样本均值的标准误

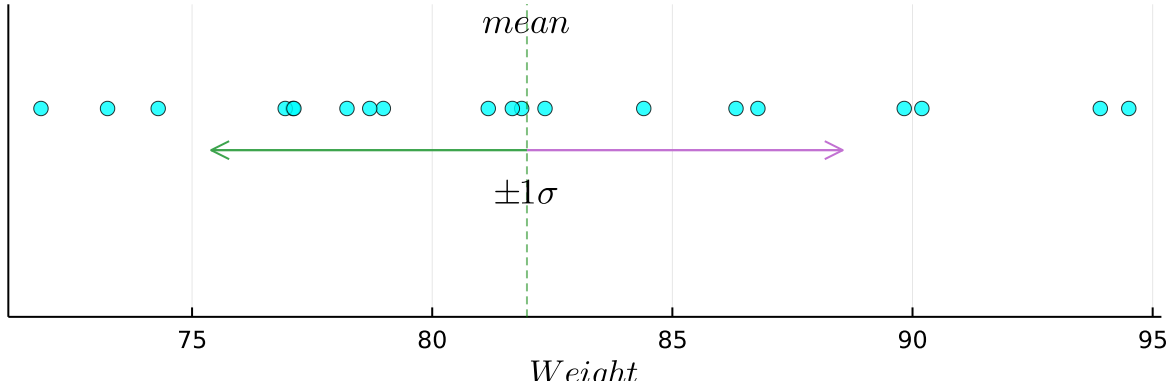

```
begin
    using Latexify    ,PlutoUI    ,RDatasets    ,DataFrames    ,Random    ,FileIO    ,HTTP
       ,Images    ,Plots    , StatsPlots    ,Statistics    ,StatsBase    ,LaTeXStrings
    ,Symbolics    ,CSV    ,Distributions
    TableOfContents(title="样本均值的标准误")
end
```

ch07 样本均值标准误

> **Info**
>
> 标准误是能否迈入假设检验大门的钥匙. 后续的假设检验, 方差检验方法都以标准误作为基础.
>
> 在中心极限定理和大数定律一节, 我们已经看到样本均数分布的一些特征.
>
> 抽样实验中从总体中抽取的个体数量对样本均数分布的性质影响巨大. 在面对一个总个体数庞大的总体, 样本信息是我们唯一可以获取的信息.
>
> 总的原则只有一个, 要尽可能多的从总体中抽取可用个体. 当然在多年的统计实验研究中, 已经发现了一些规律, 比如样本容量下限在$(n = 30)$, 当抽样中获取的个体数达到30 的时候, 样本均数分布基本符合正态分布. 也就是说随机抽取的样本均数不会离总体均数差太多.

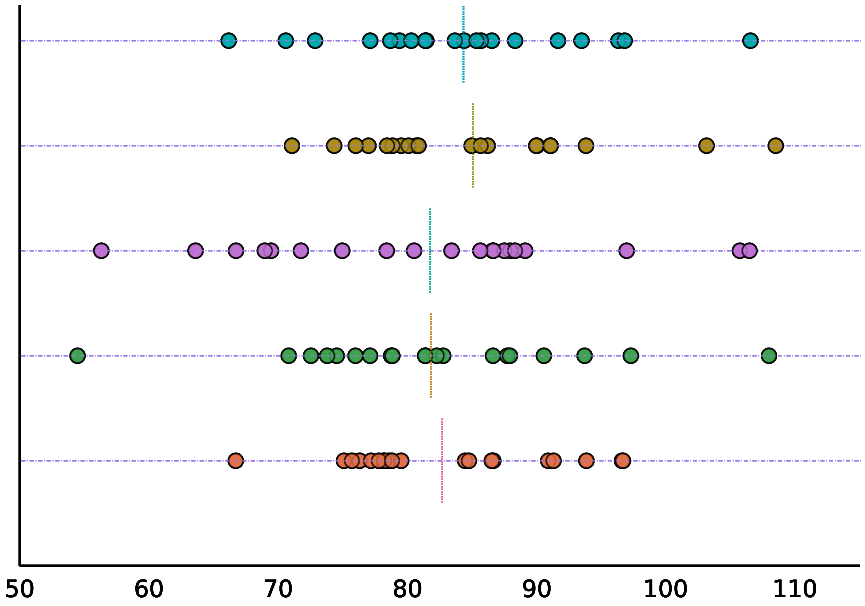
```
P = Distributions.Normal{Float64}(μ=83.0, σ=10.0)
```

```
P= Normal(83.0,10.0)  #定义正态分布对象，作为数据采样总体
```




```
begin
    y_val,offset=2, 0.2
    # sample_data=[58, 66, 71, 73, 74, 77, 78, 82, 84, 85, 88, 88, 88, 90, 90, 92,
    92, 94, 96, 98]
    sample_data=rand(P,20)
    sample_mean,sample_std=mean_and_std(sample_data)
    ann=[(sample_mean,y_val+offset,
text(L"mean", pointsize=12, halign=:center, valign=:center)),
     (sample_mean,1.8,
text(L"\pm  1 \sigma", pointsize=12, halign=:center, valign=:center))
    ]
    scatter(sample_data, [repeat([y_val],length(sample_data))] ,label=false,
    color=:aqua, alpha=0.8,yticks=false,ylims=(1.5,2.25),size=
    (600,200),xlabel=L"Weight")
    vline!([sample_mean],ls=:dash, color=:green, lw=1,alpha=0.5,label=false)
    plot!(ann=ann)
    plot!([sample_mean,sample_mean-sample_std],[1.9,1.9],label=false,arrow=true)
    plot!([sample_mean,sample_mean+sample_std],[1.9,1.9],label=false,arrow=true)

end
```



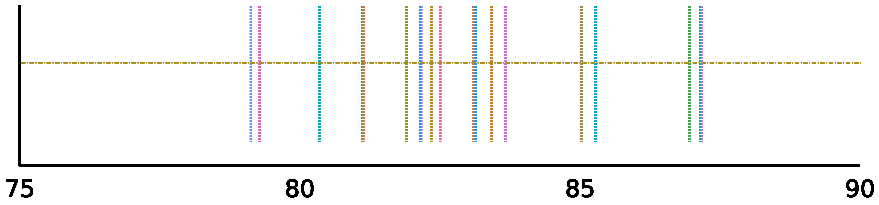

```
begin

    y_unit=0.2
    function plot_fun(data, idx)
        return scatter!(data[idx],[repeat([idx*y_unit],20)],label=false,alpha=0.8)
    end
    function plot_mean(data,idx)

        return plot!([data[idx], data[idx]],[(idx*y_unit)+0.08,
        (idx*y_unit)-0.08],label=false, lw=0.8,ls=:dot)
    end

    sample_arr=(1:1:5).|>x->rand(P,20)
    sample_mean_arr=sample_arr.|>mean
    plot(title=L"sample dist",yticks=false,size(1800,500),ylims=(0,1.6),xlims=
    (50,115))
    plotarr=[plot_fun(sample_arr,idx) for idx in 1:5]
    mean_segment_arr=[plot_mean(sample_mean_arr,idx) for idx in 1:5]
    plot!(plotarr... ,grid=false,tick_direction=:none)
    plot!(mean_segment_arr...)
    hline!([idx*y_unit for idx in 1:5],label=false,ls=:dashdot,lw=0.5)

end
```




```
begin
    sampe_no=20
    sample_arr2=(1:1:20).|>x->rand(P,sampe_no)
    sample_mean_arr2=sample_arr2.|>mean
    function plot_mean2(data,idx)
        return plot!([data[idx], data[idx]],[1+0.05,1-0.05],label=false,
        lw=1,ls=:dot)
    end
    plot()
    mean_segment_arr2=[plot_mean2(sample_mean_arr2,idx) for idx in 1:1:sampe_no]
    hline!([0.96],label=false,ls=:dashdot,lw=0.5)
    plot!(mean_segment_arr2..., size=(600,
    200),tick_direction=:none,grid=false,xlims=(75,90),yticks=false)
end
```

在上图中当我们把每次抽样的均值按照一维形式表示的时候,每个均值数字都像是一个新的集合的一个个体.

这个集合中,每一个样本均值点与集合的均值都有离差, 离差的平方和就是均值抽样分布的方差, 均值抽样分布的方差开方就得到均值抽样分布的标准差. 这个标准差就定义为标准误

均值标准误的计算公式

以总体均值计算;

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

以样本均值计算:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

其中:

σ – –总体的标准差

s – –从抽样样本中获取的标准差

n – –样本容量

从公式可以看出标准误由标准差和样本容量决定. 根据大数定律, 样本容量越大,样本均值越趋近于总体均值

总体的数据离散程度大, 样本的离散程度越大. 标准误总和度量了样本采样时的两个方面的影响因素.

标准误表示我们抽样获取的样本均值在均值抽样分布中的位置和排名的范围.

标准误实例

在推断性统计中, 假设检验是重要的方法, 假设检验的依据就是标准误.

先来看看如何在 Julia 中计算标准误. 使用 `StatsBase.jl` 的 `sem` 方法

样本容量相同的情况

> **Example**
>
> example 1 计算下面样本数据的标准误

```
data1 =  (3, 4, 4, 5, 7, 8, 12, 14, 14, 15, 17, 19, 22, 24, 24, 24, 25, 28, 28, 29)
```

```
data1=(3, 4, 4, 5, 7, 8, 12, 14, 14, 15, 17, 19, 22, 24, 24, 24, 25, 28, 28, 29)
```

```
mean_error1 = 2.001446845080881
```

```
mean_error1=sem(data1)  # 可以提供总体均值，经过普查的数据集可以提供总体均值，例如人口普查
```

当个的标准误数据没有意义, 数据进行比较才能反应出意义, 例如数据集2

```
data2 =  (3, 4, 4, 5, 7, 8, 12, 14, 14, 15, 17, 19, 22, 24, 24, 24, 25, 28, 28, 150)
```

```
data2=(3, 4, 4, 5, 7, 8, 12, 14, 14, 15, 17, 19, 22, 24, 24, 24, 25, 28, 28, 150)
```

```
mean_error2 = 6.978265128993476
```

```
mean_error2=sem(data2)
```

data2 样本中最后一个测量值变为了 150, 标准误一下从2 跳变到7. 在样本容量一样的情况下, 标准误捕捉到了两个数据集合里数据的离散程度差异.

样本容量不同的情况

```
md"""

### 样本容量不同的情况
"""
```

```
((1, 2, 3, 4, 5), (1, 2, 3, 4, 5, 1, 2, 3, 4, 5))
```

```
data3,data4=(1, 2, 3, 4, 5),(1, 2, 3, 4, 5, 1, 2, 3, 4, 5)
```

```
[5, 10]
```

```
length.([data3,data4])
```

```
[0.707107, 0.471405]
```

```
sem.([data3,data4])
```

当样本容量增大的时候, 标准误会下降. 针对同一个总体, 采样的时候使用的个体越多, 捕捉的信息越接近于总体参数. 可以看到大数定律在抽样统计中无时无刻不在.

当总体信息未知的时候, 抽取的个体越多, 统计数值越接近于总体参数信息. 这就是推断统计.

Question

如果我们要通过抽样的方法来看看IQ 的得分情况,

-.一种随机抽取了10个人,测量IQ,获取平均IQ,

-.一种随机抽取1000个人, 测量IQ,获取平均IQ.

两种情况, 你更愿意相信那种规模的数据?

样本标准误和样本标准差的区别

Info

1. 样本标准差度量的是样本数据的平均离散程度.
2. 样本标准误度量的是样本均值代表样本均值分布均值的能力