中心趋势的测度

ch02 sec2.1

```
begin
    using Latexify    ,PlutoUI    ,RDatasets    ,DataFrames    ,Random    ,FileIO    ,HTTP
       ,Images    ,Plots    , StatsPlots    ,Statistics    ,StatsBase    ,LaTeXStrings
    ,Symbolics
    TableOfContents(title="中心趋势的测度")
end
```

ch02 sec2.1

> **Info**
>
> 统计学中中心趋势的度量用一个数字来代表数据的中心. 有三种表示方法:
>
> 1. 平均数(mean)
> 2. 中位数(median)
> 3. 众数(mode)

使用反映中心趋势的统计量有利有弊, 有利的是计算简单,只有一个数字, 不利的地方是对中心趋势的度量并不能反映出一个总体或者样本的完整信息. 所以只能作为快速判断趋势的依据, 不能作为最终决策的依据.

> **Example**
>
> example 1 学生考试成绩

10个考生在某考试中的成绩

```
begin
    score=[86, 90, 95, 100, 100, 100, 110, 110, 115, 120]

    score_df=DataFrame(id=1:1:10,score=score)

    describe(score_df[:,2])

end
```

```
Summary Stats:
Length:         10
Missing Count:  0
Mean:           102.600000
Minimum:        86.000000
1st Quartile:   96.250000
Median:         100.000000
3rd Quartile:   110.000000
Maximum:        120.000000
Type:           Int64
```

100

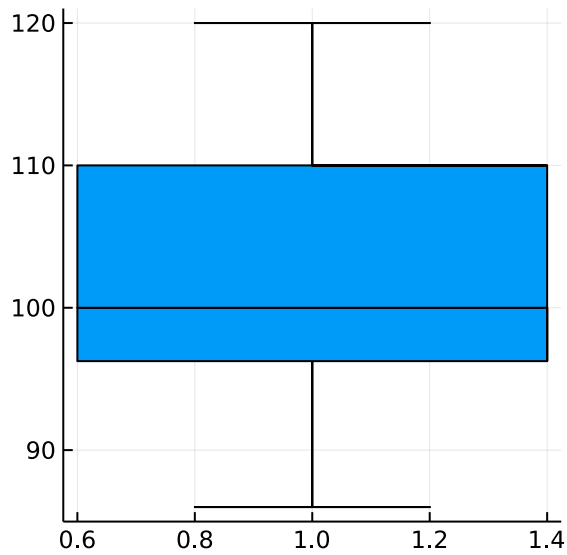
```
mode(score_df[:,2])    # 众数是考试乘积里出现最多的分数
```

100.0

```
median(score_df[:,2])  #中位数也是 100
```

102.6

```
mean(score_df[:,2])    #考试分数均数
```




```
boxplot(score_df[:,2], label=false,size=(300, 300))
```

Example

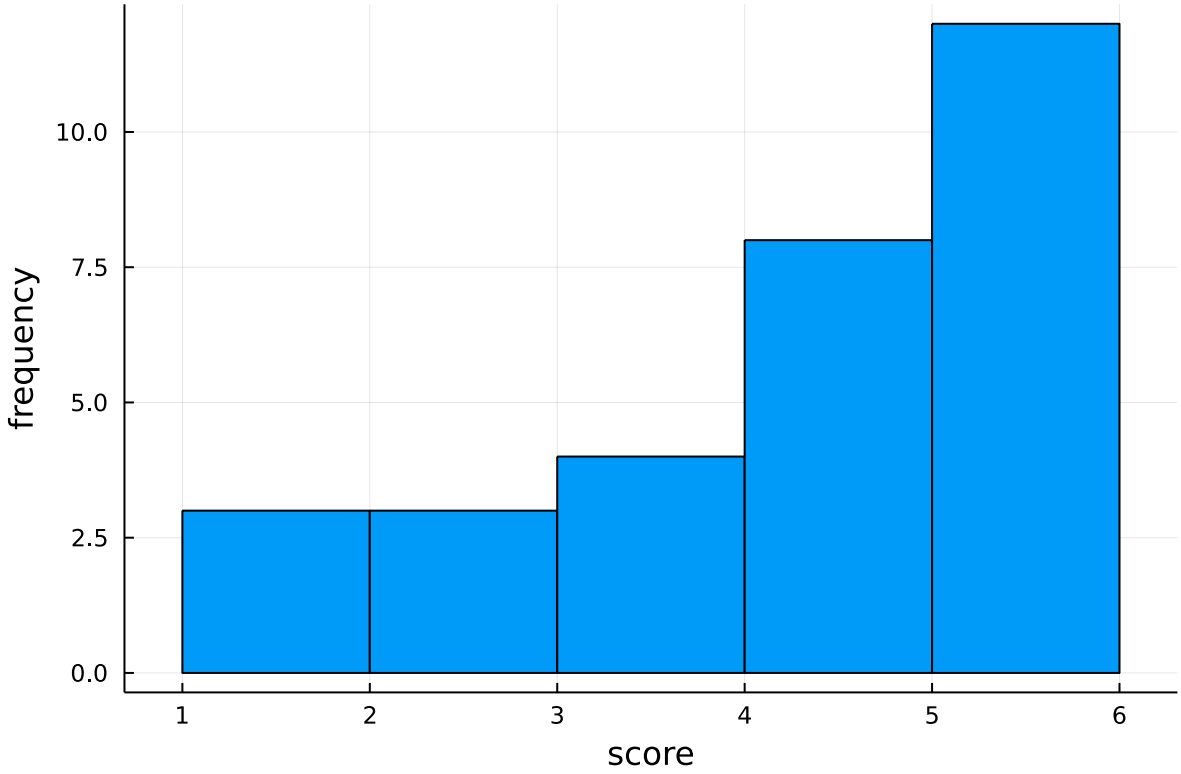
example 2 假设我们想要了解五年级学生对于学习的态度, 给定一个量表, 从不重要-> 重要设置为 1-5 分的等级 随机选取了 30 个学生进行调查.

首先要说明, 这个调查设置是欠考虑的, 无法获取学生的真实意图. 这里只是为了说明统计方法

30 名学生学习态度打分如下:

```
begin
    attitude=[1 1  1   2 2 2  3 3 3 3;
              4 4  4   4 4 4  4 4 5 5;
              5 5  5   5 5 5  5 5 5 5
     ]

    arr=vcat(attitude[1,:],attitude[2,:],attitude[3,:])

     attitude_df=DataFrame(id=1:1:30,score=arr)

     describe(attitude_df[:,2])


end
```

```
Summary Stats:
Length:         30
Missing Count:  0
Mean:           3.766667
Minimum:        1.000000
1st Quartile:   3.000000
Median:         4.000000
3rd Quartile:   5.000000
Maximum:        5.000000
Type:           Int64
```




```
histogram(attitude_df[:,2],xlabel="score", ylabel="frequency",label=false)
```

通过柱形图可以看到得分偏向于较高的一边, 在这种情况下, 均数的反映出的信息就大打折扣了, 单凭均数一个数值无法反映出这个趋势

```
md"通过柱形图可以看到得分偏向于较高的一边， 在这种情况下， 均数的反映出的信息就大打折扣了， 单凭均数一个数值无法反映出这个趋势"
```

```
-0.8325943208748224
```

```
skewness(attitude_df[:,2])    # skewness   为偏度方法,可以反映出数据偏向,负数表示较小得分部分尾部长
```

```
5
```

```
mode(attitude_df[:,2])  #学生选择最多的分数为满分
```

```
4.0
```

```
median(attitude_df[:,2])  #中位数4
```

```
3.7666666666666666
```

```
mean(attitude_df[:,2])  #均数 Wie3.7
```

Definition

总体和样本均值公式

总体均值:

$$\mu = \frac{\sum X}{N}$$

样本均值:

$$\bar{X} = \frac{\sum X}{n}$$

其中符号表示含义: $\bar{X}$: 样本均值, μ : 总体均值, $\sum$:求和符号, X: 单个对象的取值, n:一次采样中获取样本中对象的个数, N 总体中单个对象的个数.

在统计学里费力的定义两套很相似的符号表达式是有深刻内涵的, 有很多的涉及统计的内容混用两种符号形式, 使用者认为两者区别不大, 但是如果理解了样本推断统计学的概念就会觉着这么做是非常错误的. 在选取符号之前一定要思考我现在针对的是一个总体的统计还是针对样本的统计

在两种符号中,都会用到对象的个数, 对于一个有无限个对象的总体, 我们获取不到N,所以根据公式得不到相应的总体均值. 例如工厂里一台机器加工的管道, 由于生产会延续很长时间, 具体会生产多少管道,不得而知. 我们想看看管道的直径的总体均值, 由于N 未知,所以无法计算.

但是总是可以在采样中获取的一个样本中n个对象的值. 因此样本均值μ 总是可以计算出.

上面的工厂里,虽然不能获取所有生产管道的数据, 但是可以抽取一天里生产的管道, 测量直径, 例如随机抽取$n = 30$ 管道, 测量直径. 使用样本均值计算公式来求这次采样的均值. 可以预料到的是在不同工作日抽取的样本的均值是不同的. 生产过程中会有一些未知的因素导致数值会有些变化.

Example

example 3

计算 一组数据的$[6, 2, 4, 2]$ 下列表达式结果

KeySet([$\sum X^2$, $\sum X$, $\sum (X-2)^2$, $\sum X - 2$, $(\sum X)^2$])

结果如下:

$\sum X^2$	60
$\sum X$	14
$\sum (X-2)^2$	20
$\sum X-2$	6
$(\sum X)^2$	196

```
begin
    data2=[6,2,4,2]   # 数据
    substract2(x)=x-2  # 减法
    square(x)=x^2     #平方
    a=data2|>sum     #求和
    b=data2.|>square|>sum  # 对数组元素分别平方，然后求和，针对每个元素的操作用"." 操作符
    c=data2|>sum|>square   # 求和然后平方
    d=data2.|>substract2|>sum  # 每个元素-2，然后求和
    e=data2.|>substract2.|>square|>sum  #每个元素-2,每个元素平方，然后求和
    answer=Dict(L"\sum X"=>a,L"\sum X^2"=>b,L"{(\sum X)}^2"=>c,L"\sum X-2"=>d,L"\sum {(X-2)}^2"=>e)
    description=Dict(L"\sum X"=>"求和",L"\sum X^2"=>"平方->求和",L"{(\sum X)}^2"=>"求和->平方",L"\sum X-2"=>"减2->求和",L"\sum {(X-2)}^2"=>"减2->平方->求和")
    key=keys(answer)  # 取键名
    latexify(answer, env=:mdtable)

end
```

实例3 表达式方法描述如下:

Dict($\sum X^2$ => "平方->求和", $\sum X$ => "求和", $\sum (X-2)^2$ => "减2->平方->求和", $\sum X-2$ => "减2->求和", $(\sum X)^2$ => "求和->平方")

```
description
```