统计分布介绍

统计分布介绍

1. 伯努利分布
2. 二项式分布

   二项式分布, 成功的概率并非只有 (p=0.5)
3. 泊松分布
4. 均一分布
5. 多项式分布
6. 正态分布

   正态分布的性质

   样本抽样数据如何形成抽样分布

```
begin
    using Latexify    ,PlutoUI    ,RDatasets    ,DataFrames    ,Random    ,FileIO    ,HTTP
        ,StatsPlots    ,Statistics    ,StatsBase    ,LaTeXStrings    ,Symbolics    ,CSV
       ,Distributions

    using Plots    :plot,bar
    TableOfContents(title="统计分布介绍")

end
```

```
TaskLocalRNG()
```

```
Random.seed!(123)
```

统计分布介绍

```
md"""
#  统计分布介绍
"""
```

1. 伯努利分布

伯努利分布(Bernoulli distribution) 实际是描述只有两种状态的随机变量的方法.

最常见的场景有足球比赛开始时抛硬币, 胜者可以挑选场地. 对于一个公平的硬币, 正反面的机会均等, 一些随机因素, 包括抛硬币时用的力, 高度等等造成结果的不确定性.只进行一次实验.

伯努利分布(Bernoulli distribution) 实际是二项式分布(Binomial distribution) 的一个特例, 当二项式分布实验次数为1时就是伯努利分布.

```
md"""
## 1. 伯努利分布

伯努利分布(Bernoulli distribution)  实际是描述只有两种状态的随机变量的方法.


最常见的场景有足球比赛开始时抛硬币， 胜者可以挑选场地.  对于一个公平的硬币， 正反面的机会均等， 一些随机因素， 包括抛硬币时用的力， 高度等等造成结果的不确定性.只进行一次实验.

![](https://tva1.sinaimg.cn/orj360/e6c9d24egy1h5gugep58lj20go0m80tw.jpg)

伯努利分布(Bernoulli distribution) 实际是二项式分布(Binomial distribution) 的一个特例, 当二项式分布实验次数为$1$时就是伯努利分布.
"""
```

```
Distributions.Bernoulli{Float64}(p=0.5)
```

```
begin
    success_rate=0.5
    Bernoulli(success_rate)
end
```

2. 二项式分布

二项式分布(Binomial distribution)

二项式分布与伯努利分布不同点在于, 二项式分布 分布进行多次实验

二项式分布的规范为:

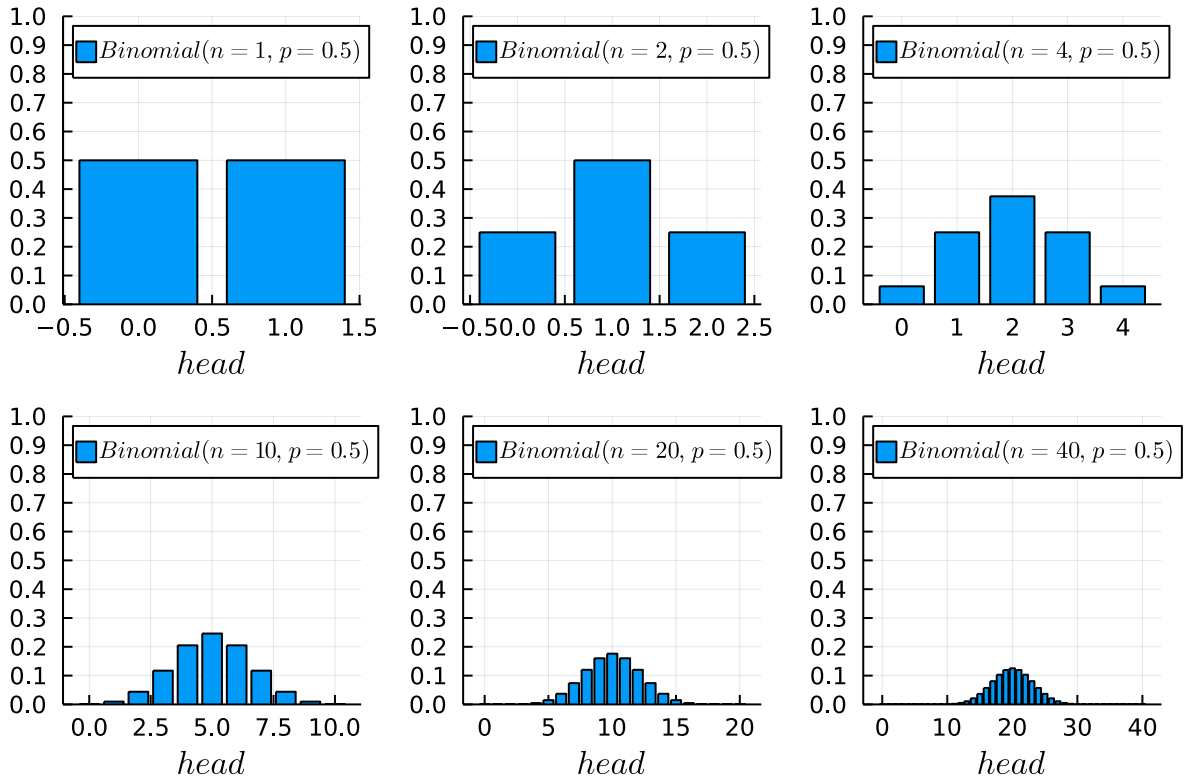
Def

- 试验重复进行n 次
- 每次试验只用两种可能, 一种定义为成功, 一种定义为失败
- 成功的概率表示为p
- 每次试验都是独立的, 互不影响

```
md"""
## 2. 二项式分布

二项式分布(Binomial distribution)

二项式分布与伯努利分布不同点在于，  二项式分布  分布进行多次实验

二项式分布的规范为:

!!! def

    - 试验重复进行$n$ 次
    - 每次试验只用两种可能，一种定义为成功，一种定义为失败
    - 成功的概率表示为$p$
    - 每次试验都是独立的，互不影响


"""
```

```
Distributions.Binomial{Float64}(n=1000, p=0.5)
```

```
begin
    n,p=1000,0.5
    d=Binomial(n,p)
end
```

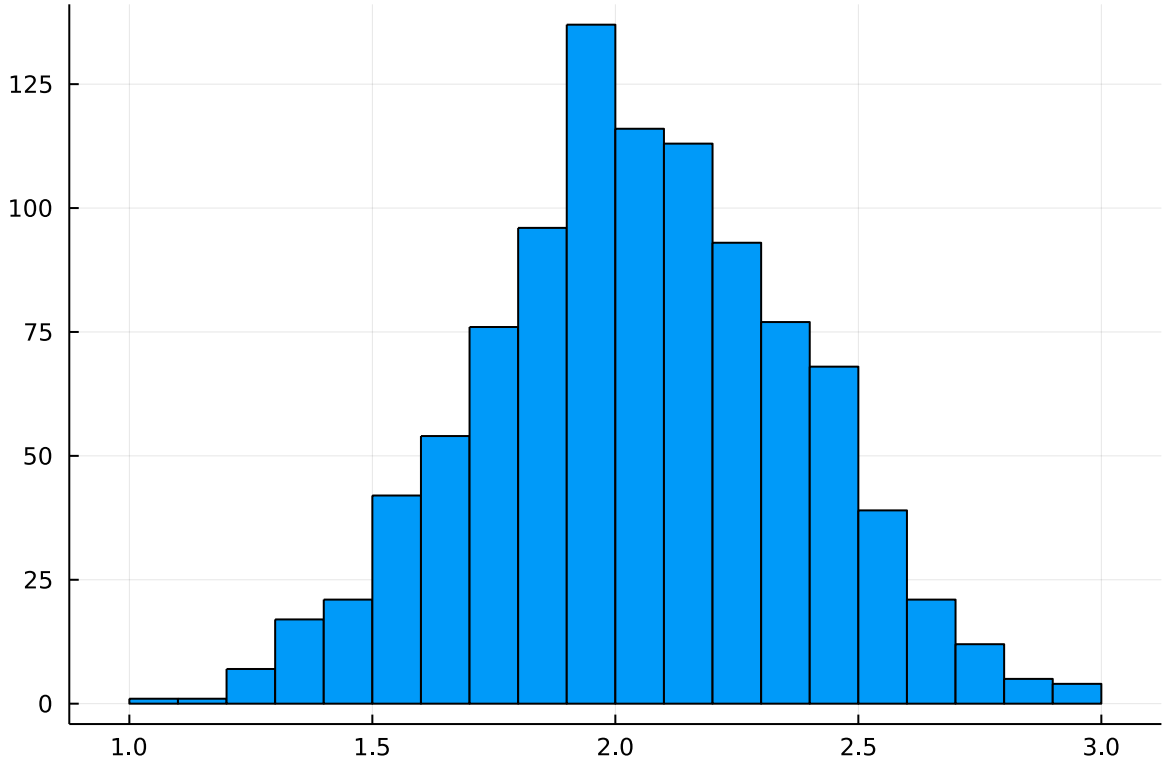



```
plot(binomial_pdf.((1,2,4,10,20,40))...)
```

上图为n 取不同值的二项式分布的图. 以$Binnomial(4, p = 0.5)$ 为例, 图中从左至右表示投出 $0, 1, 2, 3, 4, 5$ 次正面的几率. 因为成功的几率为$p = 0.5$, 所以期望最高为抛出两次正面. 因为每次抛硬币是独立, 并且随机的, 一次正面也没有的概率也存在.

要注意这是理论分布的, 分布的均值为 $\frac{0+1+2+3+4+5}{5} = 2$, 意思是如果我们把抛$4$次硬币作为一次抽样, 那么执行多次抽样形成的抽样分布是一个正态分布, 均值围绕在2 周围. 这就是我们只执行一次实验, 抛四次硬币, 很大可能会是两次正面, 两次反面, 当然具体到每次实际操作, 不会完全一样, 有抽样误差.

```
md"""
上图为n 取不同值的二项式分布的图. 以$Binnomial(4,p=0.5)$ 为例, 图中从左至右表示投出 $0,1,2,3,4,5$ 次正面的几率.  因为成功的几率为$p=0.5$, 所以期望最高为抛出两次正面. 因为每次抛硬币是独立, 并且随机的, 一次正面也没有的概率也存在.

要注意这是理论分布的, 分布的均值为$\frac{0+1+2+3+4+5}{5}=2$, 意思是如果我们把抛$4$次硬币作为一次抽样, 那么执行多次抽样形成的抽样分布是一个正态分布, 均值围绕在$2$ 周围. 这就是我们只执行一次实验, 抛四次硬币, 很大可能会是两次正面, 两次反面,
当然具体到每次实际操作, 不会完全一样, 有抽样误差.
"""
```




```
let
    data=[mean(rand(Binomial(4,0.5),10)) for i in 1:1000]
    histogram(data,label=false)
end
```

二项式分布, 成功的概率并非只有($p = 0.5$)

例如蓝球比赛投篮就符合二项式分布. 例如投 10 个球, 成功率 0.3

```
md"""
### 二项式分布, 成功的概率并非只有$(p=0.5)$

例如蓝球比赛投篮就符合二项式分布. 例如投 10 个球, 成功率 0.3
"""
```

```
throw = Distributions.Binomial{Float64}(n=10, p=0.3)
```

```
throw=Binomial(10, 0.3)  # 投篮命中率为 0.3
```

```
["1次正面的概率" => 0.00976563, "2次正面的概率" => 0.0439453, "3次正面的概率" => 0.117188, "4次
```

```
begin
    coin=Binomial(10,0.5)
    res=["$(i)次正面的概率"=>pdf(coin, i)  for i in 1:10]
end
```

```
0.9990234375000017
```

```
    res.|>(x->x[2])|>sum    #概率密度累积为 1
```

3. 泊松分布

泊松分布的属性是:

- 试验成功的数量可以计算
- 一段特定时间段内事件成功的均数是已知的
- 每个结果是独立的
- 事件成功的概率和时间间隔成比例

例如某医院一段时间内出生的婴儿数量平均为每小时 10 人, 这就符合泊松分布的要求, 知道一段时间内成功的均数, 母亲分娩都是独立时间,之间互不影响. 随着间隔时间延长, 出生率会增加

```
md"""
## 3. 泊松分布

泊松分布的属性是:

 - 试验成功的数量可以计算
 - 一段特定时间段内事件成功的均数是已知的
 - 每个结果是独立的
 - 事件成功的概率和时间间隔成比例


例如某医院一段时间内出生的婴儿数量平均为每小时 10 人, 这就符合泊松分布的要求, 知道一段时间内成功的均数, 母亲分娩都是独立时间,之间互不影响. 随着间隔时间延长, 出生率会增加
"""
```

```
Distributions.Poisson{Float64}(λ=4.0)
```

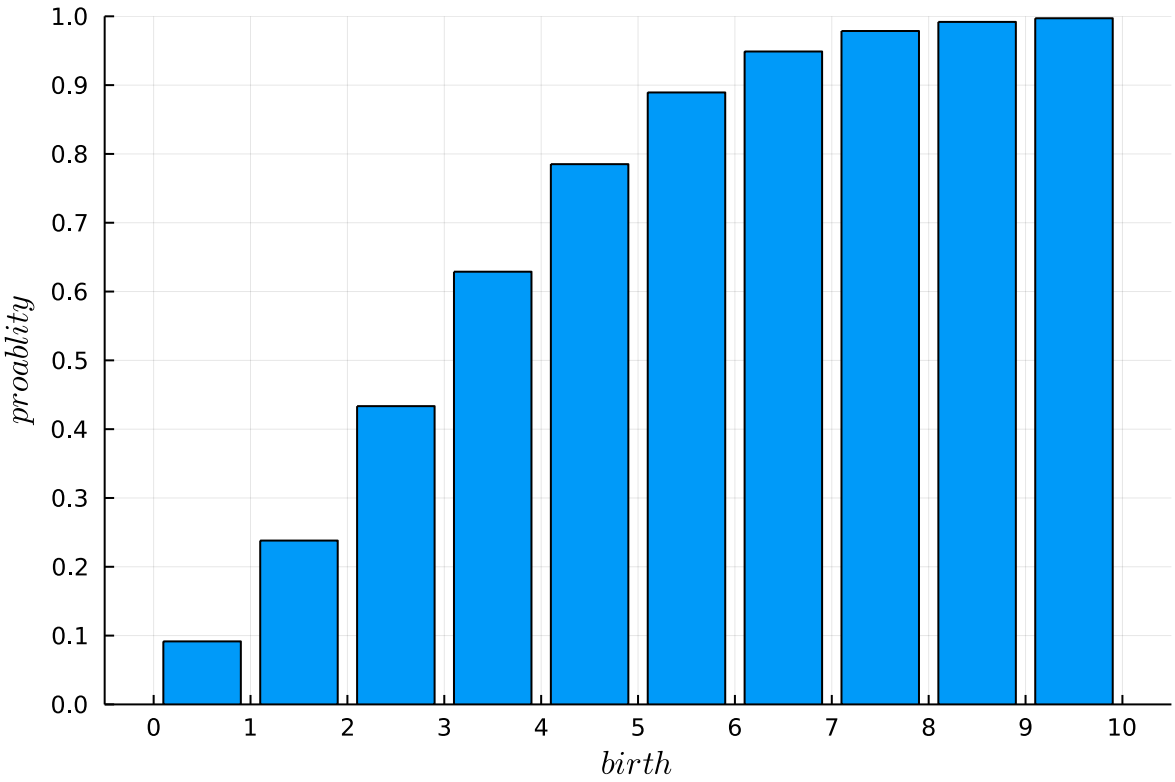
```
begin
    λ=4  #医院统计每小时出生婴儿 4 人
    birth_possion=Poisson(λ)  # 建立泊松分布
end
```

```
0.19536681481316454
```

```
pdf(birth_possion,4)    #每小时出生 4 人的概率
```

```
0.9971602338794863
```

```
cdf(birth_possion,10)    #出生数 0-10 人的累积概率接近1
```




```
possion_cdf(birth_possion,10)  #累加的概率
```

4.均一分布

在均一分布中, 每个点发生的机会均等. 掷色子就是均一分布, 每个点数出现的机会一样.

实例 公交车到站间隔时间为20分钟,8分钟以内公交车会会到站的概率为多少?

因为间隔时间为 $0-20$, 最幸运的是立刻就有车到站, 最不走运的时,到站是车刚好开走. 也就是公交车在$(0\ 20)$分钟时间间隔内随时会来, 每分钟到站的机会均等. 所以这是均一分布. 计算8分钟以内公交到站是累积$(0-8)$分钟的概率

```
md"""
## 4.均一分布

在均一分布中， 每个点发生的机会均等．  掷色子就是均一分布， 每个点数出现的机会一样．

实例　公交车到站间隔时间为20分钟,8分钟以内公交车会会到站的概率为多少?

因为间隔时间为 $0-20$， 最幸运的是立刻就有车到站， 最不走运的时,到站是车刚好开走． 也就是公交车在$(0~20)$分钟时间间隔内随时会来， 每分钟到站的机会均等． 所以这是均一分布． 计算8分钟以内公交到站是累积$(0-8)$分钟的概率
"""
```

```
bustop = Distributions.Uniform{Float64}(a=0.0, b=20.0)
```

```
bustop=Uniform(0, 20)  #建立均值分布
```

```
bus_at_prob =
  [0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05,
```

```
bus_at_prob=(0:20).|>t->pdf(bustop,t)  # 20分钟内,每分钟车到站机会均等
```

公交到站这个例子看似在谈时间, 其实和时间没有关系. 等一分钟和等 20 分钟的机会相等, 假设我们去公交站 1000 次, 两种等待时间的理论次数都是 50 次(1000X0.05) . 试验与频率有关与时间无关

我们换一种观点, 假设有一个 20 面的色子, 累积掷出 1-8的概率为多少?

```
less_than_8 = 0.4
```

```
less_than_8=cdf(bustop,8)  #等待 8 分钟车回来的概率是 0.4
```

5. 多项式分布

多项式分布是对二项式分布的泛化,

成功的概率是一个概率向量, 标明了在一次实验中各项成功的概率.

例如一个罐子中共有 10 个球, 黄色球 6 个, 红色球 2 个, 粉色球 2 个. 当我们随机从中抽取一个球的时候, 概率为:$[0.6, 0.2, 0.2]$,

如果我们采用放回时抽样(因为总体中个体少,如果总体个数庞大可以不放回), 抽取 4 个球, 计算每种组合的概率, 就会组成一个 以 概率向量为成功率, 实验次数为 4的多项式抽样分布

从这个抽样分布中连续抽取四个黄球的概率为多少呢?

```
md"""
## 5. 多项式分布
 多项式分布是对二项式分布的泛化,

 成功的概率是一个概率向量, 标明了在一次实验中各项成功的概率.

 例如一个罐子中共有 10 个球, 黄色球 6 个, 红色球 2 个, 粉色球 2 个. 当我们随机从中抽取一个球的时候, 概率为:$[0.6,0.2,0.2]$,

如果我们采用放回时抽样(因为总体中个体少,如果总体个数庞大可以不放回), 抽取 4 个球, 计算每种组合的概率, 就会组成一个 以 概率向量为成功率, 实验次数为 4的多项式抽样分布

从这个抽样分布中连续抽取四个黄球的概率为多少呢?
"""
```

```
begin
    struct Pv
         Ye:: Float64
         Re:: Float64
         Pi:: Float64
    end
 pv=Pv(0.6,0.2,0.2)    # 黄, 红,粉球的概率向量
 trial=Pv(4,0,0)       #  从布袋中取出4 个球,都是黄球的试验
 marble=Multinomial(4, [pv.Ye,pv.Re,pv.Pi]) #以概率向量执行4次试验的多项式分布
 allyellow=pdf(marble,[trial.Ye,trial.Re,trial.Pi])
 println("概率向量为 ",[0.6,0.2,0.4])
 println("连续抽取4次黄球的概率为 ", allyellow)
end
```

```
概率向量为 [0.6, 0.2, 0.4]
连续抽取4次黄球的概率为 0.1296
```

6. 正态分布

在统计学中使用最为广泛的是正态分布(normal distribution), 不仅仅是因为很多现实的总体的分布接近于正态分布. 更为重要的是无论总体分布如何, 针对总体抽样获取的均数分布均呈现正态分布. 这是正态分布对推断统计最重要的作用.

每一次抽样获取的均数都位于正态分布中, 并且大多数情况下抽取的均数都围绕在均数附近(抽取样本的均数分布在抽样分布均数的附近). 这就是抽样推断统计的理论依据.

对于统计中的正态分布, 要理解两个方面的内容, 一是正态分布自身的特性, 第二是抽样的均数是怎么形成正态分布的. 一旦明确了抽样均数分布的特性, 后续的一些推断统计难度就会下降, 因为很多的统计都是以抽样均数分布为基础展开讨论的.

> Notice
>
> 样本抽样均数分布是为以总体均数为中心, 并不是以总体分布为中心. 在大多数的统计教程中, 样本数据是模拟从某个分布获取的, 但是实际上我们不需要考虑总体是什么分布. 我们唯一能得到的可能只是样本数据.

从逻辑上上, 为什么推断统计要声明零假设呢? 因为总体未知, 既然总体未知, 总体分布情况怎么会知道呢?

有反反复复的理解这其中的关系

```
md"""
## 6. 正态分布

在统计学中使用最为广泛的是正态分布(normal distribution)，不仅仅是因为很多现实的总体的分布接近于正态分布. 更为重要的是无论总体分布如何，针对总体抽样获取的均数分布均呈现正态分布. 这是正态分布对推断统计最重要的作用.

每一次抽样获取的均数都位于正态分布中，并且大多数情况下抽取的均数都围绕在均数附近(抽取样本的均数分布在抽样分布均数的附近).  这就是抽样推断统计的理论依据.

对于统计中的正态分布，要理解两个方面的内容，一是正态分布自身的特性，第二是抽样的均数是怎么形成正态分布的.  一旦明确了抽样均数分布的特性，后续的一些推断统计难度就会下降，因为很多的统计都是以抽样均数分布为基础展开讨论的.


!!! notice
    样本抽样均数分布是为以总体均数为中心，并不是以总体分布为中心. 在大多数的统计教程中，样本数据是模拟从某个分布获取的，但是实际上我们不需要考虑总体是什么分布.  我们唯一能得到的可能只是样本数据.

从逻辑上上，为什么推断统计要声明零假设呢? 因为总体未知，既然总体未知，总体分布情况怎么会知道呢?


有反反复复的理解这其中的关系
"""
```

正态分布的性质

- 形成钟形曲线
- 左右对称
- 均数和中位数一样, 位于中心
- 68% 的曲线下面积集中在± 1 个标准差区间内
- 95% 的曲线下面积集中在± 2 个标准差区间内
- 99.7% 的曲线下面积集中在± 3 个标准差区间内

在正态分布中, 位置和百分比比实际的分数更能反映出总体的信息. 个体分数转化为标准分以后就可以使用标准正态分布来处理位置和概率问题. 如下图

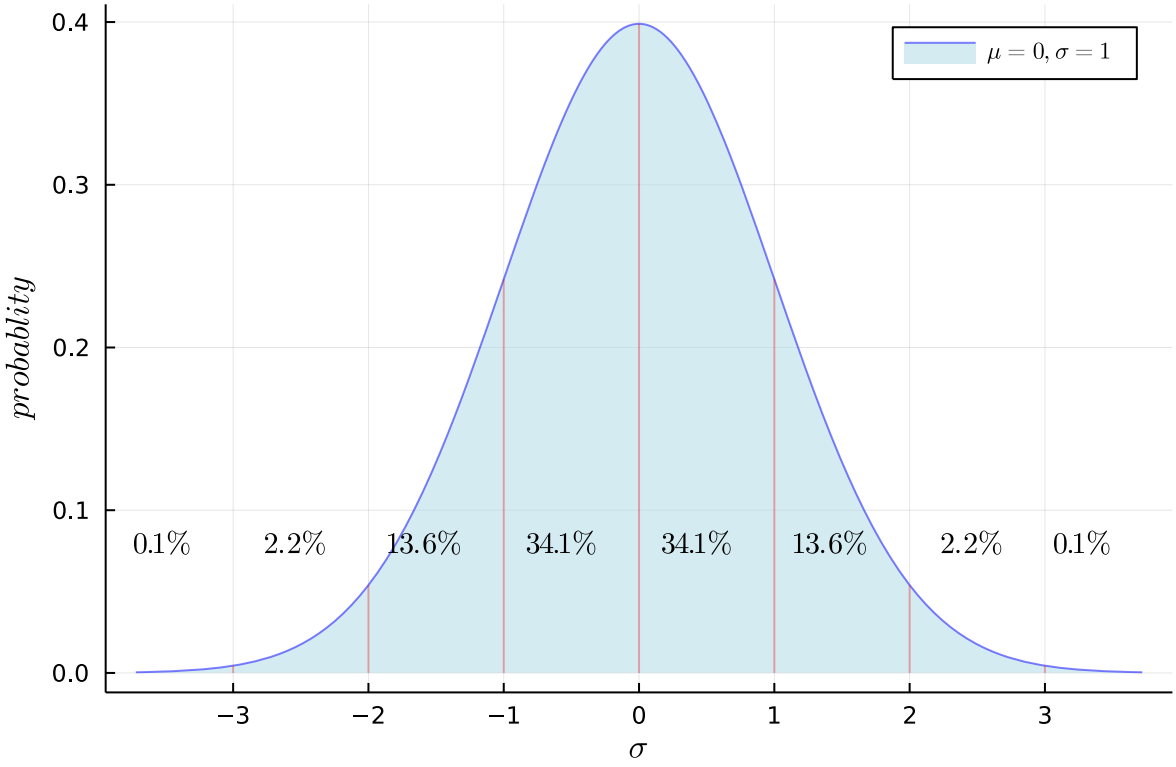
- 在$\pm 1\sigma$ 区间内集中了68% 的数据
- 在$\pm 2\sigma$ 区间内集中了95% 的数据
- 在$\pm 3\sigma$ 区间内集中了99.7% 的数据

落在三个标准差之外的数据可以认为是变异性非常大的数据 可以认为离散程度过大不能代表总体

这就是 $68-95-99.7$ 规则. 为统计推断提供了一个定量的依据.

但是这个定量依据只是一个手段. 最终要能使用好这个规则,需要理以样本均数如何形成正态分布

```
md"""
在正态分布中, 位置和百分比比实际的分数更能反映出总体的信息. 个体分数转化为标准分以后就可以使用标准正态分布来处理位置和概率问题.  如下图

- 在$\pm 1 \sigma$ 区间内集中了$68 \%$ 的数据
- 在$\pm 2 \sigma$ 区间内集中了$95 \%$ 的数据
- 在$\pm 3 \sigma$ 区间内集中了$99.7 \%$ 的数据

落在三个标准差之外的数据可以认为是变异性非常大的数据 可以认为离散程度过大不能代表总体

这就是 $68-95-99.7$ 规则. 为统计推断提供了一个定量的依据.

但是这个定量依据只是一个手段. 最终要能使用好这个规则,需要理以样本均数如何形成正态分布
"""
```




```
let
    μ,σ=0,1
    d=Normal(μ,σ)
    range=-3σ:1:3σ
    round1(x)=round(x,digits=1)
    res=range.|>x->cdf(d,x)*100|>round1
    diff= [n==0 ? res[1] : res[n+1]-res[n] for n in 0:length(res)-1]

    make_ann(str,x,y=0.08)=(x-0.3,y,
    text(round1(str)|>x->L"%$(x) \%", pointsize=10, halign=:right, valign=:center))
    ann= [make_ann(diff[n],range[n]) for n in 1:length(range)]
    push!(ann, make_ann(0.1,3.8σ))

    plot(d,label=L"\mu=%$(μ),\sigma=%$(σ)",alpha=0.5,color=:blue,xticks=
(-3:1:3),fill=(0,:lightblue))
   plot!(repeat(range', 2), [zeros(1, length(range)); (range.|>σ->pdf(d,σ))'], label
 = "", color = :red, alpha = 0.3,lw=1)

    snormal=plot!(ann=ann,xlabel=L"\sigma",ylabel=L"probablity")
end
```

样本抽样数据如何形成抽样分布

当我们进行抽样试验时不管样本量多少,都是总体的一个缩影, 后续有偏差,但是大体反映了总体的一些特征. 如果反复的抽样, 每次抽样都是总体的一个特征反映, 但是样本并没有提供总体的所有信息, 会有偏差,但是大体上会形成对总体的一个特征集中的描述, 这个描述就以正态分布的形式出现.

以样本均数统计量为例, 反复抽样多次就可以看到大多数的样本都捕获了总体的信息. 大部分的抽样试验属于抽样分布, 从$68-95-99.7$ 规则可以知道,每次抽样都有极大的概率位于总体均数的两个标准差以内.

计算时你可以不用考虑这个问题, 但是理解样本抽样分布形成机理以后, 统计学的学习就彻底明了

```
md"""
### 样本抽样数据如何形成抽样分布

当我们进行抽样试验时不管样本量多少,都是总体的一个缩影， 后续有偏差,但是大体反映了总体的一些特征. 如果反复的抽样， 每次抽样都是总体的一个特征反映， 但是样本并没有提供总体的所有信息， 会有偏差,但是大体上会形成对总体的一个特征集中的描述， 这个描述就以正态分布的形式出现.

以样本均数统计量为例， 反复抽样多次就可以看到大多数的样本都捕获了总体的信息. 大部分的抽样试验属于抽样分布， 从$68-95-99.7$ 规则可以知道,每次抽样都有极大的概率位于总体均数的两个标准差以内.

计算时你可以不用考虑这个问题， 但是理解样本抽样分布形成机理以后， 统计学的学习就彻底明了
"""

```

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

```
function binomial_pdf(n)
    bar(0:n, pdf.(Binomial(n), 0:n),
         ylim = (0, 1), yticks = 0:0.1:1,
        label = L"Binomial(n=%$n, p=0.5)", legend = :topleft,xlabel=L"head")
 end
```

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

```
function possion_cdf(dist,n)
    bar(0:n,(1:n).|>n->cdf(dist,n), ylim = (0, 1), yticks =
    0:0.1:1,xticks=0:10,ylabel=L"proablity",xlabel=L"birth",label=false)
end
```