样本均值的分布

ch07 样本均值的分布

sec7.1 样本和总体的关系

如果理解充分认识样本和总体的关系, 在进行推断统计的时候, 对流程的理解就很清晰.

从提供的信息角度看, 总体的信息完整, 样本的信息是不完整的信息.

如果把总体看成一本资料, 通过碎纸机以后, 变成纸条. 如果从这些纸条有些遗失了, 那么我们拼出的信息可能与原来信息有差别. 简答来说, 纸条的数量越多, 获得的信息就越完整. 缺失的那部分信息可以认为是误差.

Definition

抽样误差 : 样本统计量与对样总体参数间的差异, 或者误差总和

sec7.2 样本均值的分布

在一个总体中, 个体的的某属性值不同.当我们从总体中挑选出一些不同的个体组成样本,可以计算出某属性测量值的均值. 如果我们反复抽取个体组成不同的样本, 每个样本都会获得一个均值. 每一个样本获取的均值和总体的均值都有抽样误差.

那么问题接着就来了:

这些样本的均值有什么特性

我们直接定义结论, 然后来详细解释. 如果在统计实验中重复的抽样次数多, 样本容量很大. 那么样本的均值会形成正态分布.

对于样本均值有很多问题需要搞清楚:

> Question
>
> 1. 总体某个参数的分布是什么样的?
> 2. 总体的某属性参数的均值有分布图吗?
> 3. 从总体抽取的不同样本的均值有分布吗? 是什么分布
> 4. 样本均值分布和总体的分布之间的关系.

> Answer
>
> 1. 总体某参数测量值的分布有多种类型
> 2. 总体单个参数的均值没有分布, 因为这仅仅只是一个数字
> 3. 不同样本均值一般不同. 符合正态分布
> 4. 样本均值的分布和总体分布没有任何关系. 总体分布描述的是个体取值的图像, 样本均值反映的是多个不同样本的均值的图像.

在统计中有三个数量关系要分清楚. 给三者赋予符号来理解:

总体: P=1

总体中个体数量: T

样本数 : m

每样本中个体数 :n

总体只有一个,总体中个体数量为一个定值,或者无限

样本数指的是进行的抽样实验次数

每样本中的个体数是每次抽样从总体中抽出的个体数量

Example

example 1 :用上面的符号描述一下假定为 对一个人口为1000万的国家进行身高抽样统计

总体数 P=1 个

总体中个体数量为 T=1000 万人

执行 100 次抽样, m=100 次

每次抽样从 T=1000万人总体中抽出 n=1000 人

Notice

1. 抽样统计要做的是通过在个体数量庞大的总体中抽取少量个体来近似获得某个总体的参数值.
2. 上述抽样统计是想通过抽样获得总体**身高均值** 这一参数的近似值
3. 我们要做的并不是去了解总体的数据是什么分布.
4. 获取总体个分布是可以的, 通过普查的方式获取总体中每个人的身高也可以.
5. 进行 100 次抽样目的是尽量减少抽样统计中的误差.

普查的问题: 理论上说对一个总体的所有个体进行统计, 会获取完整的信息,不会有误差. 但实际面对一个数量庞大的总体, 在测量属性值的时候会出现测量错误. 信息实际也会发生偏差.

样本均值的一般特性:

1. 样本的均值总是出现在总体均值附近
2. m 次抽样获取的均值分布大致是正态分布,
3. n 值越大, 每次抽出的个体的均值越接近于总体均值

Example

example 2

对一个总体 $\{2.0, 4.0, 6.0, 8.0\}$ 的总体进行抽样统计的描述:进行 16 次抽样,每次抽出 2 个数字

```
["总体" => 1, "总体集合" => [2.0, 4.0, 6.0, 8.0], "抽样次数" => 16, "样本容量" => 2]
```

```
begin
    @variables P, T,m,n,μ,X,X̄
    P=1
    T=[2.0,4.0,6.0,8.0]
    m=16
    n=2
    ["总体"=>P,"总体集合"=>T,"抽样次数"=>m,"样本容量"=>n]
end
```

$$\mu = 5.0$$

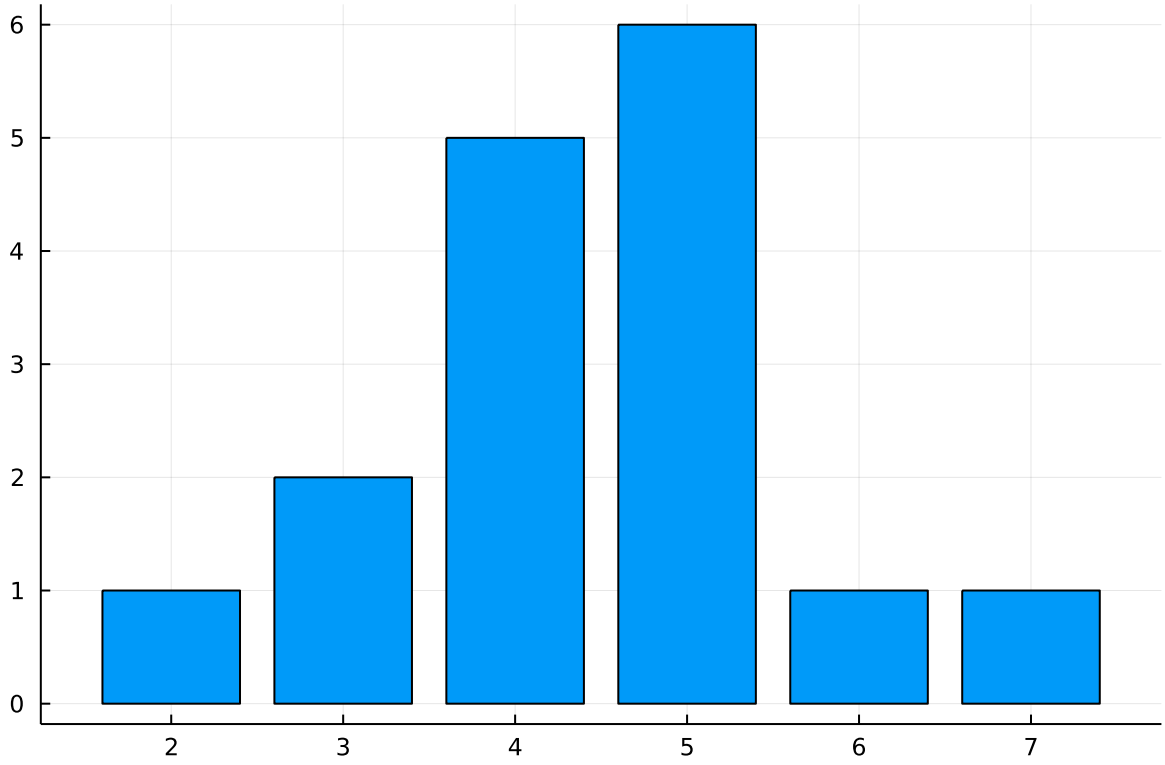
```
latexify(μ~mean(T))  # 总体均值作为参数
```

```
sample_data =
  [[4.0, [4.0, 4.0]], [2.0, [2.0, 2.0]], [4.0, [4.0, 4.0]], [3.0, [2.0, 4.0]], [4.0, [4.0, 4
```

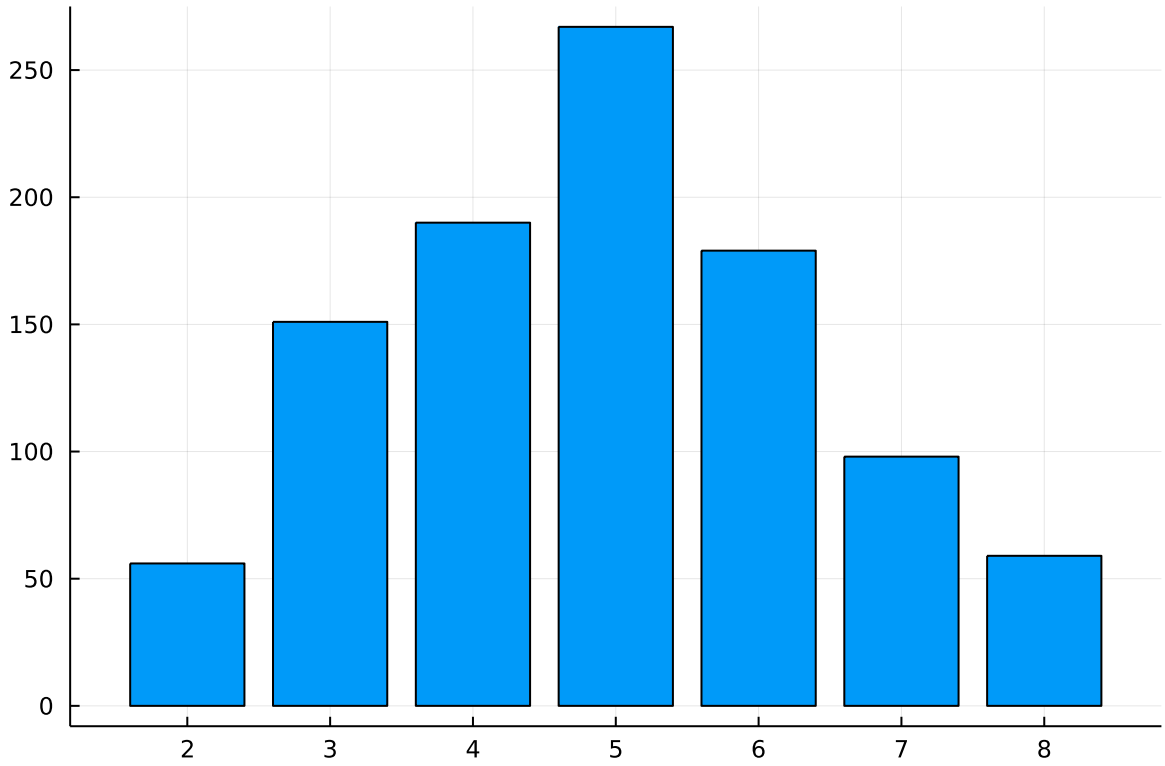
```
sample_data=[rand(T,2)|>X->[mean(X),X] for i = 1:m] # 流程:1, 进行 16 次抽样, 每次从总体抽出 2 个数字, 计算均值
```

	sample_no	number	sample_mean
1	"sample-1"	[4.0, 4.0]	4.0
2	"sample-2"	[2.0, 2.0]	2.0
3	"sample-3"	[4.0, 4.0]	4.0
4	"sample-4"	[2.0, 4.0]	3.0
5	"sample-5"	[4.0, 4.0]	4.0
6	"sample-6"	[6.0, 8.0]	7.0
7	"sample-7"	[6.0, 4.0]	5.0
8	"sample-8"	[6.0, 4.0]	5.0
9	"sample-9"	[2.0, 6.0]	4.0
10	"sample-10"	[6.0, 2.0]	4.0
	more		
16	"sample-16"	[6.0, 6.0]	6.0

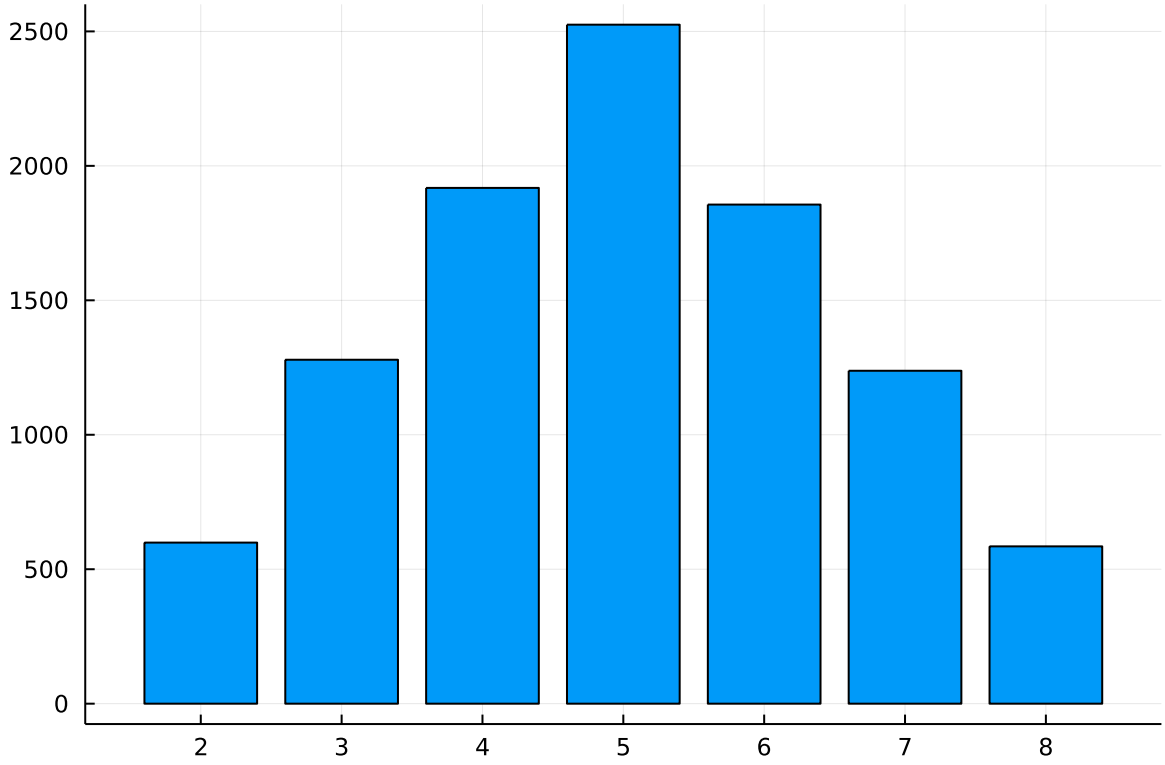
```
begin
    range=1:16
    sample_no=range.|>i->"sample-$(i)"
    mean_value=sample_data.|>x->x[1]
    num=sample_data.|>x->x[2]
    df=DataFrame(sample_no=sample_no, number=num, sample_mean=mean_value)
end
```




```
sample_plot(sample_data)  #16次抽样,如果抽样 1000 次
```




```
[rand(T,2)|>X->[mean(X),X] for i = 1:1_000]|>sample_plot  #1000 次实验
```




```
[rand(T,2)|>X->mean(X) for i = 1:10_000]|>sample_plot  # 10_000 实验
```

Info

当我们针对总体的抽样实验次数越多, 得到的均数分布越靠近与总体的均数. 这就是抽样统计的理论依据. 抽样实验越多, 获取的信息越多. 中心极限定律和大数定律就是对这里理论依据的总结归纳.

在统计学中要理解总体和抽样样本之间的关系. 中心极限定律和大数定律把两者联系起来.

```
md"""

!!! info

    当我们针对总体的抽样实验次数越多， 得到的均数分布越靠近与总体的均数。  这就是抽样统计的理论依据。
    抽样实验越多， 获取的信息越多。
    中心极限定律和大数定律就是对这里理论依据的总结归纳。

    在统计学中要理解总体和抽样样本之间的关系。 中心极限定律和大数定律把两者联系起来。
"""
```

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

```
begin
    function sample_plot(data)
    sample_data=data
    meanval=sample_data.|>x->x[1].|>Int
    min,max=extrema(meanval)
    feq=counts(meanval)
    bar(min:max,feq,label=false)
    end
end
```