中心极限定律

ch07 中心极限定律和大数定律

中心极限定理

均一分布总体的均数抽样分布

大数定理(law of large numbers)

```
1000
```

```
begin
    Random.seed!(0);
    n=1000
end
```

ch07 中心极限定律和大数定律

> **Warining**
>
> 在抽样统计中我们只是想获得总体的参数近似信息,不需要考虑总体的其他信息, 包括总体的分布如何. 例如从一个均一分布得到的样本均值符合正态分布. 见后续实例

理解这个任务目标很重要

中心极限定理

中心极限定理(central limit theorem) 说明, 只要每次抽样中抽取的个体很多. 例如每次抽样抽取 30 个个体进行测量($n = 30$), 重复多次抽样实验, 每次抽样的均数组成的样本均数集合符合正态分布.

注意表述:多次抽样获取的不同均值组成一个集合.每次抽样获取的均数是样本均数总体里的一个个体

因为样本(单位为:次)均数符合正态分布, 大部分进行的抽样实验获取的均值都会分布在这个正态分布的均值附近.

这为们进行少量的抽样实验就可以获取近似的总体均值信息提供依据.

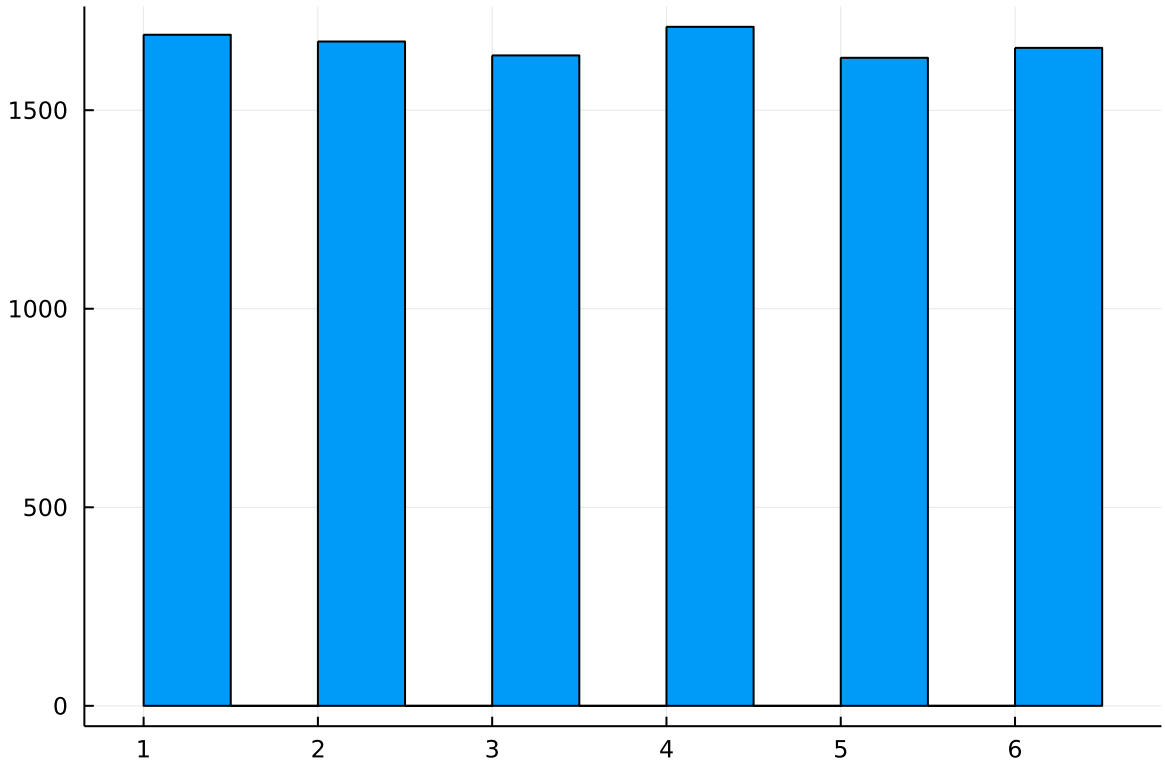
均一分布总体的均数抽样分布

最常见的均一分布是掷色子

公平的筛子,投掷的时候每一面的机会均等都为 $\frac{1}{6}$, 见下图

筛子投掷的总体实际是无限的, 由于各面出现的机会均等, 总体的均值为**3.5**, **3.5** 的含义并不是我们会投出一个 3.5 的花色值. 根据均值的定义**3.5** 度量了筛子投掷的总体点数的集中的趋势.

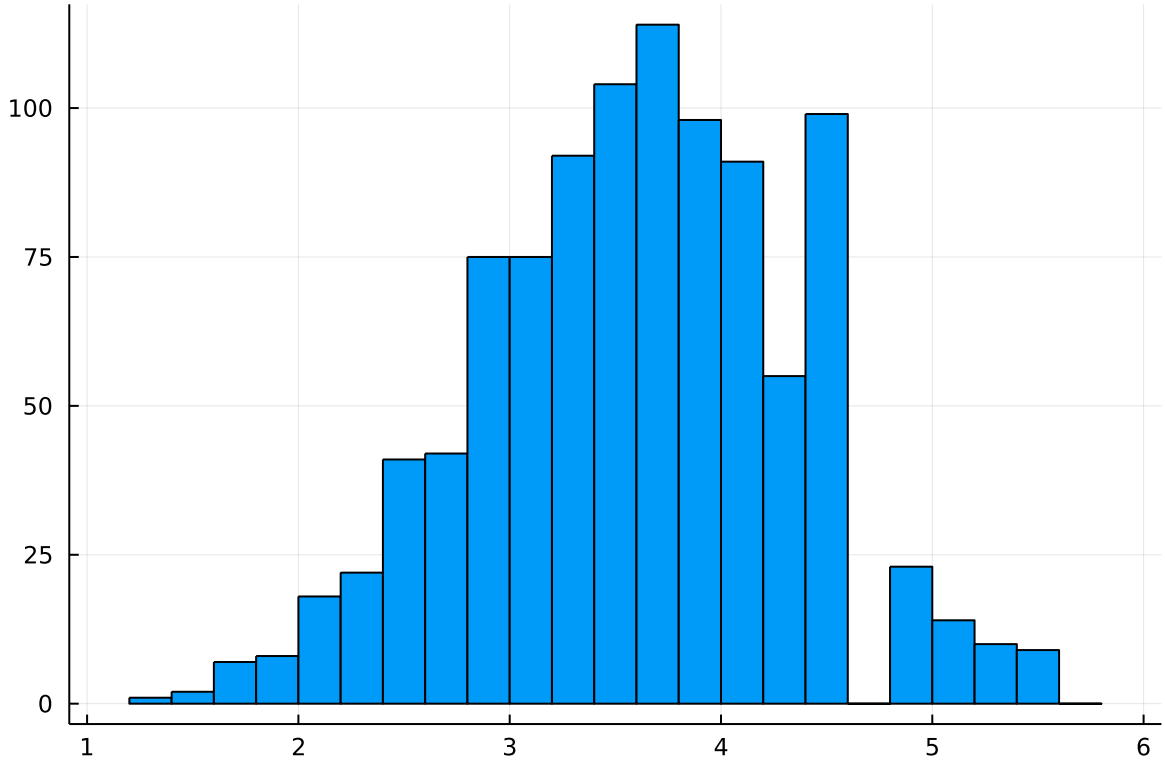
按照中心极限定理, 当掷色子次数增多以后, 样本的均值会趋近于总体的均值.



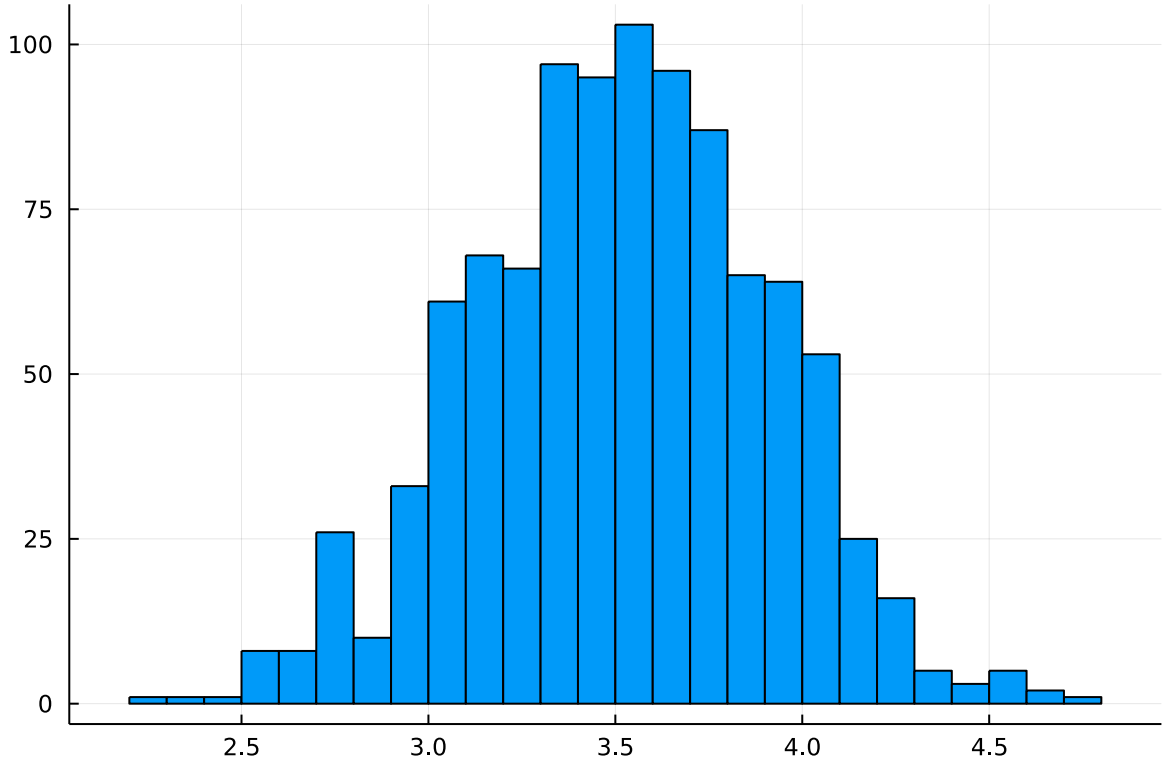

```
begin
    data=sample(1:6,10000)    # 模拟掷骰子 ， 各面出现的机会均等
    histogram(data,xticks=(0:6),label=false)
end
```

下面的代码都是样本均值统计的分布图. 在系列图中从上到下有一个趋势. 当采样次数都为($n = 1000$) 时,每次采样选取的个体越多, 最后得到的图形越接近于正态分布. 重复采样次数多显示图中的趋势不是偶然的因素, 是由固定的模式的. 可见采样的时候选择的个体数对于样本均数的分布有直观重要的影响.

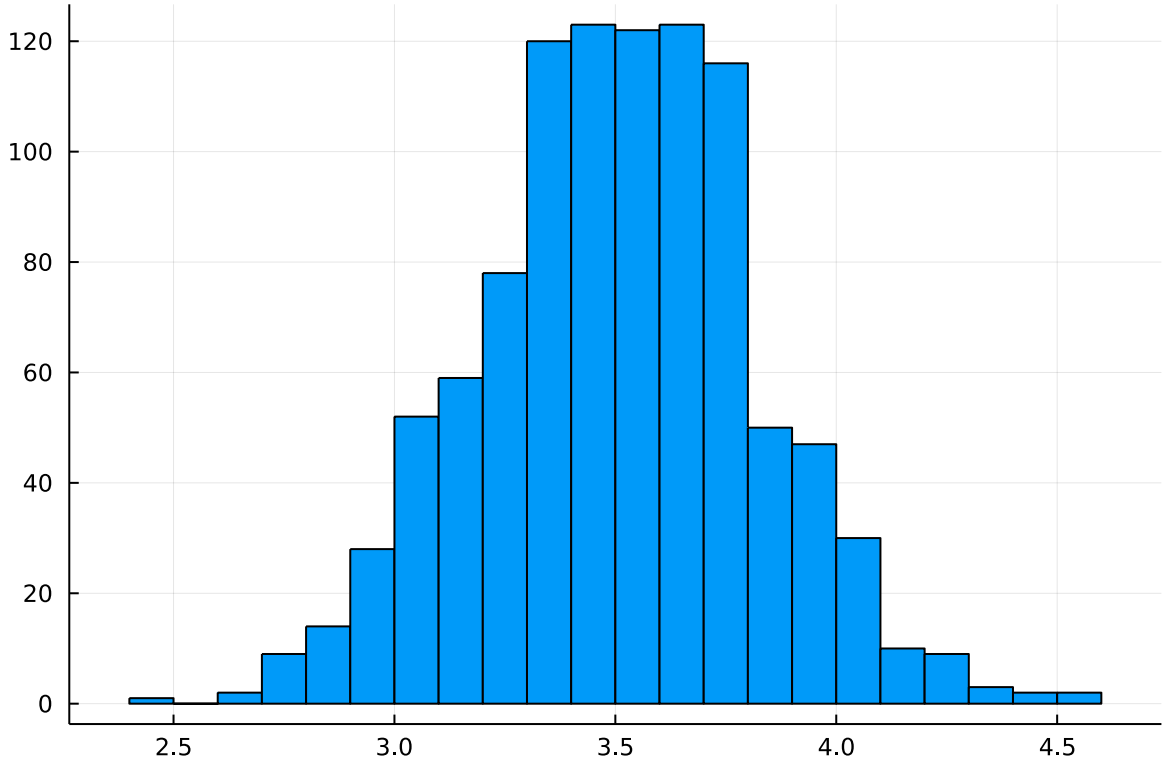
在样本均值分布的离散度量中引入的属性叫标准误.



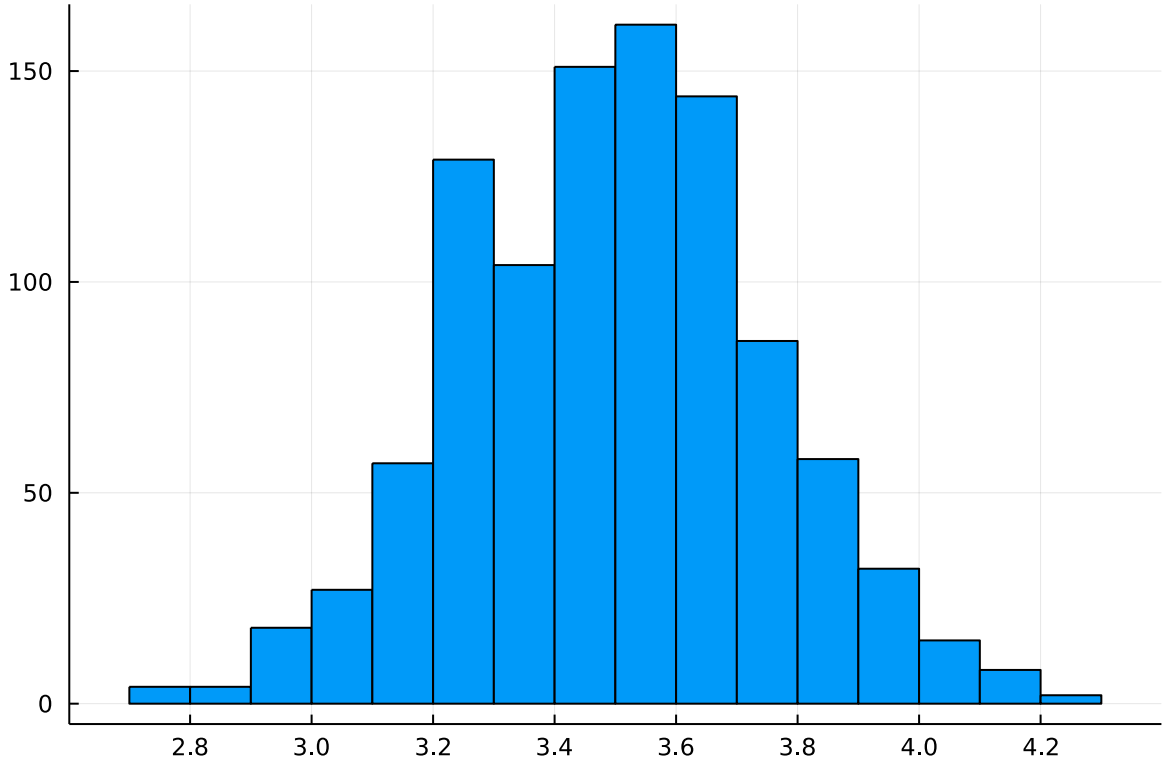

```
let
    sample5=[]

    for i in 1:n
        push!(sample5, mean(sample(data, 5,))) #sample 函数点击右下角Live docs 查看文档
    end

     histogram(sample5, label=false)
end
```



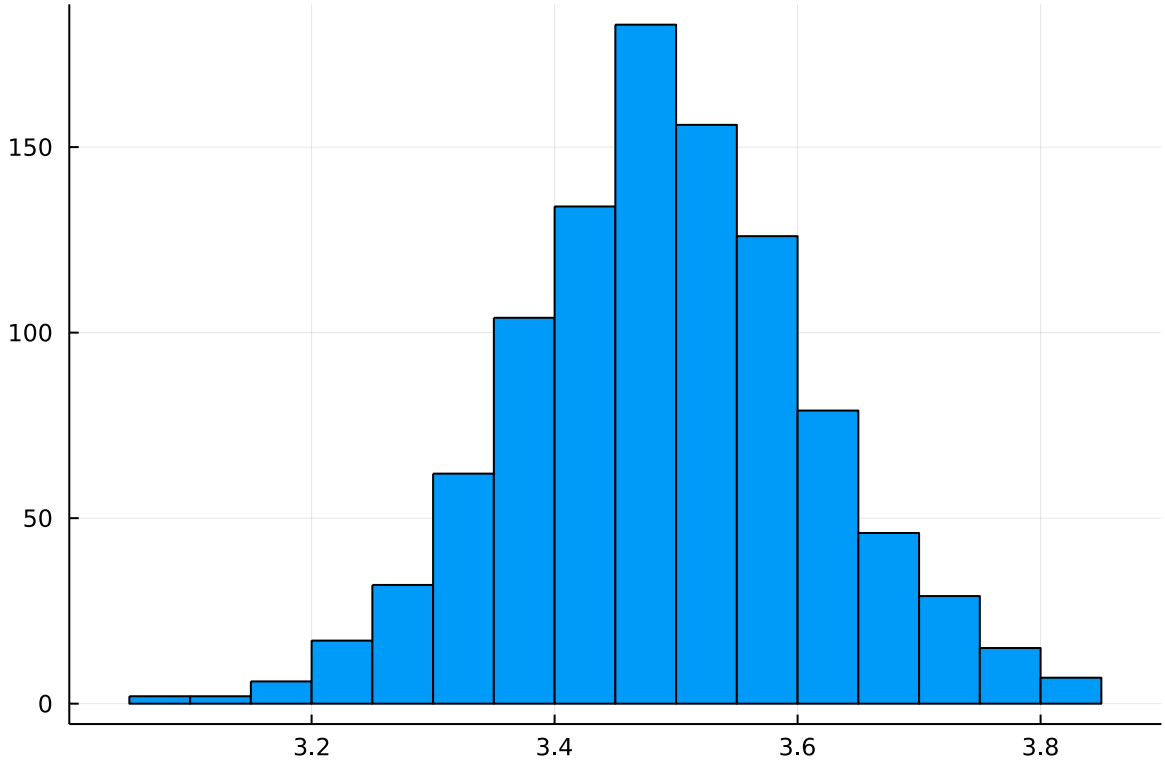

```
let
    sample20=[]

    for i in 1:n
        push!(sample20, mean(sample(data, 20,)))
    end

     histogram(sample20, label=false)
end
```



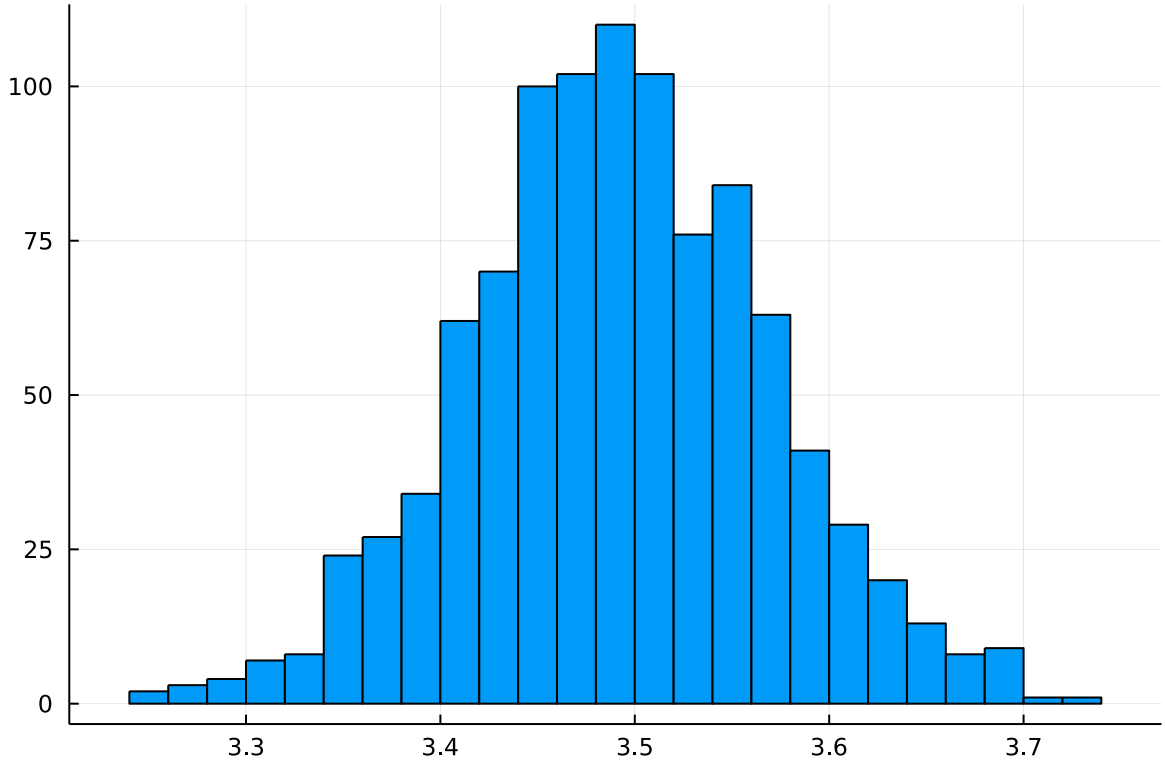

```
let
    sample30=[]

    for i in 1:n
        push!(sample30, mean(sample(data, 30,)))
    end

     histogram(sample30, label=false)
end
```



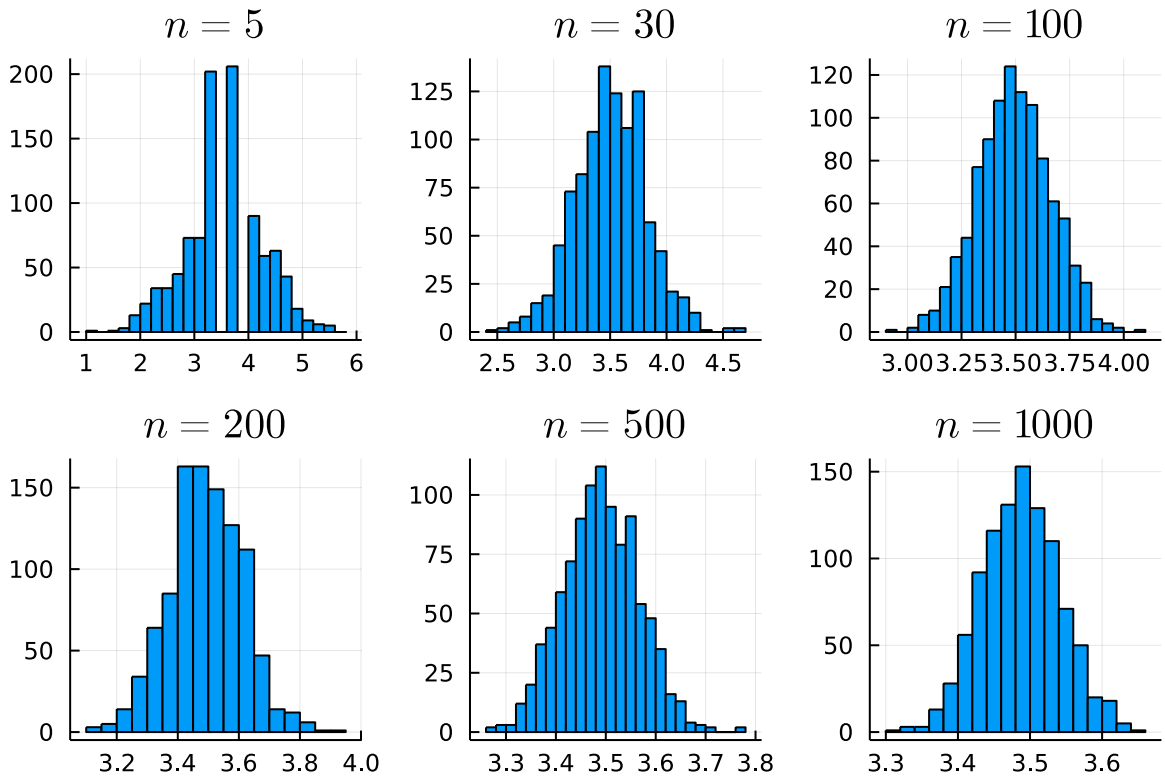

```
let
    sample50=[]

    for i in 1:n
        push!(sample50, mean(sample(data, 50,)))
    end

     histogram(sample50, label=false)
end
```




```
let
    sample200=[]
    freq=0

    for i in 1:n
        mean,std=sample(data, 200)|>mean_and_std
        push!(sample200,mean)

    end

     histogram(sample200, label=false)
end
```




```
let
    sample500=[]
    for i in 1:n
        mean,std=sample(data, 500)|>mean_and_std
        push!(sample500,mean)

    end

     histogram(sample500, label=false)
end
```

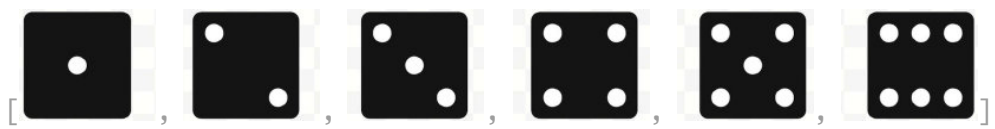



```
begin
    n_collection=[5,30,100,200,500,1000]
    title=hcat([L"n=%$n" for n in n_collection]...)
    plot_arr=n_collection.|>num->plot_dist(data,num,1000)
    plot(plot_arr...,layout = (2, 3), title=title,legend = false)
end
```

大数定理(law of large numbers)

上面的一组图说明就是大数定律(law of large numbers), 当采样采用过程中选取的对象数越多, 样本均值分布的的均值收敛于总体均数. 采样中选取的个体(对象, n) 度量了从总体中获取信息的水平.

举个简单的例子, 比如一个地方风景作为总体, 照片的像素点越多, 照片越清楚, 反映总体的细节信息就越多. 这就是大数定律.



```
begin
    url="https://tva1.sinaimg.cn/mw690/e6c9d24egy1h5aq35fwbej222k0a6q47.jpg"
    black_img=load(download(url))
    blackdice=black_dice=sprite(black_img,1,6)

end
```

```
sprite (generic function with 3 methods)
```

```
begin
    function sprite(img,row=2,col=2)
         arr=[]
          height,width=size(img)
          w,h=width÷col,height÷row
            for r in 0:row-1
                for c in 0: col-1
                    a= r==0 ? 1 : r*h
                    b= c==0 ? 1 : c*w
                    push!(arr,img[a:(r+1)*h,b:(c+1)*w])
                end
            end
       return arr
    end
end
```

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

```
begin
     get_mean(data,num,n)=[sample(data,num,)|>mean for i in 1:n]
     plot_dist(data,num,n)=get_mean(data,num,n)|>x->histogram(x,label=false)
end
```