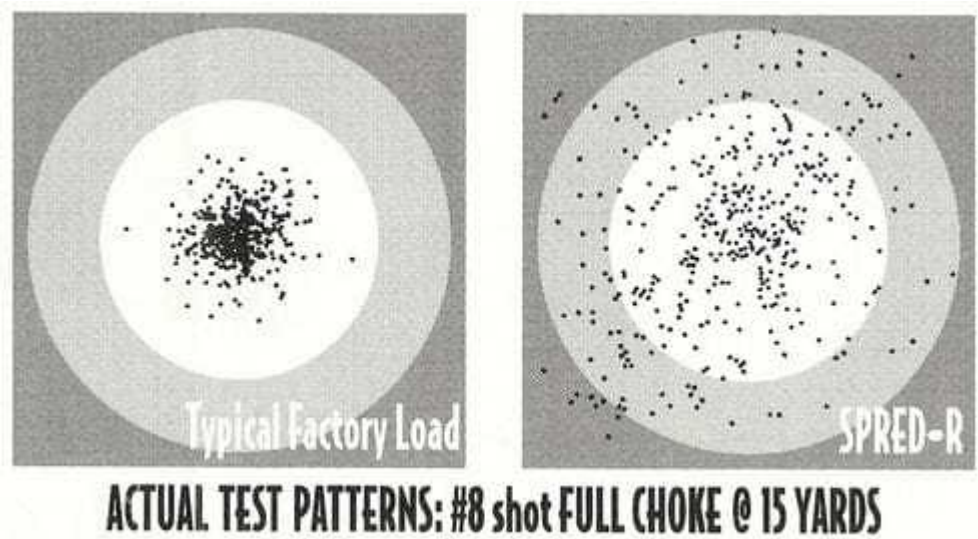
变异程度的测量

ch03 数据集合变异程度的测量

- 极差
  - 离差
- 方差和标准差
  - 平方的意义
  - 样本方差和标准差中的分母
  - 计算机程序的差异



ch03 数据集合变异程度的测量

通过数据集合的均值,可以便捷的提供数据的总体信息, 但是关于个体之间的差异并没有表现出来. 所以需要考虑个体的差异.

从题图可以直观的看到, 每次射击都会偏离圆心, 每个弹着点到靶心的距离就刻画了每次射击的准确程度.

> Info
>
> 统计学中对于数据离散程度有三种表示方法:
>
> 1. 极差(range)
> 2. 方差(variance)
> 3. 标准差(standard deviation)

与统计中心趋势的方法一样, 统计离散程度的方法也要区分是针对总体还是针对抽样样本.

极差

极差统计的是数据集合中某属性最大测量值和最小测量值的差值, 常用的$25\%, 75\%$处的数字也属于这个范畴.

离差

在计算方差和标准差的时候首先要计算离差, 离差的定义: 当个测量值与均值的差值

$$总体离差 = X - \mu$$

$$样本离差 = X - \bar{X}$$

> **Example**
>
> example 1 就散下面一组样本数据的离差

```
sample_data =  [3, 5, 6, 7, 8, 4, 5, 6]
```

```
sample_data=[3,5, 6, 7,8, 4, 5, 6]
```

```
dev_arr =  [-2.5, -0.5, 0.5, 1.5, 2.5, -1.5, -0.5, 0.5]
```

```
dev_arr=sample_data.-mean(sample_data)
```

	X	X-X̄
1	3	-2.5
2	5	-0.5
3	6	0.5
4	7	1.5
5	8	2.5
6	4	-1.5
7	5	-0.5
8	6	0.5

```
begin
df=DataFrame(i=sample_data,x=dev_arr)
rename!(df, Dict(:i => "X", :x => "X-X̄"))
end
```

如果直接统计离差的和, 无论什么数据集合结果都会是 0,所以需要解决这个问题,引入方差就是要解决这个问题

```
md" 如果直接统计离差的和， 无论什么数据集合结果都会是 $0$,所以需要解决这个问题,引入方差就是要解决这个问题"
```

```
0.0
```

```
sum(dev_arr)
```

方差和标准差

方差和标准差直接从公式开始

> **Notice**
>
> 使用符号时, 时刻注意统计的是总体还是样本数据, 一定不要误用公式

	总体	根据样本的估计
方差	$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$ 式中：$\sum$——求和； X——分布中的一个取值； μ——总体均值； N——总体中的对象个数。	$s^2 = \frac{\sum(X-\overline{X})^2}{n-1}$ 式中：$\sum$——求和； X——分布中的一个取值； $\overline{X}$——样本均值； n——样本中的对象个数。
标准差	$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$ 式中：$\sum$——求和； X——分布中的一个取值； μ——总体均值； N——总体中的对象个数。	$s = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}}$ 式中：$\sum$——求和； X——分布中的一个取值； $\overline{X}$——样本均值； n——样本中的对象个数。

平方的意义

在总体和样本的方差, 标准差公式里都出现了单个测量值和均值的平方这一项? 为什么要使用平方?

可以认为当个个体的对量值与均值的差异是一个标量值, 没有方向. 例如在一组考试成绩的数据集合上, 有的学生分数比均值高, 有的学生比均值低. 如果来累积差值, 正负值差异才计算中会被消减掉,这样就反映不出变异程度.

样本方差和标准差中的分母

在 样本方差和标准差中的分母中出现了 $n-1$ 项, 为什么要减 1?

大部分情况下, 采样时获取样本中对象的个数n 会少于 总体中的对象数量N, 如果每个个体都代表一些信息, 那么在绝大多数情况下, 从样本的个体中统计信息都是不完整的. $n-1$ 就是为了减少样本信息和总体信息之间的差异.

要理解这个问题其实也不是太容易, 需要在采样和计算公式上概念上反复多次才能明白到底是怎么一回事.

Example

example 1 下面数据为工作日随机抽取的管道的直径测量值.

```
tube =
  [53.3567, 53.4589, 52.2828, 52.9875, 51.9932, 53.3733, 55.7541, 50.2795, 53.636, 53.5177,
```

```
Summary Stats:
Length:          20
Missing Count:   0
Mean:            52.956206
Minimum:         50.279496
1st Quartile:    52.254033
Median:          53.250832
3rd Quartile:    53.525917
Maximum:         55.754105
Type:            Float64
```

```
1.39282827061105
```

```
• var(tube)  #  方差
```

```
1.1801814566459896
```

```
• std(tube)  # 标准差
```

自定义样本方差和标准差方法

在自定义的计算公式里首先计算离差平方和(sum of squared deviation,SS)

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

对上面的一组数据进行操作

```
SS = 18.0
```

```
SS=dev_arr.^2|>sum
```

方差定义为:离差平方和除以样本数减 1

$$Var = \frac{SS}{n-1}$$

```
Var = 2.5714285714285716
```

```
Var=SS/(length(sample_data)-1)
```

" 标准差定义为方差执行开方操作

$$Std = \sqrt{\frac{SS}{n-1}}$$

```
Std = 1.6035674514745464
```

```
Std=sqrt(Var)
```

看起来比较复杂, 实际从内部到外层执行, 写成一个管道操作形式比较好理解, 管道中每个组件完成一项操作

```
var2 (generic function with 1 method)
```

```
function var2(data)
    n=length(data)    #样本中数据个数
    X̄=mean(data)      # 样本中对象的均值
    sub_mean(x)=x-X̄         # 单个对象与均值的差
    square(x)=x^2     # 平方操作
    div(SS)=SS/(n-1)

    data.|>sub_mean.|>square|>sum|>div
end
```

```
1.3928282706110502
```

```
var2(tube)
```

标准差也采用管道操作

```
std2 (generic function with 1 method)
```

```
std2(data)=data|>var2|>sqrt
```

```
1.1801814566459898
```

```
std2(tube)
```

```
svar (generic function with 1 method)
```

```
svar(data::Vector)=(sum((data.-mean(data)).^2))/(length(data)-1) #另一而种定义
```

```
stdev (generic function with 1 method)
```

```
stdev(data::Vector)=sqrt(svar(data))
```

```
["var" ⇒ 1.39283, "std" ⇒ 1.18018]
```

```
["var"=>svar(tube), "std"=>stdev(tube)]
```

当方差接近于1 的时候, 标准差也接近于 1, 这有点特殊, 但是方差和标准差的单位其实不同, 方差开方以后单位才和原统计值一样. 也就是标准差的单位和统计值单位一样.

利用标准差可以直观的表示测量值和均值之间的平均差异. 但是方差并不是多余的, 因为计算标准差之前一定要先计算方差, 而且在很多统计方法中需要首先计算方差作为中间步骤.

```
[1.0, 1.00499, 1.04881, 1.09545, 1.41421, 2.0]
```

```
[1.0,1.01, 1.1, 1.2,2,4].|>sqrt #接近于1的值开方后也接近于1,但是如果带有物理单位,两者量纲不同
```

计算机程序的差异

我们在表格中提到总体和样本的公式不要混淆使用, 在julia语言方法中没有单独针对总体或者样本的方法,总体和样本的方法作为方法的选项

```
std(itr; corrected::Bool=true, mean=nothing[, dims])
```

当 `corrected::Bool=false` 时计算的就是总体的值.

实际计算机并不知道你的数据是来自总体还是来自样本, 当样本中对象很多时, $n-1$ 与 n 差异不大. 所以关于符号使用主要是从概念上考虑的,这是我们作为使用者要考虑的问题

Summary

方差和标准差计算了总体或者样本中每个个体和均值差异的平均值. 度量的是集合中测量值的平均差异.

使用均值和方差(标准差) 就能度量一个某个属性的大部分信息.

其中要反复思考的是:个体之间一定存在差异, 样本采样中由于随机选择的个体会不同, 所以计算的均值和方差(标准差)会出现差异. 在统计学中要把这个微妙地方的细节彻底弄明白才行