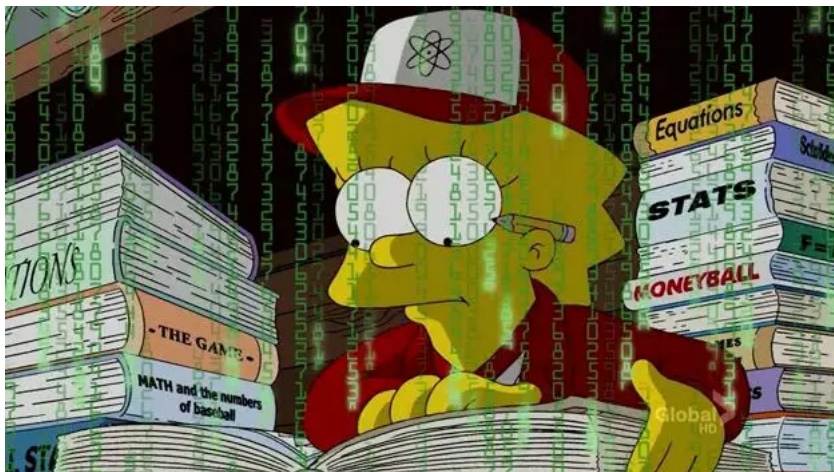


手把手 | Python代码和贝叶斯理论告诉你，谁是最好的棒球选手

原创：文摘菌 大数据文摘 4月18日



大数据文摘作品

编译：李雷、张馨月、王梦泽、小鱼

除了文中所附的代码块，你也可以在文末找到整个程序在Jupyter Notebook上的链接。

在数据科学或统计学领域的众多话题当中，我觉得既有趣但又难理解的一个就是贝叶斯分析。在一个课程中，我有机会学习了贝叶斯统计分析，但我还需要对它做一些回顾和强化。

从个人观点出发，我就是想更好地理解贝叶斯理论，以及如何将它应用于现实生活中。

本文主要是受到了RasmusBååth在Youtube上的系列节目“贝叶斯数据分析入门”的启发。RasmusBååth非常善于让你直观地理解贝叶斯分析，不是抛各种复杂的公式给你，而是引导你一步步去思考。

RasmusBååth的视频链接：

<https://www.youtube.com/user/rasmusab/feed>

本文将通过贝叶斯理论来分析棒球选手的命中率，手把手教你如何利用贝叶斯理论进行分析。说实话，我不是一个体育迷，也很少看体育比赛。

那么为什么选择棒球？

“不论你了解与否，但棒球的魅力就在于精确度。没有其他运动像棒球这样完全依赖于运动数据的连续性，统计性和有序性。棒球球迷比注册会计师还要关注数字。”

——体育记者Jim Murray

有人说棒球可能是世界上数据记录得最详细的运动。历史上已经累积了近百年来的棒球统计数据。

然而，仅仅收集统计数据并不会让棒球在统计方面变得有趣，也许更重要的是这项运动本身的特点。

举例来说，在完成一次打数(At Bats，是棒球运动中的一个成绩计算名词，指击球手完成打击的次数)过程中，谁在外野打球对于击球手是否可以击中本垒打影响甚微。

在其他体育运动，尤其是足球和篮球运动中，球员统计数据的意义可能会因球场内其他地方发生的重要事件而被淡化。而棒球这项运动中，统计数据在比较球员表现上发挥了重要作用。

棒球统计数据包含很多指标，有些指标的定义很直观，有些则比较复杂。我选择观察的测量指标是打击率(Batting Average, AVG)。

在棒球中，打击率由安打(Hits，安打是棒球运动中的一个名词)次数除以打数来定义，通常精确到小数点后三位。

有人质疑打击率的作用，但正如C. Trent Rosecrans所说，“尽管如此，打击率相较于其他统计数据而言确实是有历史和背景意义。我们都知道AVG为0.300打者的水平怎么样，我们也知道AVG为0.200打者有多糟，以及AVG为0.400打者有多棒。”

在美国职业棒球大联盟(MLB)中，春季训练是在常规赛开始之前的一系列练习和表演赛。

我会尝试解决以下两个问题：

- 如何解读2018年春季训练中的打击率
- 怎么比较两名球员的打击率

在进入代码内容之前，我会简要介绍一下Rasmus Bååth在他的视频中所讲的内容。

首先，我们需要三样东西来完成贝叶斯分析。

1.数据

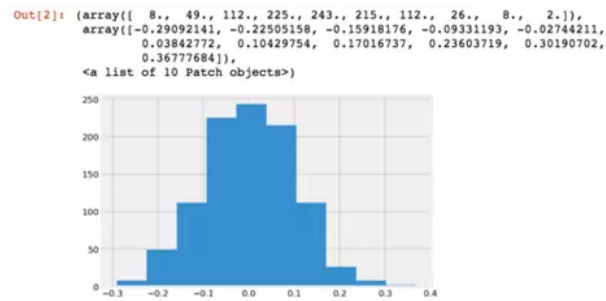
2.生成模型

3.先验概率

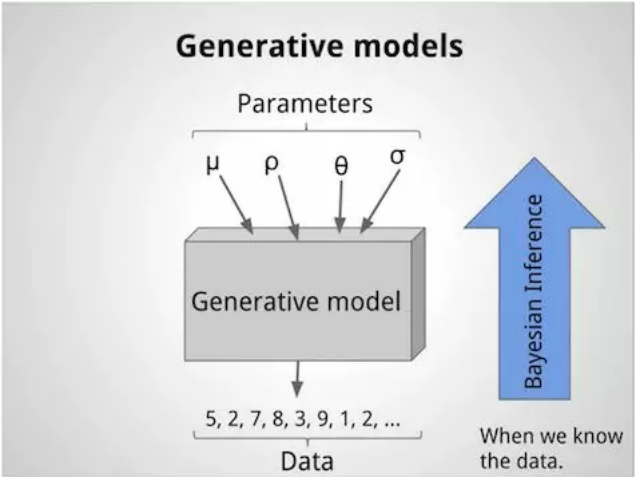
就我的例子而言，数据就是我们所观察到的2018年春季训练的打击率。

生成模型就是当给定参数作为输入时生成数据的模型。这些输入参数用于生成一个概率分布。例如，如果知道平均值和标准差，则可以通过运行以下代码轻松生成所选数据集的正态分布数据。稍后我们会看到其他类型的分布在贝叶斯分析中的运用。

```
import matplotlib.pyplot as plt
import numpy as np
mu, sigma = 0, 0.1 # mean and standard deviation
s = np.random.normal(mu, sigma, 1000)
plt.hist(s)
```



就贝叶斯分析而言，我们会逆向生成模型并尝试用观测数据推断参数。



最后，先验概率是指模型在处理数据之前就已有的信息。比如，事件是否等概率？是否有一些先前的数据可以利用？是否可以做出有依据的推测？

首先我将定义一个从Fox Sports抓取球员数据的函数，然后抓取球员的春季训练或常规赛季的击球统计数据。

Fox Sports链接：

<https://www.foxsports.com/mlb/stats>

The screenshot shows the Fox Sports website interface. At the top, there's a navigation bar with various sports categories like NFL, MLB, NASCAR, Soccer, NCAA FB, NBA, UFC, NCAA BK, and NHL. Below this, a search bar and a 'Watch Shows More' link are visible. The main content area features a player profile for Kevin Plawecki, a Catcher for the New York Mets. His stats for the 2017 season are displayed: HR 3, RBI 13, AVG .260. Below the profile, there's a 'STATS' tab selected, showing a table of batting statistics for the Spring Training season. The table has columns for Year, Team, G, PA, AB, R, H, 2B, 3B, HR, RBI, CS, BB, K, BB%, K%, and OPS. Data is provided for the years 2014, 2015, 2016, and 2017.

Year	Team	G	PA	AB	R	H	2B	3B	HR	RBI	CS	BB	K	BB%	K%	OPS
2014	NYM	5	5	4	2	2	1	0	0	2	0	0	0	.500	.667	1.167
2015	NYM	11	22	20	3	5	2	0	1	5	0	0	2	.250	.318	.568
2016	NYM	19	34	30	5	11	3	0	1	6	0	0	2	.344	.314	.658
2017	NYM	14	30	25	4	7	1	0	1	7	0	0	1	.240	.280	.520

```
import pandas as pd
import seaborn as sns
import requests
from bs4 import BeautifulSoup
plt.style.use('fivethirtyeight')
%matplotlib inline
%config InlineBackend.figure_format = 'retina'
def batting_stats(url, season):
    r = requests.get(url)
    soup = BeautifulSoup(r.text, 'lxml')
    table = soup.find_all("table", {"class": "wisbb_standardTable tablesorter"})[0]
    table_head = soup.find_all("thead", {"class": "wisbb_tableHeader"})[0]
    if season == 'spring':
        row_height = len(table.find_all('tr'))[-1]
    else:
        row_height = len(table.find_all('tr'))[-2]
    result_df = pd.DataFrame(columns=[row.text.strip() for row in table_head.find_all('th')], index = range(0, row_height))

    row_marker = 0
    for row in table.find_all('tr')[1:-1]:
        column_marker = 0
        columns = row.find_all('td')
        for column in columns:
            result_df.iat[row_marker, column_marker] = column.text.strip()
            column_marker += 1
        row_marker += 1
    return result_df
```

接着，我们选择一个感兴趣的球员，并对其统计数据进行分析。

2018		Spring Training		APPLY		RESET		Mets		Game Log	
SPRING ADVANCED											
BATTING											
		G	AB	R	H	2B	3B	HR	RBI	SB	CS
1	Smith, Dominic (B)	1	2	1	2	1	0	0	0	0	0
2	Cecchini, Gavin (C)	5	10	7	3	4	1	0	2	5	1
3	Plawecki, Kevin (C)	5	9	7	0	4	1	0	0	4	0
4	Goussone, Luis (B)	5	11	8	4	4	1	0	1	5	0
5	Nirino, Brandon (P)	5	15	12	4	5	1	1	1	4	0
6	Bryce, Jay (P)	3	5	5	0	2	1	0	0	1	0
7	Calbreath, Adriel (SS)	4	10	10	2	4	1	0	0	1	0

纽约大都会队春季训练的统计页面

如果按照球员的打击率 (AVG) 排列，你可以看到第一名是Dominic Smith (DS)，而Gavin Cecchini (GC) 则排第二。那他们是优秀球员吗？我不知道。但如果仅看AVG，DS以1.000的AVG 值位于榜首。

在谷歌上搜了一下，我发现“近年来，全联盟的平均击打率通常在0.260左右”。如果是这样，那么DS和GC的AVG似乎太高了。通过进一步观察两位选手的打数 (AB) 和安打 (H)，显然DS只有1个AB而GC有7个。并且在查看其他选手的AB后发现，2018年最高的AB为13，而2017年纽约大都会队的最高AB为60。

场景一

假设我对球员们过去的表现一无所知，2018年春季训练是唯一的数据来源，因此我不并知道AVG的取值范围。那么，我应该如何解读2018年春季训练的统计数据？

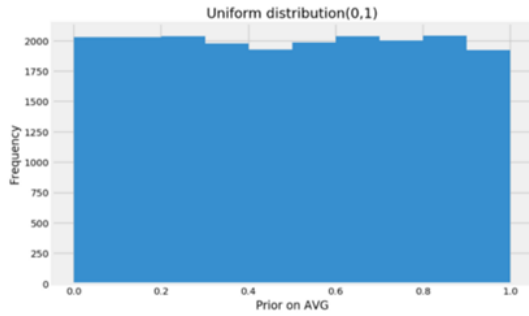
首先我们来抓取DS的春季训练数据。

```
ds_url_st = "https://www.foxsports.com/mlb/dominic-smith-player-stats?seasonType=3"
dominic_smith_spring = batting_stats(ds_url_st, 'spring')
dominic_smith_spring.iloc[-1]
```

```
Out[5]: BATTING      2018
Team      NYM
G          1
PA         2
AB         1
R          2
H          1
2B         0
3B         0
HR         0
RBI        0
SB         0
CS         0
BB         0
SO         0
AVG        1.000
OBP        1.000
SLG        1.000
OPS        2.000
Name: 4, dtype: object
```

```
n_draw = 20000
prior_ni = pd.Series(np.random.uniform(0, 1, size = n_draw))
plt.figure(figsize=(8,5))
plt.hist(prior_ni)
plt.title('Uniform distribution(0,1)')
plt.xlabel('Prior on AVG')
plt.ylabel('Frequency')
```

```
Out[6]: <matplotlib.text.Text at 0x1190cef50>
```



先验概率代表了我们在得到具体数据之前对某事物的普遍看法。在上述分布中，所有概率几乎相同（由于是随机生成，所以存在轻微差异）。

因此，这意味着我对球员一无所知，甚至无法对AVG做任何合理的猜测。我假设AVG是0.000和AVG是1.000的概率相同，或者等于AVG值为0和1之间任何数值的概率。

现在我们观察到的数据表明当有1个AB和1个H时，AVG是1.000，这可以用二项分布来表示。具有二项式分布的随机变量X表示在n次独立的是/非试验序列中成功的次数，其中每次试验成功的概率是p。

在我们的例子里，AVG是成功的概率，AB是试验次数，H是成功次数。

先记住这些术语，然后我们来定义逆向生成模型。

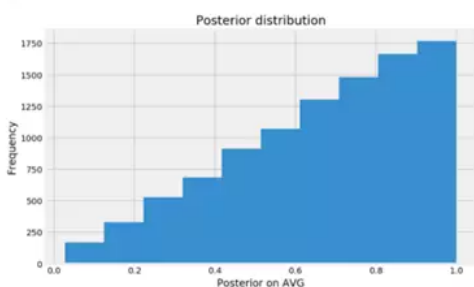
我们将从定义的均匀分布中随机选取一个概率值，并将此概率用作生成模型的参数。假设我们随机挑选的概率值为0.230，这意味着在二项分布中成功的概率为23%。

试验次数为1（DS有1个AB），如果生成模型的结果与我们观察到的结果相匹配（DS有1个H），那么概率值为0.230保持不变。如果我们重复这个过程并进行过滤，最终将得到一个概率分布，由它所得到的结果与我们观察到的结果相同。

这就是后验概率。

```
def posterior(n_try, k_success, prior):
    hit = list()
    for p in prior:
        hit.append(np.random.binomial(n_try, p))
    posterior = prior[list(map(lambda x: x == k_success, hit))]
    plt.figure(figsize=(8,5))
    plt.hist(posterior)
    plt.title('Posterior distribution')
    plt.xlabel('Posterior on AVG')
    plt.ylabel('Frequency')
    print('Number of draws left: %d, Posterior mean: %.3f, Posterior median: %.3f, Posterior 95% quantile interval: %.3f-%.3f' %
          (len(posterior), posterior.mean(), posterior.median(), posterior.quantile(.025), posterior.quantile(.975)))
ds_n_trials = int(dominic_smith_spring[['AB', 'H']].iloc[-1][0])
ds_k_success = int(dominic_smith_spring[['AB', 'H']].iloc[-1][1])
posterior(ds_n_trials, ds_k_success, prior_ni)
```

```
Number of draws left: 9917, Posterior mean: 0.665, Posterior median: 0.705, Posterior 95% quantile interval: 0.154-0.986
```



后验概率分布中95%的分位数区间称为可信区间，这与频率统计中的置信区间略有不同。还有另一种可以使用的可信区间，我后面讲到Pymc3时会提到。

贝叶斯统计中的可信区间和频率统计的置信区间的主要区别是二者的释义不同。**贝叶斯概率反映了人的主观信念**。根据这种理论，我们可以认为真实参数处于可信区间内的概率是可测量的。这种说法很吸引人，因为它使我们能够直接用概率对参数进行描述。

许多人认为这个概念是理解概率区间的一种更自然的方式，也很容易解释。置信区间使你能判断某区间是否包含真实的参数。

如果我们收集一个新样本，计算置信区间，并多次重复这个过程，那么我们计算出的95%的置信区间将包含真实的AVG值。

- 可信区间：根据观察数据，AVG的真实值落在可信区间内的概率为95%。
- 置信区间：当我们用这类数据计算置信区间时，有95%的置信区间会包含AVG的真实值。

注意两者的区别，可信区间是在给定固定边界情况下对参数值的概率描述，置信区间是在给定固定参数值情况下的边界概率。

在现实生活中，我们想知道的是真实的参数而不是边界，因此，贝叶斯可信区间是更合适的选择。在这种情况下，我们只对球员的真实AVG感兴趣。

有了上面的后验分布，我有95%的把握断定DS真正的AVG将在0.155到0.987之间。但这个范围太大了。换句话说，在没有先验知识并且在只观察了一次试验的情况下，我不太确定DS的真实AVG是多少。

场景二

对于第二个场景，我们假设知道上一年的春季训练的统计数据。

```
dominic_smith_spring.iloc[-2:]
```

Out[8]:

	BATTING	Team	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	AVG	OBP	SLG	OPS
3	2017	NYM	20	42	36	4	6	1	0	0	4	0	0	5	13	.167	.286	.194	.480
4	2018	NYM	1	2	1	2	1	0	0	0	0	0	0	0	0	1.000	1.000	1.000	2.000

现在有了2017年春季训练的统计数据，我们的先验假设应该反映这方面的信息。注意到2017年春季训练时DS的AVG是0.167，因此2017年的统计数据不呈均匀分布。

Beta分布是一个连续概率分布，它有两个参数，alpha和beta。Beta分布最常见的用途之一是对一个实验的成功概率的不确定性进行建模。

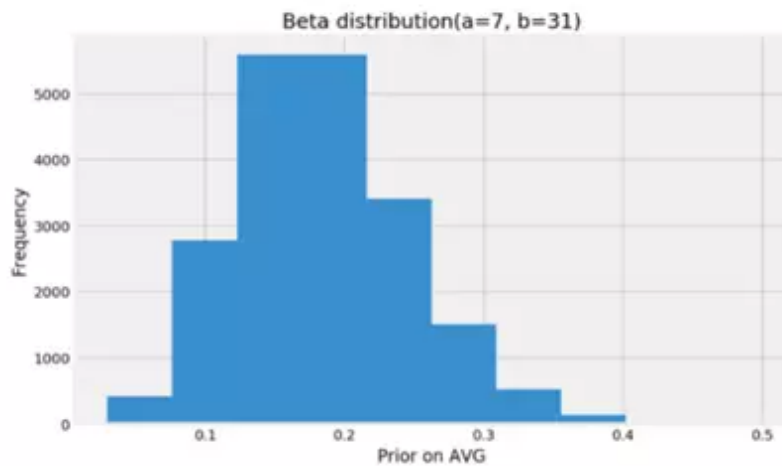
特别地，在已知n次试验中观察到k次成功的条件下，X的条件分布是一个 $\alpha=k+1$ 、 $\beta=n-k+1$ 的Beta分布。

Beta分布相关内容：

<https://www.statlect.com/probability-distributions/beta-distribution>

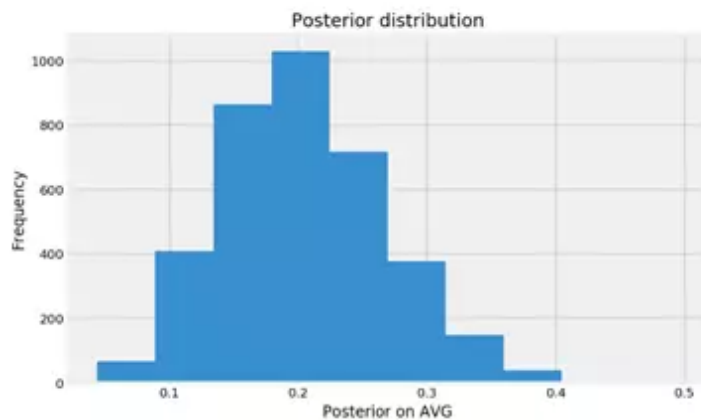
```
n_draw = 20000
prior_trials = int(dominic_smith_spring.iloc[3].AB)
prior_success = int(dominic_smith_spring.iloc[3].H)
prior_i = pd.Series(np.random.beta(prior_success+1, prior_trials-prior_success+1, size = n_draw))
plt.figure(figsize=(8,5))
plt.hist(prior_i)
plt.title('Beta distribution(a=%d, b=%d)' % (prior_success+1,prior_trials-prior_success+1))
plt.xlabel('Prior on AVG')
plt.ylabel('Frequency')
```

```
Out[9]: <matplotlib.text.Text at 0x11969b1d0>
```



```
posterior(ds_n_trials, ds_k_success, prior_i)
```

```
Number of draws left: 3662, Posterior mean: 0.205, Posterior median: 0.200, Posterior 95% quantile interval: 0.095-0.340
```



和在场景一中使用均匀分布的先验假设得到的后验结果相比，这里95%的分位数区域已经被缩小了。现在我有95%的把握断定DS的AVG会在0.095到0.340之间。

然而，一般来说AVG超过0.300已经是优秀的打者了，这里对AVG的估计意味着这名球员可以是最差或是最好的打者。所以我们需要更多的数据来缩小可信区间的范围。

场景三

在这个场景中，假设我不仅有2017年春季训练的统计数据，还有2017年常规赛的统计数据。那么这会如何影响后验结果和结论呢？

```
ds_url = "https://www.foxsports.com/mlb/dominic-smith-player-stats?seasonType=1"
dominic_smith_reg = batting_stats(ds_url, 'regular')
dominic_smith = dominic_smith_reg.append(dominic_smith_spring.iloc[3], ignore_index=True)
dominic_smith
```

```
Out[12]:
```

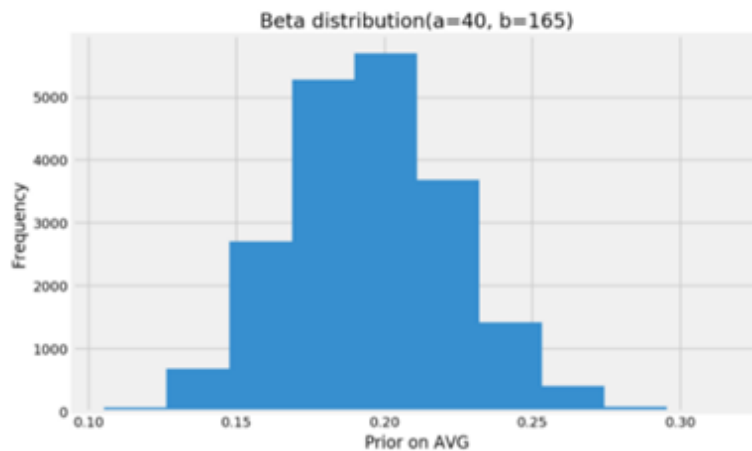
	BATTING	Team	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	AVG	OBP	SLG	OPS
0	2017	NYM	49	183	167	17	33	6	0	9	26	0	0	14	49	.198	.262	.395	.657
1	2017	NYM	20	42	36	4	6	1	0	0	4	0	0	5	13	.167	.286	.194	.480

```
ds_prior_trials = pd.to_numeric(dominic_smith.AB).sum()
ds_prior_success = pd.to_numeric(dominic_smith.H).sum()
n_draw = 20000
```



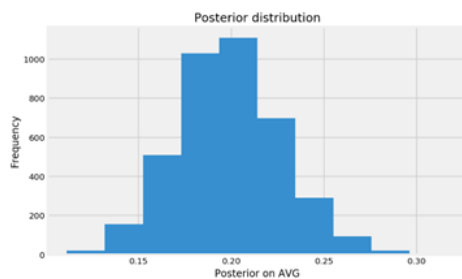
```
prior_i_02 = pd.Series(np.random.beta(ds_prior_success+1, ds_prior_trials-ds_prior_success+1, size = n_draw))
plt.figure(figsize=(8,5))
plt.hist(prior_i_02)
plt.title('Beta distribution(a=%d, b=%d)' % (ds_prior_success+1,ds_prior_trials-ds_prior_success+1))
plt.xlabel('Prior on AVG')
plt.ylabel('Frequency')
```

Out[15]: <matplotlib.text.Text at 0x11a46acd0>



```
posterior(ds_n_trials, ds_k_success, prior_i_02)
```

Number of draws left: 3929, Posterior mean: 0.199, Posterior median: 0.198, Posterior 95% quantile interval: 0.146-0.258



现在我有95%的把握断定DS的真正AVG将在0.146到0.258之间。虽然范围不是很精确，但与场景一和场景二相比，场景三的可信区间要窄得多。

场景四

我想比较两名选手，看看谁在AVG方面表现得更好。我观察的数据来自2018年春季训练，先验知识是2017年的春季训练和常规赛。现在我要比较DS和GC这两名选手的打击率。

在场景三中，我剔除了所有生成的结果与观察数据不一致的参数，然后进行模拟采样。但是这种类型的随机样本生成和过滤计算量很大，并且运行缓慢。

因此，我们可以借助一些工具使采样器在高概率的区域花费更多的时间以提高效率。像Pymc3这样的概率编程工具可以通过使用诸如HMC-NUTS之类的巧妙算法来有效地处理采样过程。

Pymc3链接：

<https://github.com/pymc-devs/pymc3>

HMC-NUTS链接：

<http://blog.fastforwardlabs.com/2017/01/30/the-algorithms-behind-probabilistic-programming.html>

我们先从Fox Sports中抓取Gavin Cecchini的统计数据。

```
gc_url_st = "https://www.foxsports.com/mlb/gavin-cecchini-player-stats?seasonType=3"
gc_url_reg = "https://www.foxsports.com/mlb/gavin-cecchini-player-stats?seasonType=1"
gavin_cecchini_spring = batting_stats(gc_url_st, 'spring')
gavin_cecchini_reg = batting_stats(gc_url_reg, 'regular')
gc_n_trials = int(gavin_cecchini_spring.iloc[1].AB)
gc_k_success = int(gavin_cecchini_spring.iloc[1].H)
gc_prior = pd.DataFrame(gavin_cecchini_reg.iloc[1]).transpose().append(gavin_cecchini_spring.iloc[0])
gc_prior
```

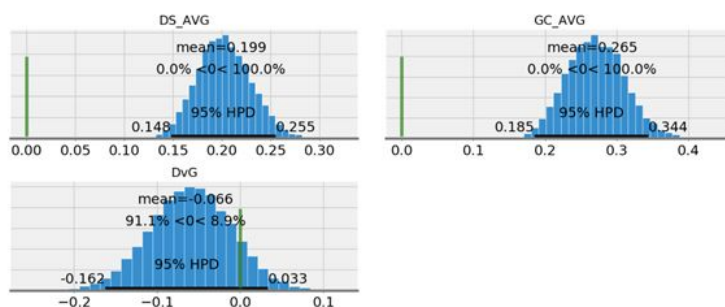
```
Out[20]:
```

	BATTING	Team	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	AVG	OBP	SLG	OPS
1	2017	NYM	32	82	77	4	16	2	0	1	7	0	1	4	19	.208	.256	.273	.529
0	2017	NYM	11	29	27	6	9	2	0	1	3	1	1	2	5	.333	.379	.519	.898

```
gc_prior_trials = pd.to_numeric(gc_prior.AB).sum()
gc_prior_success = pd.to_numeric(gc_prior.H).sum()
def observed_data_generator(n_try, observed_data):
    result = np.ones(observed_data)
    fails = n_try - observed_data
    result = np.append(result, np.zeros(fails))
    return result
ds_observed = observed_data_generator(ds_n_trials, ds_k_success)
gc_observed = observed_data_generator(gc_n_trials, gc_k_success)
```

接着，我们拟合一个Pymc3模型。

```
import pymc3 as pm
with pm.Model() as model_a:
    D_p = pm.Beta('DS_AVG', ds_prior_success+1, ds_prior_trials-ds_prior_success+1)
    G_p = pm.Beta('GC_AVG', gc_prior_success+1, gc_prior_trials-gc_prior_success+1)
    DS = pm.Bernoulli('DS', p=D_p, observed=ds_observed)
    GC = pm.Bernoulli('GC', p=G_p, observed=gc_observed)
    DvG = pm.Deterministic('DvG', D_p - G_p)
    start = pm.find_MAP()
    trace = pm.sample(10000, start=start)
pm.plot_posterior(trace, varnames=['DS_AVG', 'GC_AVG', 'DvG'], ref_val=0)
```



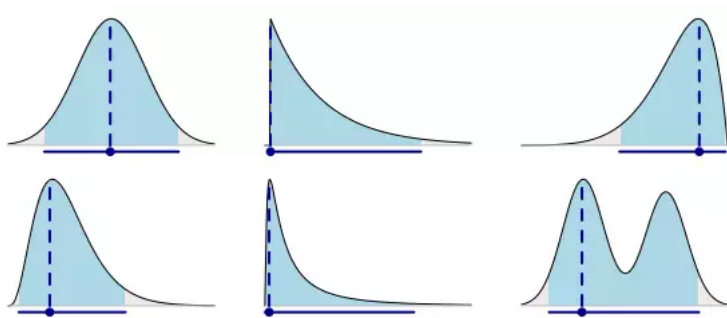
如果我们用Pymc3中的plot_posterior函数绘制DS_AVG、GC_AVG和DvG (DS_AVG - GC_AVG)的后验分布，我们可以看到图中出现的术语是HPD而不是分位数 (quantile)。

最大后验密度 (Highest Posterior Density, HPD) 区间是我们可以对后验密度函数使用的另一种可信区间。HPD区间会选择包括众数在内的最大后验概率密度值所在的最窄区间。

在Rasmus Bååth的另一篇文章中，比较了分位数区间和最高密度区间，并提供了简单明晰的对比图。以下是六种不同后验分布中的众数和覆盖了95%的概率密度的最高密度区间。

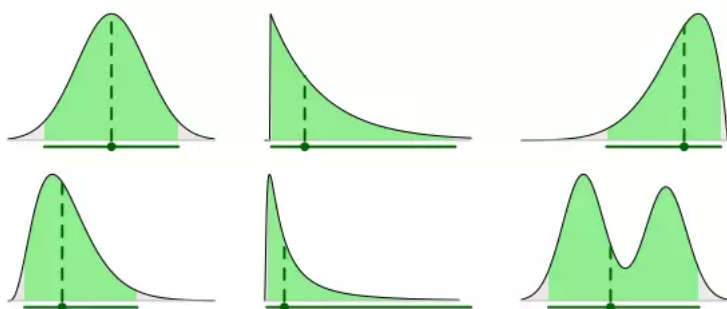
文章链接：

<http://www.sumsar.net/blog/2014/10/probable-points-and-credible-intervals-part-one/>



可能的点和可信区间，第1部分：图形总结

分位数区间包含中位数，中位数落在区间左侧的概率是50%，落在右侧的概率也是50%，同时以95%的可信区间为例，落在区间任意一侧的概率是2.5%。



可能的点和可信区间，第1部分:图形总结


就DS和GC的AVG来看，它们的众数和中位数看起来并没有多大区别，若实际情况确实如此，两位选手AVG的HPD区间和分位数区间应该也大致相同。让我们看看它们到底长什么样。

```
pm.summary(trace)
```

DS_AVG:

Mean	SD	MC Error	95% HPD interval	
0.199	0.028	0.000	[0.148, 0.255]	
Posterior quantiles:				
2.5	25	50	75	97.5
<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>				

GC_AVG:

Mean	SD	MC Error	95% HPD interval	
0.265	0.041	0.000	[0.185, 0.344]	
Posterior quantiles:				
2.5	25	50	75	97.5
				
0.189	0.237	0.264	0.293	0.349

DvG:

Mean	SD	MC Error	95% HPD interval	
-0.066	0.050	0.000	[-0.162, 0.033]	
Posterior quantiles:				
2.5	25	50	75	97.5
----- ===== ===== -----				
-0.166	-0.100	-0.066	-0.033	0.031

我们可以看到，对于DS和GC两名选手，HPD区间和分位数区间要么完全相同，要么仅在小数点后略有不同。

问题是，我想根据AVG来判断谁是更好的球员，目前看来，我还不能确定。至少我有95%的把握判定这两名球员的AVG相差无几。

计算的结果及生成的图形显示出这两名球员AVG的差异在-0.162到0.033之间（我们用DvG（DS-GC）表示他们AVG的差异，如果DvG为正表示DS更好，反之则GC更好）。

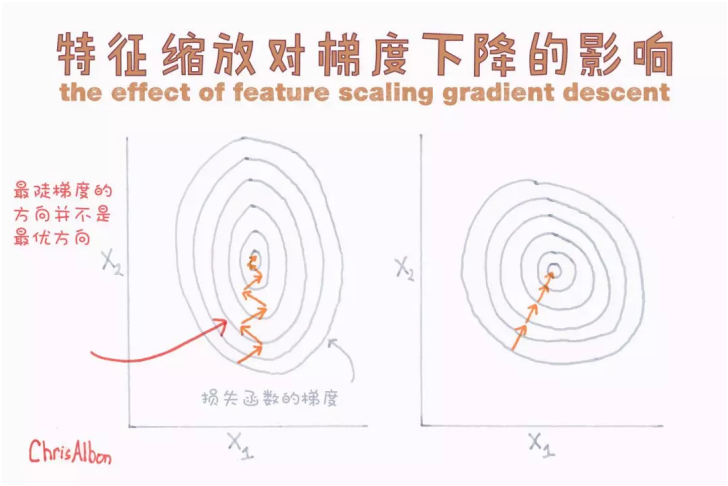
从结果来看区间包括了0.000，这代表两名球员的AVG不存在差异。因此，即使有证据表明GC比DS更优秀（因为DvG的后验分布在负值区域的面积比在正值区域的面积更大），但是我有95%的把握判定**这两名球员的AVG并无差异**。

也许有了更多的数据后，我可以确定他们之间的差异。毕竟，这就是贝叶斯理论的精髓所在。并不是说真相不存在，而是了解真相的过程很缓慢，随着技术的不断进步，我们能做的就是不断修正我们的认知。

完整版Jupyter Notebook的链接：
<https://github.com/tthustla/Bayesball/blob/master/Bayesball.ipynb>

原文链接：
<https://towardsdatascience.com/bayesball-bayesian-analysis-of-batting-average-102e0390c0e4>

【今日机器学习概念】



Have a Great Definition