

美国前总统小布什政府的说法，减税政策将惠及绝大多数的美国家庭。在这项政策推行后，将会有 9200 万美国人享受减税待遇，人均减税额超过 1000 美元（具体数字应该是 1083 美元）。这个关于减税政策的概括准确吗？

解析

《纽约时报》评价说：“数据本身并没有撒谎，只不过有些数据没有发出声音罢了。”是不是会有 9200 万美国人将享受减税待遇？答案是肯定的。那么，这些人中的大部分人都可以少缴纳约 1000 美元的税款吗？不是的。因为减税额的中位数还不足 100 美元。只有数量相对少的巨富们才有资格享受大额减税，而正是这些人拉高了平均值，让人均减税额看起来比绝大多数美国人真正享受到的要高。

减税政策

美国前总统小布什政府的说法，减税政策将惠及绝大多数的美国家庭。在这项政策推行后，将会有 9200 万美国人享受减税待遇，人均减税额超过 1000 美元（具体数字应该是 1083 美元）。这个关于减税政策的概括准确吗？

解析

《纽约时报》评价说：“数据本身并没有撒谎，只不过有些数据没有发出声音罢了。”是不是会有 9200 万美国人将享受减税待遇？答案是肯定的。那么，这些人中的大部分人都可以少缴纳约 1000 美元的税款吗？不是的。因为减税额的中位数还不足 100 美元。只有数量相对少的巨富们才有资格享受大额减税，而正是这些人拉高了平均值，让人均减税额看起来比绝大多数美国人真正享受到的要高。

1 平均数

2 花言巧语的收入

3 其他案例分析

4

安全的旅行方式

5

知识拓展

6

图说天下

7

统计知识



案例 | 跳绳成绩

例 3 下面是四年级一班9个男生1分钟跳绳成绩的记录单。

编 号	1	2	3	4	5	6	7	8	9
成 绩 / 下	102	170	96	90	97	106	110	182	100

平均数是117下

7号男生跳了110下，他的成绩处在
这组同学中的什么位置？



为什么跳的比平均数少，成绩还是第三名？



182 170 110 106 102 100 97 96 90

—

正中间的一个数是102，102是这组数据的中位数



同中位数比，7号男生的成绩怎么样？



案例 | 小新的成绩



妈妈，我这次数学考试考了78分，而全班平均分只有77分，每次都最后几个，这次我在中上水平，进步很大啊！



你真棒！成绩
单给妈妈看一下
好吗？
www.liubb.net

六（1）班第三单元数学成绩单

姓名：小新 成绩：**78分** 平均分：**77分**



分数	100	91	90	86	78	15	6
人数	1	15	4	13	1	4	2

我很生气！
碰碰.....

问：小新撒谎了吗？



为什么妈妈
那么生气呢？
我没有撒谎
啊！

聪明的你能帮小新解开疑惑
吗？小新哪句话讲的不合实
际？



案例 | 2017 年金球奖

金球奖候选人具体得分：C 罗分数接近梅西和内马尔总和
C 罗获得了 2017 年金球奖，B/R 透露了候选人具体的得分，最终 C 罗的得分接近于梅西和内马尔的总和。

规则

共有 173 名记者参与投票，每张选票填写 5 个名字，其中该选票中的第一名得 6 分，第二名得 4 分，第三名得 3 分，第四名 2 分，第五名 1 分。

案例 | 2017 年金球奖

金球奖候选人具体得分：C 罗分数接近梅西和内马尔总和

C 罗获得了 2017 年金球奖，B/R 透露了候选人具体的得分，最终 C 罗的得分接近于梅西和内马尔的总和。



规则

共有 173 名记者参与投票，每张选票填写 5 个名字，其中该选票中的第一名得 6 分，第二名得 4 分，第三名得 3 分，第四名 2 分，第五名 1 分。

案例 | 2017 年金球奖

金球奖候选人具体得分：C 罗分数接近梅西和内马尔总和

C 罗获得了 2017 年金球奖，B/R 透露了候选人具体的得分，最终 C 罗的得分接近于梅西和内马尔的总和。



规则

共有 173 名记者参与投票，每张选票填写 5 个名字，其中该选票中的第一名得 6 分，第二名得 4 分，第三名得 3 分，第四名 2 分，第五名 1 分。

其他

- 我们每个人的脚，几乎都比平均脚数多。
-



案例 | 统计 vs. 战术 vs. 谎言



我们是以身高的算术平均数计数来吓唬对手，还是采取中位数计数来哄骗对手。



案例 | 《时代》杂志的订户

“编者的话”

旧订户 年龄的中位数是 41 岁，家庭平均年收入为 9535 美元。

新订户 年龄的中位数是 34 岁，家庭平均年收入为 7270 美元。

疑问

为什么两次谈到年龄时都指出采用了中位数，而关于收入却不明晰平均数的类型？

猜想

也许收入使用的是数值较大的均值，以达到利用高收入读者群吸引广告商的目的。



案例 | 《时代》杂志的订户

“编者的话”

旧订户 年龄的中位数是 41 岁，家庭平均年收入为 9535 美元。

新订户 年龄的中位数是 34 岁，家庭平均年收入为 7270 美元。

疑问

为什么两次谈到年龄时都指出采用了中位数，而关于收入却不明確平均数的类型？

猜想

也许收入使用的是数值较大的均值，以达到利用高收入读者群吸引广告商的目的。



案例 | 《时代》杂志的订户

“编者的话”

旧订户 年龄的中位数是 41 岁，家庭平均年收入为 9535 美元。

新订户 年龄的中位数是 34 岁，家庭平均年收入为 7270 美元。

疑问

为什么两次谈到年龄时都指出采用了中位数，而关于收入却不明確平均数的类型？

猜想

也许收入使用的是数值较大的均值，以达到利用高收入读者群吸引广告商的目的。



可怜的斯威士兰人？

斯威士兰人的平均寿命低得吓人，男性为 32 岁，女性为 33 岁（32 岁并不是一个常见的死亡年龄，大多数活过 32 岁的人都可以活得更久）

为了要计算平均值，他们与因为缺乏医疗资源而夭折的婴儿并在一起计算，而这些早逝天使的数量又很多，于是斯威士兰人的平均寿命就被婴儿的高死亡率拉低了。



可怜的斯威士兰人？

斯威士兰人的平均寿命低得吓人，男性为 32 岁，女性为 33 岁（32 岁并不是一个常见的死亡年龄，大多数活过 32 岁的人都可以活得更久）

夭折的婴儿！

为了要计算平均值，他们与因为缺乏医疗资源而夭折的婴儿并在一起计算，而这些早逝天使的数量又很多，于是斯威士兰人的平均寿命就被婴儿的高死亡率拉低了。



可怜的斯威士兰人？

斯威士兰人的平均寿命低得吓人，男性为 32 岁，女性为 33 岁（32 岁并不是一个常见的死亡年龄，大多数活过 32 岁的人都可以活得更久）

夭折的婴儿！

为了要计算平均值，他们与因为缺乏医疗资源而夭折的婴儿并在一起计算，而这些早逝天使的数量又很多，于是斯威士兰人的平均寿命就被婴儿的高死亡率拉低了。



可怜的斯威士兰人？

斯威士兰人的平均寿命低得吓人，男性为 32 岁，女性为 33 岁（32 岁并不是一个常见的死亡年龄，大多数活过 32 岁的人都可以活得更久）

夭折的婴儿！

为了要计算平均值，他们与因为缺乏医疗资源而夭折的婴儿并在一起计算，而这些早逝天使的数量又很多，于是斯威士兰人的平均寿命就被婴儿的高死亡率拉低了。



案例 | 平均寿命

可怜的斯威士兰人？

斯威士兰人的平均寿命低得吓人，男性为 32 岁，女性为 33 岁（32 岁并不是一个常见的死亡年龄，大多数活过 32 岁的人都可以活得更久）

夭折的婴儿！

为了要计算平均值，他们与因为缺乏医疗资源而夭折的婴儿并在一起计算，而这些早逝天使的数量又很多，于是斯威士兰人的平均寿命就被婴儿的高死亡率拉低了。

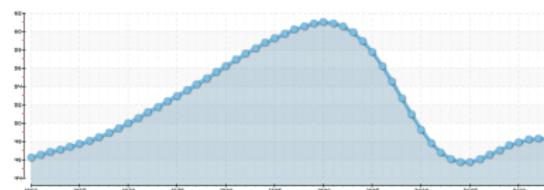


案例 | 平均寿命

斯威士兰 - 男性-出生时的预期寿命 (岁)



斯威士兰 - 女性-出生时的预期寿命 (岁)

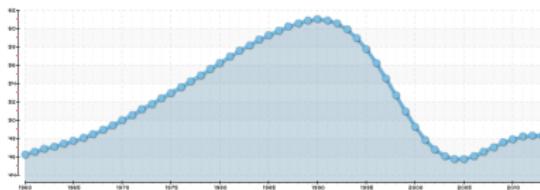


案例 | 平均寿命

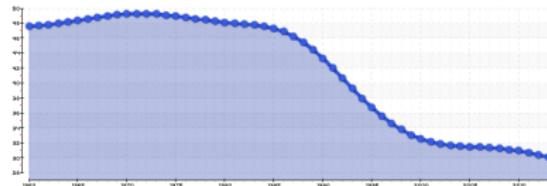
新西兰 - 男性 - 出生时的预期寿命 (岁)



新西兰 - 女性 - 出生时的预期寿命 (岁)

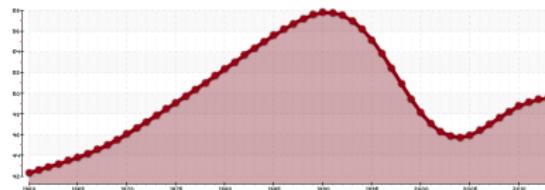


新西兰 - 出生率 (每1000人)

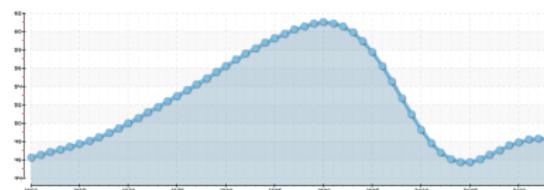


案例 | 平均寿命

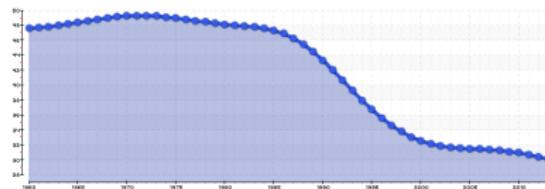
斯威士兰 - 男性-出生时的预期寿命 (岁)



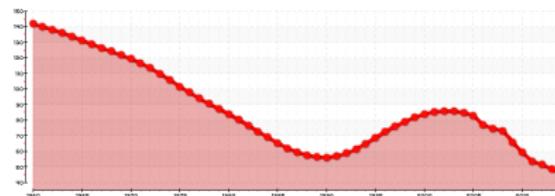
斯威士兰 - 女性-出生时的预期寿命 (岁)



斯威士兰 - 出生率 (每1000人)



斯威士兰 - 婴儿死亡率 (每1000个活产婴儿)

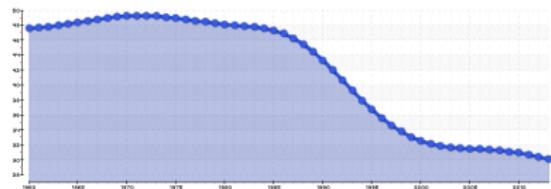


案例 | 平均寿命

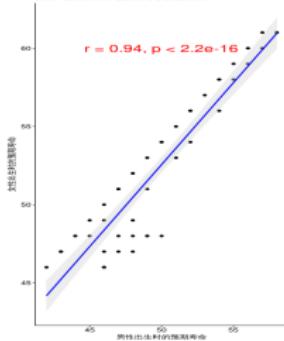
斯威士兰 - 男性-出生时的预期寿命 (岁)



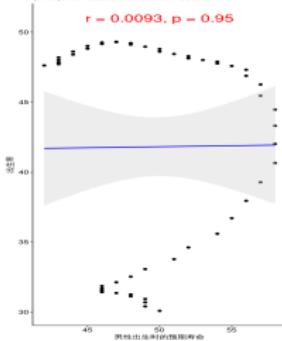
斯威士兰 - 出生率 (每1000人)



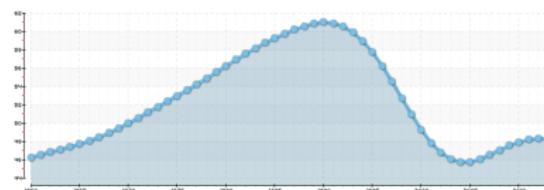
男性与女性出生时预期寿命的相关性:



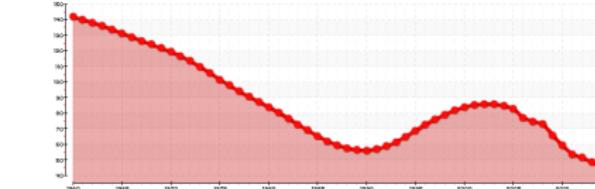
(男性)出生时预期寿命与出生率的相关性:



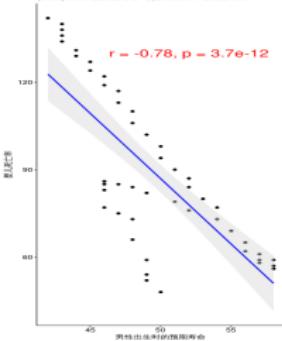
斯威士兰 - 女性-出生时的预期寿命 (岁)



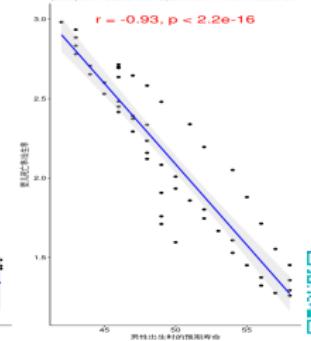
斯威士兰 - 婴儿死亡率 (每1000个活产婴儿)



(男性)出生时预期寿命与婴儿死亡率的相关性:



(男性)出生时预期寿命与婴儿死亡率与出生率的相关性:



引言

事物实际上通常呈现两极化，但平均值把这两极拉在一起。

癌症存活率

- 古尔德，腹部间皮癌（1985 年提出“相同生物个体间存在差异”的观点。他指出，以往对癌症病人存活率的统计让人忽略了病人个体情况的研究。）
- 存活时间的中位数只有 8 个月：有一半的患者活不过 8 个月，另一半的人可以活得更久
- 存活时间的排行，在中位数右方的曲线，有一条延伸数年的尾巴
- 存活了 20 年（因为另一种不相关的癌症而去世）



引言

事物实际上通常呈现两极化，但平均值把这两极拉在一起。

癌症存活率

- 古尔德，腹部间皮癌（1985 年提出 “相同生物个体间存在差异”的观点。他指出，以往对癌症病人存活率的统计让人忽略了病人个体情况的研究。）
- 存活时间的中位数只有 8 个月：有一半的患者活不过 8 个月，另一半的人可以活得更久
- 存活时间的排行，在中位数右方的曲线，有一条延伸数年的尾巴
- 存活了 20 年（因为另一种不相关的癌症而去世）



案例 | 预产期为什么都不准？

预产期

从最后一次月经来潮的第一天开始，往后推算 280 天（平均值）。

预产期不准确的原因

- 有些孕妇会早产
- 几乎没有孕妇可以过了预产期两周以上，而医生还不实施人工催生

早产会让怀孕期的平均值下降，延迟则会拉高平均值的天数，但我们用人为的力量介入，阻止宝宝晚于平均值两周以上出生。这种只计算每个早产儿却排除晚生的宝宝所造成的不平衡效果，会使怀孕期的平均值低于自然的天数。

280 天理论的由来

我们的行为以平均值作为依据，但这个数字之所以会变成平均值，却是我们的行为所造成的。

案例 | 预产期为什么都不准？

预产期

从最后一次月经来潮的第一天开始，往后推算 280 天（平均值）。

预产期不准确的原因

- 有些孕妇会早产
- 几乎没有孕妇可以过了预产期两周以上，而医生还不实施人工催生

早产会让怀孕期的平均值下降，延迟则会拉高平均值的天数，但我们用人为的力量介入，阻止宝宝晚于平均值两周以上出生。这种只计算每个早产儿却排除晚生的宝宝所造成的不平衡效果，会使怀孕期的平均值低于自然的天数。

280 天理论的轮回

我们的行为以平均值作为依据，但这个数字之所以会变成平均值，却是我们的行为所造成的。

案例 | 预产期为什么都不准？

预产期

从最后一次月经来潮的第一天开始，往后推算 280 天（平均值）。

预产期不准确的原因

- 有些孕妇会早产
- 几乎没有孕妇可以过了预产期两周以上，而医生还不实施人工催生

早产会让怀孕期的平均值下降，延迟则会拉高平均值的天数，但我们用人为的力量介入，阻止宝宝晚于平均值两周以上出生。这种只计算每个早产儿却排除晚生的宝宝所造成的不平衡效果，会使怀孕期的平均值低于自然的天数。

280 天理论的轮回

我们的行为以平均值作为依据，但这个数字之所以会变成平均值，却是我们的行为所造成的。

案例 | 预产期为什么都不准？

预产期

从最后一次月经来潮的第一天开始，往后推算 280 天（平均值）。

预产期不准确的原因

- 有些孕妇会早产
- 几乎没有孕妇可以过了预产期两周以上，而医生还不实施人工催生

早产会让怀孕期的平均值下降，延迟则会拉高平均值的天数，但我们用人为的力量介入，阻止宝宝晚于平均值两周以上出生。这种只计算每个早产儿却排除晚生的宝宝所造成的不平衡效果，会使怀孕期的平均值低于自然的天数。

280 天理论的轮回

我们的行为以平均值作为依据，但这个数字之所以会变成平均值，却是我们的行为所造成的。

案例 | 预产期为什么都不准？

预产期

从最后一次月经来潮的第一天开始，往后推算 280 天（平均值）。

预产期不准确的原因

- 有些孕妇会早产
- 几乎没有孕妇可以过了预产期两周以上，而医生还不实施人工催生

早产会让怀孕期的平均值下降，延迟则会拉高平均值的天数，但我们用人为的力量介入，阻止宝宝晚于平均值两周以上出生。这种只计算每个早产儿却排除晚生的宝宝所造成的不平衡效果，会使怀孕期的平均值低于自然的天数。

280 天理论的轮回

我们的行为以平均值作为依据，但这个数字之所以会变成平均值，却是我们的行为所造成的。

案例 | 预产期为什么都不准？

总结

合并过早与过晚两个极端边界值的结果，产生了不准确的预产期。

大规模研究

40 万名的瑞典妇女，大部分的准妈妈们都超过 280 天还没生产。到了第 282 天（中位数），有一半的婴儿已经出生；但人数最多的怀孕期，也是最能够当做每个人可能的怀孕期，是 283 天（众数）。

结论

近期研究证实的数据：最常见也最可能的怀孕期是 283 天。



案例 | 预产期为什么都不准？

总结

合并过早与过晚两个极端边界值的结果，产生了不准确的预产期。

大规模研究

40 万名的瑞典妇女，大部分的准妈妈们都超过 280 天还没生产。到了第 282 天（中位数），有一半的婴儿已经出生；但人数最多的怀孕期，也是最能够当做每个人可能的怀孕期，是 283 天（众数）。

结论

近期研究证实的数据：最常见也最可能的怀孕期是 283 天。



案例 | 预产期为什么都不准？

总结

合并过早与过晚两个极端边界值的结果，产生了不准确的预产期。

大规模研究

40 万名的瑞典妇女，大部分的准妈妈们都超过 280 天还没生产。到了第 282 天（中位数），有一半的婴儿已经出生；但人数最多的怀孕期，也是最能够当做每个人可能的怀孕期，是 283 天（众数）。

结论

近期研究证实的数据：最常见也最可能的怀孕期是 283 天。



1 平均数

2 花言巧语的收入

3 其他案例分析

4

安全的旅行方式

5

知识拓展

6

图说天下

7

统计知识



案例 | 安全的旅行方式

问题

人们在路上旅行时究竟采用哪一种方式更加安全，是乘坐飞机还是乘坐火车？（我们从一开始就把汽车看做是头号杀手。）

新闻报道

新闻媒体的报道好像乘坐飞机是一件十分可怕的事情。其实，飞行安全与否与我们是否经常翻阅报纸和杂志并没有直接的联系。

小概率事件

小概率事件在一次试验中发生的机会非常小，但是，如果做了许多次试验，它必然发生。



案例 | 安全的旅行方式

问题

人们在路上旅行时究竟采用哪一种方式更加安全，是乘坐飞机还是乘坐火车？（我们从一开始就把汽车看做是头号杀手。）

新闻报道

新闻媒体的报道好像乘坐飞机是一件十分可怕的事情。其实，飞行安全与否与我们是否经常翻阅报纸和杂志并没有直接的联系。

小概率事件

小概率事件在一次试验中发生的机会非常小，但是，如果做了许多次试验，它必然发生。



案例 | 安全的旅行方式

问题

人们在路上旅行时究竟采用哪一种方式更加安全，是乘坐飞机还是乘坐火车？（我们从一开始就把汽车看做是头号杀手。）

新闻报道

新闻媒体的报道好像乘坐飞机是一件十分可怕的事情。其实，飞行安全与否与我们是否经常翻阅报纸和杂志并没有直接的联系。

小概率事件

小概率事件在一次试验中发生的机会非常小，但是，如果做了许多次试验，它必然发生。



案例 | 安全的旅行方式 | 飞机更安全

飞机更安全

从平均意义上来看，也就是从基本的证据来看，可以说乘坐飞机旅行而遇难的人的确比乘坐火车遇难的人少。

统计数据

每 100 亿乘客公里数：

火车 9 人遇难

飞机 3 人遇难

疑问

- 如果这种统计数据正确，那么，乘坐火车旅行而遇难的人将会是乘坐飞机旅行遇难人数的 3 倍？
- 可是为什么当我们登上飞机的时候会出冷汗，而在乘坐火车时却不会这样？

案例 | 安全的旅行方式 | 飞机更安全

飞机更安全

从平均意义上来看，也就是从基本的证据来看，可以说乘坐飞机旅行而遇难的人的确比乘坐火车遇难的人少。

统计数据

每 100 亿乘客公里数：

火车 9 人遇难

飞机 3 人遇难

疑问

- 如果这种统计数据正确，那么，乘坐火车旅行而遇难的人将会是乘坐飞机旅行遇难人数的 3 倍？
- 可是为什么当我们登上飞机的时候会出冷汗，而在乘坐火车时却不会这样？

案例 | 安全的旅行方式 | 飞机更安全

飞机更安全

从平均意义上来看，也就是从基本的证据来看，可以说乘坐飞机旅行而遇难的人的确比乘坐火车遇难的人少。

统计数据

每 100 亿乘客公里数：

火车 9 人遇难

飞机 3 人遇难

疑问

- 如果这种统计数据正确，那么，乘坐火车旅行而遇难的人将会是乘坐飞机旅行遇难人数的 3 倍？
- 可是为什么当我们登上飞机的时候会出冷汗，而在乘坐火车时却不会这样？

案例 | 安全的旅行方式 | 火车更安全

理性思考

从理性的角度出发，感兴趣的不是下一个千公里会遇难的可能性，而是在下一个小时内是否会遇难。

统计数据

每 1 亿乘客小时数：

火车 7 人遇难

飞机 24 人遇难

结论

乘坐飞机旅行每小时所产生的死亡事故是乘坐火车旅行的 3 倍以上。



理性思考

从理性的角度出发，感兴趣的不是下一个千公里会遇难的可能性，而是在下一个小时内是否会遇难。

统计数据

每 1 亿乘客小时数：

火车 7 人遇难

飞机 24 人遇难

结论

乘坐飞机旅行每小时所产生的死亡事故是乘坐火车旅行的 3 倍以上。



案例 | 安全的旅行方式 | 火车更安全

理性思考

从理性的角度出发，感兴趣的不是下一个千公里会遇难的可能性，而是在下一个小时内是否会遇难。

统计数据

每 1 亿乘客小时数：

火车 7 人遇难

飞机 24 人遇难

结论

乘坐飞机旅行每小时所产生的死亡事故是乘坐火车旅行的 3 倍以上。



案例 | 安全的旅行方式 | 总结

每10亿次出行死亡数	每10亿小时死亡数	每10亿千米死亡数
公交车: 4.3	公交车: 11.1	航空: 0.05
火车: 20	火车: 30	公交车: 0.4
面包车: 20	航空: 30.8	火车: 0.6
小轿车: 40	水上交通: 50	面包车: 1.2
步行: 40	面包车: 60	水上交通: 2.6
水上交通: 90	小轿车: 130	小轿车: 3.1
航空: 117	步行: 220	航天飞机: 5.99
自行车: 170	自行车: 550	自行车: 44.6
摩托车: 1640	摩托车: 4840	步行: 54.2
航天飞机: 14925373	航天飞机: 63,128	摩托车: 108.9



- 1 平均数
- 2 花言巧语的收入
- 3 其他案例分析

- 4 安全的旅行方式
- 5 知识拓展
- 6 图说天下
- 7 统计知识



本福特定律

本福特定律，也称为本福特法则，说明一堆从实际生活得出的数据中，以 1 为首位数字的数的出现概率约为总数的三成，接近直觉得出之期望值 $1/9$ 的 3 倍。推广来说，越大的数，以它为首几位的数出现的概率就越低。它可用于检查各种数据是否有造假。

应用

- 1972 年，Hal Varian 提出这个定律来用作检查支持某些公共计划的经济数据是否有欺瞒之处。
- 1992 年，Mark J. Nigrini 提出以它检查是否有伪帐。
- 推而广之，它能用于在会计、金融甚至选举中出现的数据，应用于欺骗检测和股票市场分析等领域。



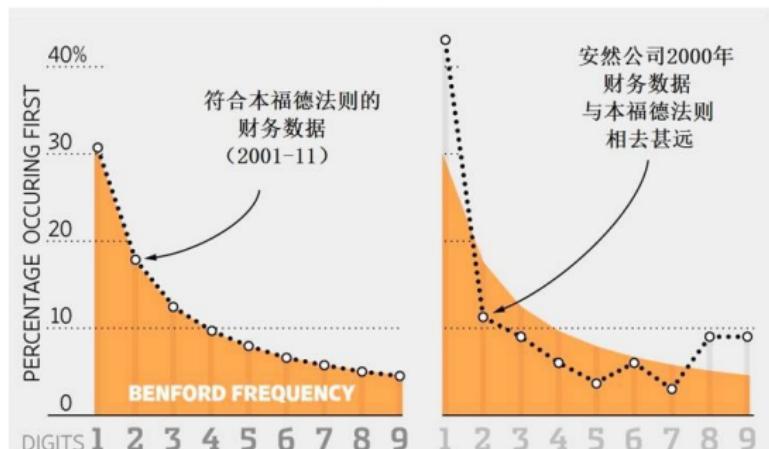
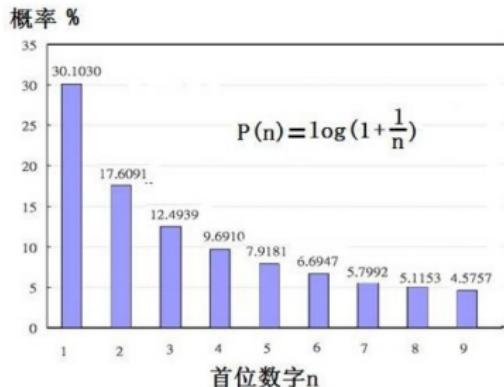
本福特定律

本福特定律，也称为本福特法则，说明一堆从实际生活得出的数据中，以 1 为首位数字的数的出现概率约为总数的三成，接近直觉得出之期望值 $1/9$ 的 3 倍。推广来说，越大的数，以它为首几位的数出现的概率就越低。它可用于检查各种数据是否有造假。

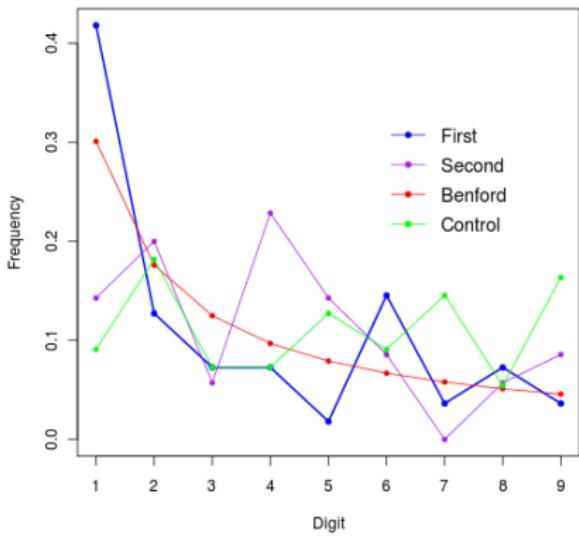
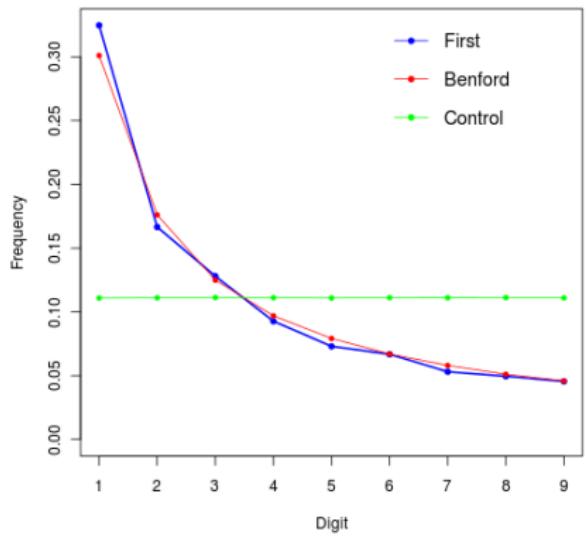
应用

- 1972 年，Hal Varian 提出这个定律来用作检查支持某些公共计划的经济数据是否有欺瞒之处。
- 1992 年，Mark J. Nigrini 提出以它检查是否有伪帐。
- 推而广之，它能用于在会计、金融甚至选举中出现的数据，应用于欺骗检测和股票市场分析等领域。

本福特定律



本福特定律



齐夫定律

齐夫定律（1949 年，实验定律，非理论定律）

- 在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。
- 频率最高的单词出现的频率大约是出现频率第二位的单词的 2 倍，而出现频率第二位的单词则是出现频率第四位的单词的 2 倍。

齐夫定律实例

在 Brown 语料库中，“the”、“of”、“and”是出现频率最前的三个单词，其出现的频数分别为 69971 次、36411 次、28852 次，大约占整个语料库 100 万个单词中的 7%、3.6%、2.9%，其比例约为 6 : 3 : 2。

遵循齐夫定律的现象

- 单词的出现频率（不仅适用于语料全体，也适用于单独的一篇文章）
- 网页访问频率、城市人口、收入前 3% 的人的收入、地震震级、固体破碎时的碎片大小

齐夫定律

齐夫定律（1949 年，实验定律，非理论定律）

- 在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。
- 频率最高的单词出现的频率大约是出现频率第二位的单词的 2 倍，而出现频率第二位的单词则是出现频率第四位的单词的 2 倍。

齐夫定律实例

在 Brown 语料库中，“the”、“of”、“and”是出现频率最前的三个单词，其出现的频数分别为 69971 次、36411 次、28852 次，大约占整个语料库 100 万个单词中的 7%、3.6%、2.9%，其比例约为 6：3：2。

遵循齐夫定律的现象

- 单词的出现频率（不仅适用于语料全体，也适用于单独的一篇文章）
- 网页访问频率、城市人口、收入前 3% 的人的收入、地震震级、固体破碎时的碎片大小

齐夫定律

齐夫定律（1949 年，实验定律，非理论定律）

- 在自然语言的语料库里，一个单词出现的频率与它在频率表里的排名成反比。
- 频率最高的单词出现的频率大约是出现频率第二位的单词的 2 倍，而出现频率第二位的单词则是出现频率第四位的单词的 2 倍。

齐夫定律实例

在 Brown 语料库中，“the”、“of”、“and”是出现频率最前的三个单词，其出现的频数分别为 69971 次、36411 次、28852 次，大约占整个语料库 100 万个单词中的 7%、3.6%、2.9%，其比例约为 6：3：2。

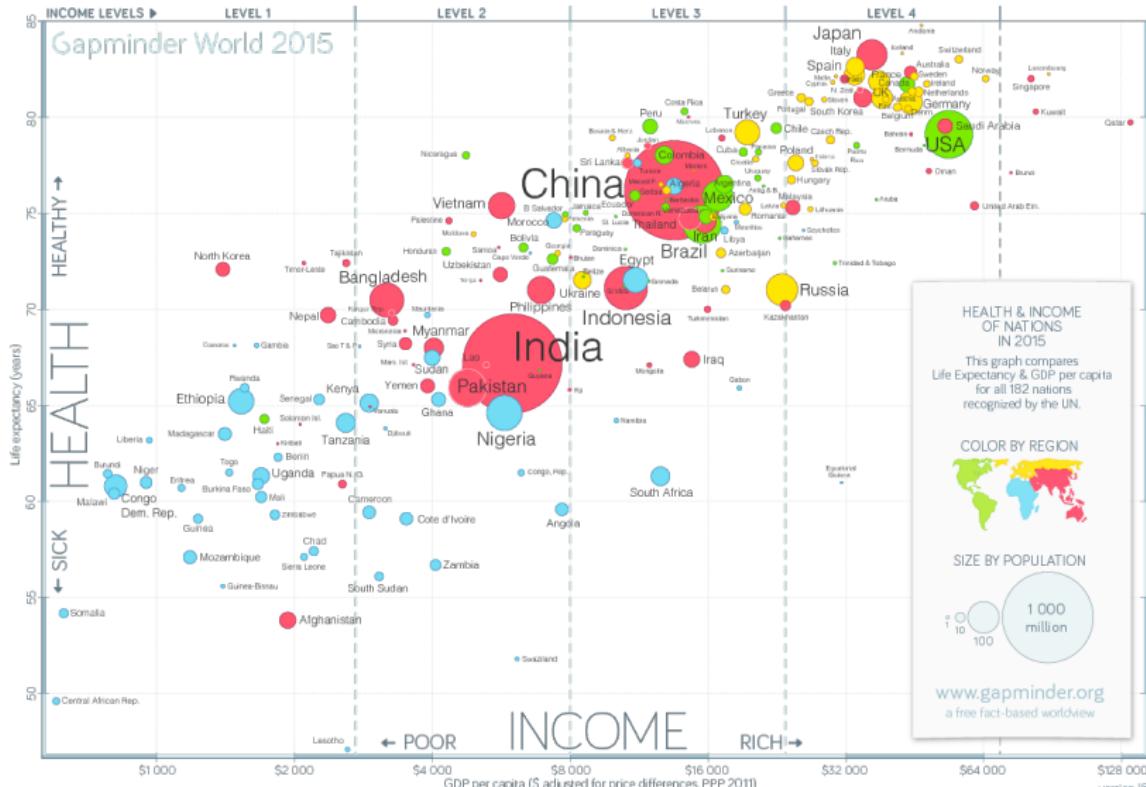
遵循齐夫定律的现象

- 单词的出现频率（不仅适用于语料全体，也适用于单独的一篇文章）
- 网页访问频率、城市人口、收入前 3% 的人的收入、地震震级、固体破碎时的碎片大小

- 1 平均数
- 2 花言巧语的收入
- 3 其他案例分析

- 4 安全的旅行方式
- 5 知识拓展
- 6 图说天下
- 7 统计知识





DATA SOURCES—INCOME: World Bank's GDP per capita, PPP (2011) International \$. Income of Syria & Cuba are Gapminder estimates. It uses log-scale to make a stacking income chart more linear on all levels. POPULATION: Data from UN Population Division. LIFE EXPECTANCY: HNIE-GDB-2015, as of Oct 2015.

ABOUT THIS GRAPH Go to www.gapminder.org/tools to see how this graph changed historically and compare 900 other indicators. LICENSING: Our charts are freely available under Creative Commons Attribution License. Please only share, modify, integrate and embed them, as long as you mention "Based on a chart from gapminder.org"

1 平均数

2 花言巧语的收入

3 其他案例分析

4 安全的旅行方式

5 知识拓展

6 图说天下

7 统计知识



总结

- 平均值的优点在于，它缩小了庞大的信息量，让人得以管理惊人的数字，但也因为这个原因，使得它容易产生误导作用，让人忘记数字的差异性。平均值将所有的失误融为一体，这是它有用的地方，也是它骗人的地方。
- 当你被告知某个数是平均数时，除非能说出它的具体种类——均值，中位数，还是众数，否则你对它的具体涵义仍知之甚少。
- 在处理诸如人类特征的数据时，各种平均数的数值十分接近。这些数据具有正态分布的形态特点。
- 当你看到某个平均收入时，首先问问：是什么样的平均？包括了哪些人？
- 遇到平均值时，要记得问：我们真正感兴趣的，是哪一个群体？在问平均薪资时，也许我们不会想知道金字塔顶端的人收入有多少，只想了解普通人的状况。
- 平均值是一个概略，有用，但所有的概略都是一个样。如果你不知道它只是个摘要，就会被它误导。平均值——是什么东西的平均值？请记得现实世界的多样性，别忘了彩虹原本是白色的。



平均值 vs. 世界的多样化

- 平均值只能说个大概，这是它们天生的说话方式。用这么简单的方式来描述一个群体，重要的细节无可避免就会被含糊带过。平均值就像晚间新闻的一句结语，简单地说出记者所归纳的重点，但是没看过报道的观众，还是不知道具体的新闻内容是什么。它向来有个问题，宣称自己为这个万花筒般的世界发声，却扼杀了我们所有的想象。**想要看穿平均值的真面目，首先你得记得世界的多样化。**
- 平均薪资、平均房价、平均寿命、平均犯罪率，这些以平均为首的名词，以及其他同等性质的平均数，如通货膨胀率等，全都把数字单一化，将一堆数据总结成一个代表总体的数字或是图上的一点。平均值既有优点，也有缺点，它铲平山丘、填满坑谷，让我们误以为地球真的是平的。事实上，地球是起伏不平的。
- 彩虹原本是白色的，经过折射和反射之后，白色的太阳光线变成了神奇的七彩颜色。每当你看见平均值时，请想起“彩虹是白色的”这句话，并想象经过折射、反射出现的七彩霓虹。

Powered by



T_EX L^AT_EX X_ET_EX Beamer

