心理统计

第四讲：集中量数和差异量数

严超赣

Chao-Gan Yan, Ph.D.

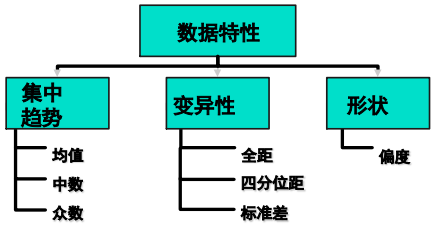
yancg@psych.ac.cn
http://rfmri.org/yan

Institute of Psychology, Chinese Academy of Sciences

1

1

数据的特性

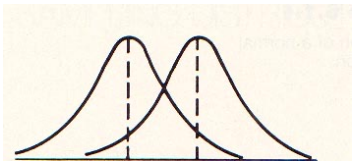


2

2

数据的特性

集中趋势指的是分布的中间位置的情况

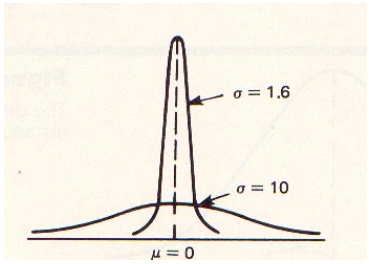


3

3

数据的特性

离散性指的是分布有多分散



4

4

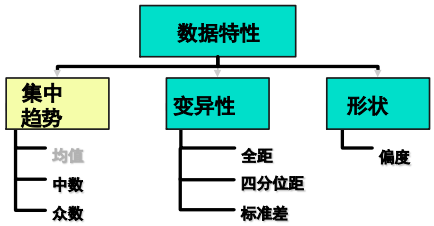
学习目标

- 学会计算均值，中数和众数
- 对于给定的分数分布，确定如何选用适宜的集中量数
- 学会计算标准差, 四分位距和全距
- 对于给定的分数分布，确定如何选用适宜的差异量数

5

5

均值



6

6

均值 （Mean）

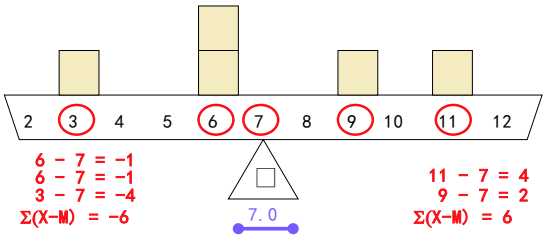
- 亦称算术平均数（arithmetic average）
- 总体的均值公式：
  μ = Σ X/ *N*
- 样本的均值公式：
  $\bar{X} = \Sigma X / n$

7

Student	Score
1	6
2	3
3	11
4	9
5	6
$\sum X$	35
$\frac{\sum X}{N}$	7

8

平衡原则



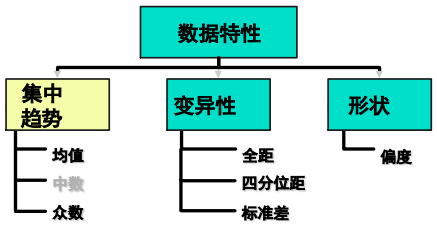
9

均值具有下列特征

- 如果改变一个给定的分数，增加一个被试，或减少一个被试，均值应当有变化.
- 如果对每一个分数都加上（或减去）一个常数，均值也会加上（或减去）这个常数。 - 如果对每个人收取20元管理费. 他们分红的均值会是200-20=180元。
- 如果对每一个分数都乘以（或除以）一个常数，均值也会乘以（或除以）这个常数。

10

中数



11

2．中数（*median*）

- 中数（*median*）是将分数分布均分为两部分的那个分数. 分布有50% 的个体等于或小于中数. 中数等价于百分位数（percentile）是50.
- 中数将分布分为两个大小相等的组

12

求中数的三种情况

- 如果分数的个数是奇数个，将其按从小到大的顺序排列．中间的数目就是中数
- 如果分数的个数是偶数个，将其按从小到大的顺序排列．然后找出中间的两个分数。将其相加后再除以2
- 当分布的中间分数有相等的分数时，用中间分数的精确上下限作插值法

13

13

中数的计算

- 计算下列连续型变量的中数？

a) 8, 10, 12, 15, 18, 19, 60
b) 8, 10, 12, 15, 16, 18, 19, 60
c) 1、2、2、3、4、4、4、4、4、5

14

14

插值法

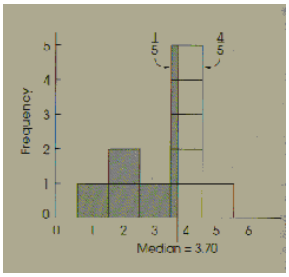
X	f	%	C%
5	1	10	100
4	5	50	90
?			50
3	1	10	40
2	2	20	30
1	1	10	10

15

15

练习

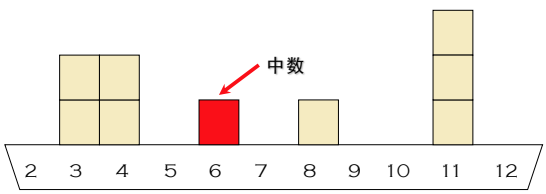
求数列1、2、2、3、4、4、4、4、4、5的中数 。



16

16

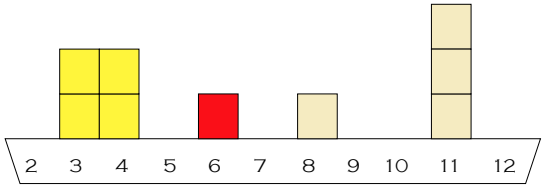
中数



17

17

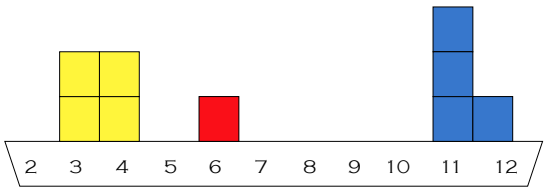
中数



18

18

中数

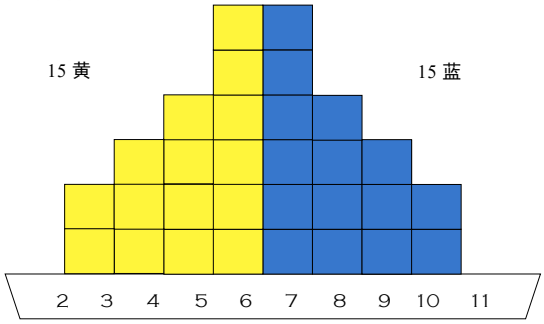


极端分数不影响中数.

19

19

对称分布中的中数

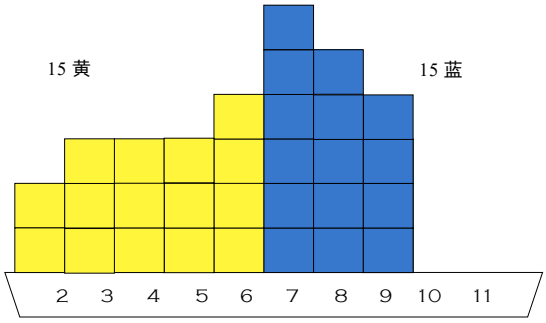


Mdn = 6.5

20

20

偏态分布中的中数

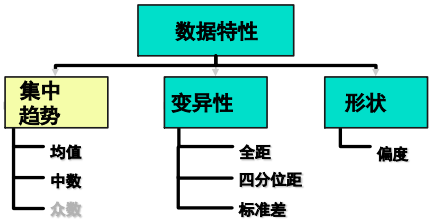


Mdn = 6.5

21

21

众数



22

22

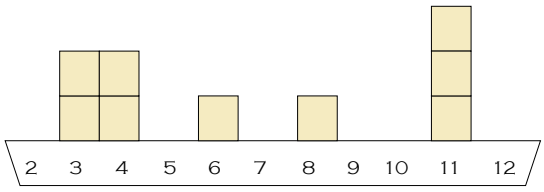
3. 众数 （mode）

- 在次数分布中, 众数是具有最多次数的那个分数或类目。
- 注意：一个次数分布可能有多个众数.
- 是类目变量可以选用的唯一集中量数

23

23

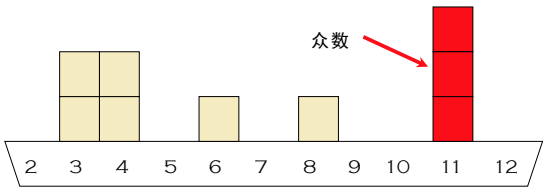
众数



24

24

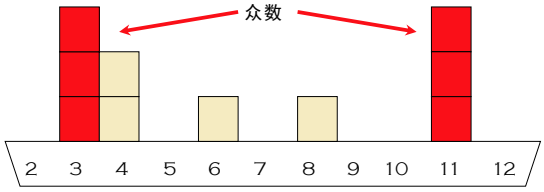
众数



25

25

众数



26

26

选择的适宜集中量数

- 命名型变量 ⇒ 众数
- 顺序型变量 ⇒ 中数
- 等距或以上变量 ⇒ 均值 (分布正态)
  ⇒ 中数(分布偏态)

27

27

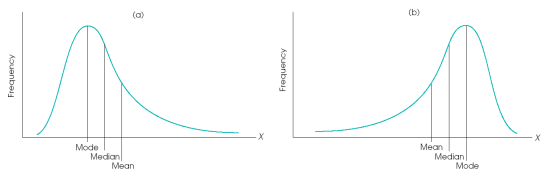
集中量数的优缺点

	+	-
众数	计算快捷 对于命名型数据特别有用	样本稳定性差
中数	不易受极端分数的影响	在一定程度上样本稳定性差
均值	样本稳定性好 与方差有关	对于离散型数据不适用 受极端数值的影响

28

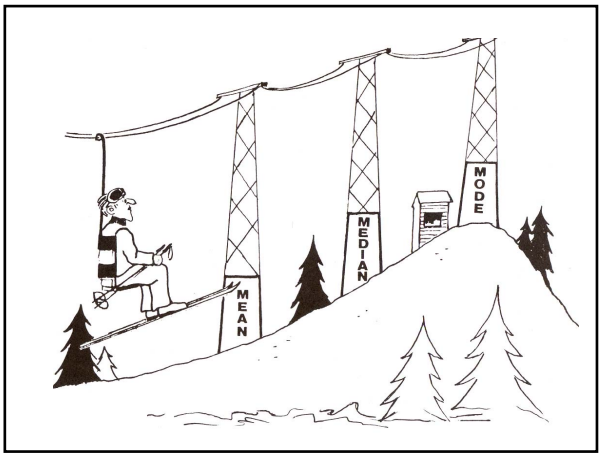
28

分布形状与集中趋势量数的关系



29

29



30

两个均值相同的正态分布

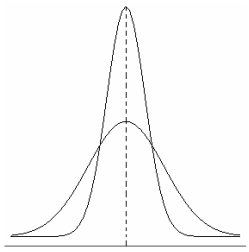


图1－4 两种分布的比较

31

31

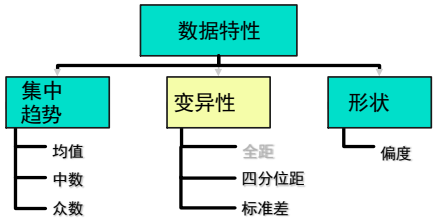
差异量数（Variability）

- 分布的第三个特征 ----变异性（Variability）.
  - 变异数是对于分布的延伸和聚合状态程度的定量化描述
  - 变异数越高，表明分数间的差别大, 变异数越小，表明分数间越近似.
- 三种差异量数： 全距 （range）, 标准差 （standard deviation）, 和四分位距（interquartile range）.

32

32

全距



33

33

1. 全距 （range）

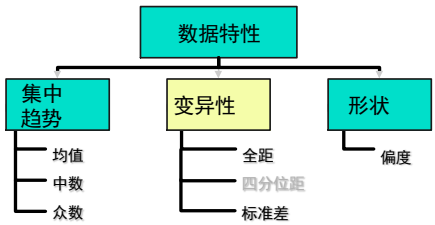
- 全距是分布分数最大值(maximum) X的精确上限与分布分数最小值(minimum) X 的精确下限的差值。
- 用全距描述分数变异性的局限：
- 该统计量只依据分布中的两个极端值，未利用到分布的大部分信息.
- 注意：如果分数是连续型，必须用精确上下限。

若 X是离散型:
range = 10 - 5 = 5
若 X是连续型:
range = 10.5- 4.5

34

34

四分位距



35

35

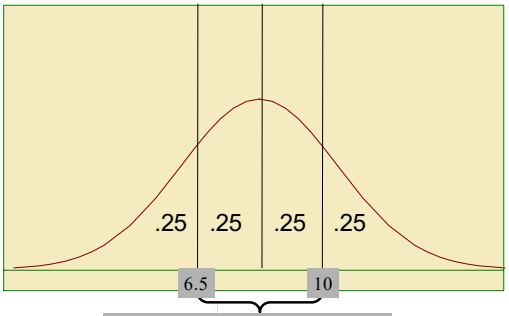
2. 四分位距(interquartile range）

- 度量变异数的另一种方法.
  - 50%，25%和75%的百分位数代表什么？
  - 用50%，25%和75%的百分位数,分布被分成4部分

36

36

Semi-interquartile Range



Interquartile range = 10 - 6.5 = 3.5

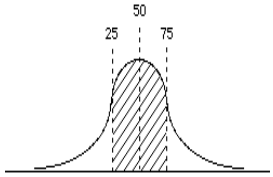
Semi-interquartile range = 3.5/2 = 1.75

37

37

四分位距的计算

X	f	%	c%
7	4	12.5	100
6	4	12.5	87.5
5	4	12.5	75
4	8	25	62.5
3	4	12.5	37.5
2	4	12.5	25
1	4	12.5	12.5



38

38

四分位距的计算

- 四分位距就是75%百分位数与25%百分位数间的距离. 它代表分布中间50%的距离.
- 如果上例是连续型变量，

median = Q2 = 4.0 -> 用插入法

25%tile = Q1 = 2.5 -> 区间2 的精确上限

75%tile = Q3 = 5.5 -> 区间5 的精确上限

四分位距 (IQR) = 5.5 - 2.5 = 3.0

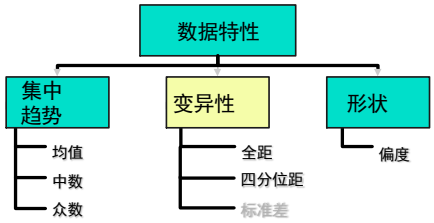
- semi-interquartile range：四分位距的一半（interquartile range）.

SIQR = (Q3 - Q1)/2

39

39

标准差



40

40

标准差 (standard deviation)

- 量度了分布中的每一个个体与某一标准偏移的距离, 这个标准就是均值
- 最重要, 最常用的差异量数
- 考虑了分布中的所有信息

41

41

方差/标准差的逻辑步骤-1

- X - μ = 离差分数（deviation score）
  - -如果分数的值大于均值 离差是正数
  - -如果分数的值小于均值 离差是负数

	Score	Deviation
张	10	-40
王	20	-30
李	30	-20
赵	40	-10
刘	50	0
胡	60	10
彭	70	20
许	80	30
陆	90	40
AVERAGE	50	

42

42

离差分数

- 为了考察分布的离散程度,我们尝试把离差加和…
- 但是却得到0

	Score	Deviation
张	10	-40
王	20	-30
李	30	-20
赵	40	-10
刘	50	0
胡	60	10
彭	70	20
许	80	30
陆	90	40
SUM		0

43

43

于是我们寻找其他的离差指标

我们将每一个离差平方

再求和

这就是著名的和方 (SS)

$SS=\sum(X-\overline{X})^2$

	Score	Deviation	Deviation^2
张	10	-40	1600
王	20	-30	900
李	30	-20	400
赵	40	-10	100
刘	50	0	0
胡	60	10	100
彭	70	20	400
许	80	30	900
陆	90	40	1600
SUM		0	6000

44

44

和方的计算公式

$$SS=\sum X^2-\frac{(\sum X)^2}{N}$$

此二者为等价。计算公式的优点为可直接利用 X 值(原始分数)。

45

45

方差: 定义公式

- 总体

$$\sigma^2=\frac{\sum(X-\mu)^2}{N}$$

"sigma"

- 样本

$$S^2=\frac{n\sum X^2-(\sum X)^2}{n(n-1)}$$

46

46

总体方差（*Population Variance*）

- 和方的平均， 即和方除以总体的容量.
- 总体方差= σ^2 = SS/N
- 上例中: σ^2 = 6000 / 9 = 666.67

47

47

总体标准差 (standard deviation)

- standard deviation = sqroot(variance) = sqroot(SS/N)
- σ = sqrt(σ^2)
- 上例中: σ = sqrt (666.67) = 25.82

48

48

样本的标准差

注意与总体标准差的不同:

- s =样本的标准差（sample standard deviation）
  用 X（不是 μ） 来计算SS)
- 需要考虑样本常常比其所属的总体较少变异性，标准差的计算需做校正.

49

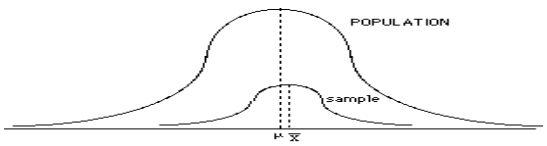
49

有偏估计

- 如果样本统计量高估或低估了总体参数，它就称为有偏估计
- 如果用样本统计量作总体方差，就低估了总体方差，是有偏估计

50

50



- 如果样本有代表性, 那么样本与总体的均值就会非常近似, 两个分布的形状也应该近似。但是, 样本的变异程度却低于总体的变异程度.
- 因此，样本方差的分母是n - 1 而不是 n

  sample variance = $s^2 = \frac{SS}{n-1}$
- 对于样本标准差也是同样
- 样本标准差= s = sqroot(SS/(n - 1))
- 这里我们所做的是用样本来估计总体的性质。因为我们不知道总体均值，我们无法真正度量每个分数与总体的标准之间距离。因此，我们在用总体均值的最佳估计，即样本均值。

51

51

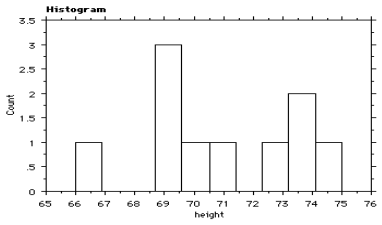
自由度 (degree of freedom)

- 用n-1 作分母，意思是利用自由度来校正样本离差，以利于对总体参数的无偏差估计。
- 什么是自由度? 样本均值是事先已知的，这样在分布中，你只需固定最后一个项目，其他的都可以变化。 n - 1意思是除了一个值，其余都可变化。
- 如: sample mean =5，如果前4 个分数是: 5, 4, 6, 2 最后一个是什么?
- 5 + 4 + 6 + 2 + X = 25
  X = 8 => X必须固定在8。

52

52

粗略估计均值和标准差



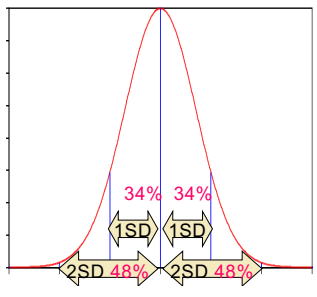
- 拇指原则：对于对称分布，均值常常在分布的中点，标准差常常在全距的1/4左右

53

53

标准差的意义

±SD	% of Pop
1	68.3
1.96	95
2	95.5
2.58	99
3	99.7



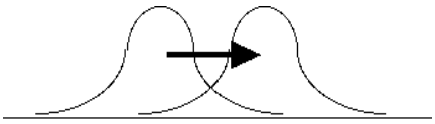
54

54

标准差的性质

1) 对分布中的每一个分数加上一个常数不会改变其标准差.

- 如果对图中的分布每个分数加上2, 均值变化了 (增加 2), 而方差不变。(因为离差不变).



55

55

标准差的性质

2) 对分布中的每一个分数乘上一个常数，所得分布的标准差是原分布的标准差乘上这个常数.

- 如果 mean = 20, 分布中的两个分数分别是21 和 23.
- 新分布中的两个分数是 42和46, 则新分布mean = 40.
- 原分布离差 (21 - 20 = 1) 和 (23 - 20 = 3).
- 新分布离差 (42 - 40 = 2) 和 (46 - 40 = 6).

56

56

比较三种离中量数

- 极端分数: 全距（range） 受影响最大, IQR 受影响最小
- 样本大小: 全距（range） 可能随n 的增加而增加 , IQR & s 不会
- 样本选取：从同一总体中多次取不同样本，全距（range） 没有稳定的值, 但 IQR 和 S 是稳定的，不应波动很大
- 对于有不确定值的分布, 全距 或 S 都无法求得, IQR (或SIQR) 是唯一的选择

57

57

差异量数的优缺点

	+	-
全距	▪ 计算快捷	▪ 样本稳定性差 ▪ 受极端数值的影响 ▪ 可能与样本量有关
四分差	▪ 不易受极端分数的影响 ▪ 适用于有不确定值的数据	▪ 在一定程度上样本稳定性差
标准差	▪ 样本稳定性好 ▪ 包含最多的信息	▪ 受极端数值的影响

58

58

Tips

- 概念和记忆
  - SS： 离差的平方和，计算公式要记忆
  - 方差： 平均和方
  - 标准差：方差的平方根
- 避免错误：
  - 计算前应先对均值和标准差作个粗略的估计
  - 计算SS应作表
  - 不要根据次数分布表计算SS
  - 总体和样本的标准差公式不同，因此应先确定数据是来自总体或样本
  - 在SS的计算公式中，无论总体或样本都是n 而不是 n-1。

59

59

在研究论文中报告集中量数和差异量数

表 4 不同情绪调节方式的执行程度

情绪调节	简单观看组 (n = 25)		评价忽视组 (n = 23)		评价重视组 (n = 24)		表情抑制组 (n = 25)		表情宣泄组 (n = 24)	
方式	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
简单观看	4.16	0.898	1.78	1.278	1.42	1.442	1.40	1.258	1.63	1.408
评价忽视	1.84	1.519	3.91	0.996	0.88	1.154	2.32	1.406	1.25	1.032
评价重视	0.88	1.054	0.70	0.822	4.00	0.885	1.32	1.145	1.17	1.007
表情抑制	1.20	1.000	1.70	1.185	1.75	1.482	4.44	0.507	0.71	0.859
表情宣泄	1.64	1.469	1.74	1.137	2.08	1.412	0.52	0.653	4.21	0.588

60

60

作业

- 有一考试成绩分布，其平均数为71，中数79。问这是一个正态分布，还是正偏态，负偏态？

61

61

作业

- 对于下面的三种情况，请指出能最佳描述其“平均”值的集中量数（平均数、中数、众数）。
  1. 样本为50个6岁儿童，关于他们最喜欢看的电视节目的研究。
  2. 研究某饮食计划对病人的影响，记录6周后他们增加或减少的体重。
  3. 一项关于动机的研究，要求被试在报纸中搜索单词“discipline”。研究者记录被试在找到单词或放弃前所用的时间（单位，分钟）。样本n=20，平均数M=29分钟,中数17分钟，众数为15分钟。

62

62

作业

- 对下面的数据

3，4，4，1，7，3，2，6，4，2，1，6，3，4，5，2，5，4，3，4

1. 画次数分布直方图
2. 指出这组数据的全距（提示：你可以使用全距公式或者只要从直方图的X轴数一下即可。）
3. 指出这组数据的四分位距和四分差。

63

63

作业

- 一个样本 n=25，样本方差 $s^2=100$

1. 求样本标准差
2. 求样本和方SS

64

64

作业

- 下列分数构成一个总体:

  8, 5, 3, 7, 5, 6, 4, 7, 2, 6

  5, 3, 6, 4, 5, 7, 8, 6, 5, 6

1. 绘制次数分布直方图
2. 在图中粗略估计分布的均值和标准差
3. 计算该总体的均值和标准差, 与粗略估计的值

65

65

作业

- 计算下列样本分数 SS, 方差, 和标准差:

  431, 432, 435, 432, 436, 431, 434

66

66