介紹

• 在上一章中，我們看到了一個

理論分佈數，在

我們提供（通過 大尺寸

觀測值）某些參數（例如

二項式中的 n 和 p，泊松中的 =np，或 

與  在正規分佈中）到完全

描述這種分佈。

• 在現實中，當大規模觀察時

是不可能的，一個必須訴諸

過程稱為取樣（取樣）以獲得

這些所需的參數。5

抽樣（維琪）

• 抽樣是統計的一部分

與

個人選擇

旨在產生一些結果的觀察結果

關於人口的知識

關注，特別是出於以下目的

統計推斷。

樣本與總體

6

示例 1

• 描述美國的膽固醇水準

20歲至74歲的男性，我們

需要供應  和  來表征

正態分佈。

• 用於產生這些結果的樣品

數量必須提供準確的

人口的代表性。（對於

例如，如果我們只抽樣60歲

男人，  會太大。

為什麼？7

重要問題 - 1

• 樣品選擇必須是隨機的;

人口中的每一個人都應該

有平等的機會存在

選擇。

- 我們如何知道如果一個特定的

批次選擇不好？

– 即基於統計特徵

在這個特定的樣本上將是

與其他一些有很大不同

我們可能從中得到的樣本

人口。8

重要問題 - 2

• 樣本量也應該很大

足以產生更可靠的

估計這些參數（例如 

和  如果基礎人群是

假設具有正態分佈）。

• 考慮的樣本量

夠大嗎？

9

抽樣分佈

• 遵循前面的例子（膽固醇

20 至 74 歲美國男性的水準

old），其中  是均值，而  是

標準差。

• 讓我們隨機選擇 n 的樣本

觀測值並計算平均值

在此示例中，將其稱為 x1。

• 讓我們隨機選擇另一個樣本

n 個觀測值並計算平均值

在此示例中，將其稱為 x2。10

•如果這個過程繼續下去，我們將有一個

xi的集合，每個都是一個樣本

隨機選擇的 n 個的平均值

觀察。

• 考慮這個系列的xi

隨機變數 X 的結果

（代表

整個總體），概率

X 本身的分佈稱為

均值的抽樣分佈

大小為 n 的樣本數。

11

人口

（體重

首先-

年 CGU

學生）

人口

平均值  （？）

人口

標準

偏差 

(?)

體重

為 100

隨機

選擇

學生

體重

為 100

隨機

選擇

學生

體重

為 100

隨機

選擇

學生

...

採樣均值 x1

採樣均值 x2

採樣均值 xm

12

續

• 當然，在實踐中，我們不會這樣做

許多採樣並選擇一個最佳

從中取樣。

• 瞭解一些重要的

理論性質

平均值的分佈

（各種 抽樣手段），

但是，允許我們進行推斷

基於大小為 n 的單個樣本。

13

總結

• 例如，我們需要知道兩個

參數  和  以便使

使用理論概率

密度分佈在描述

一般人群（一套完整的

觀察結果）遵循正常值

分配。

• 這些參數（ 和 ），

然而，通常是未知的。

14

續

• 儘管人口樣本可以是

用於計算這些參數，

我們不知道採樣是否是

一個好與否。

• 抽樣分佈有助於

選擇要使用的“理想”參數。

• 例如，獲得“理想”人口

平均值  要使用的，從

許多平均值的分佈

許多樣本。 （樣本意味著現在

成為新的隨機變數。

15

總體與樣本

• 人口統計

– 人口平均數和人口性病

– 例如，平均膽固醇水準

在美國，所有20至65歲的男性均為145.5。

• 樣本統計

– 樣本均值和樣本 STD

– 例如，一個的平均膽固醇水準

從所有20人中選擇的100人團體

在美國，65歲的男性是123.5歲。

• 如果這些

兩個統計數據是可比的。

他們是

可比較的？

16

示例 2

• 只有 3 個觀測值的極端情況

的 {1， 2， 5}。這是整個人口，與

=2.6667 和 =1.7\*

觀察 X X-mu ^2

1 -1.6667 2.778

2 -0.6667 0.444

5 2.3333 5.444

總和 = 8.667

=2.6667 平均

標準 2.0817

標準電壓 =1.7\*







 

n

我

十一 x

s n 1

2 （ ）2

（ 1）

1







n

我

十一 x

1

（ ）2





 

n

我

十一 x

n 1

2 1 （ 17）2

“樣本”的方差

“總體”的方差

觀察方差基於

“樣本”和“總體”是不同的！

• 考慮 n = 2 的樣本。

• 有9個這樣的樣品（見下一頁

幻燈片）。

• 從中觀察各種統計數據

樣本，包括樣本均值。

• 觀察這些樣品如何-

基於參數可能“目標”

總體參數。

18

19







 

n

我

十一 x

s n 1

2 （ ）2

（ 1）

1





 

n

我

十一 x

n 1

2 1 （ ）2

每個的差異

樣本 （n=2） 為

由此計算

公式。

總體方差

（n=3） 由以下公式計算

公式。

20

結論

• 採樣變異性：採樣均值取決於

在樣本中包含的特定值上，

通常 因樣品而異。

• 所有可能的樣本均值均值為

等於原件的平均值

人口。

• 所有可能的樣品標準品的手段

但是，變體 （s=1.3） 是不同的

並且小於人口標準

偏差 （=1.7）

21

>> X=[1.0 1.5 3.0 1.5 2.0 3.5 3.0 3.5 5.0]

>> 子情節（2，2，1）

>>歷史（X）

>> 子圖（2，2，2）

>>歷史（X，7）

>> 子圖（2，2，3）

>>歷史（X，5）

>> 子情節（2，2，4）

>>歷史（X，3）

樣本均值的分佈

雖然不明顯

在這裡，我們期待看到一個

正態分佈

這些樣本意味著。22

10 7

5 3

8.2 中心極限定理

23

中心極限定理

• 對於特定人群，其中  是

均值和  標準差，此

定理陳述如下：

– （1） 抽樣分佈的均值

的樣本均值與

人口平均值 。

回想一下，在示例 2 中，我們有均值

的採樣均值為 2.7， 等於

人口平均為2.7。

24

• （續）：

– （2） 抽樣的標準偏差

樣本均值的分佈為 /sqrt（n）。

（這稱為標準誤差的

取樣時的平均值 （SEM）

這告訴我們為什麼我們更喜歡更大的樣本量

（使 SEM變小），以更好地確保

我們從樣本中選擇的平均值很好

足以代表一般人群。

回想一下，在示例 2 中，我們有 s = 1.3 表示

採樣均值的均值，並且  = 1.7

人口。事實上，1.3 與 1.7 大致相同

除以 2 的平方根。

25

• （續）：

– （3） 當 n 較大時，採樣

樣本均值的分佈為

大致正常。

26

請注意，這些是

樣本均值

n=2。

評論

• 上述中心定理適用

對於任何具有有限標準的人口

偏差 （），無論形狀如何

基礎分佈。

• 離正常離譜越遠

然而，分散式將需要

較大的 n 以確保正態性的

抽樣分佈。（通常是 樣本量

30 就足夠了。

27

8.3 更多例子

28

• 考慮血清的分佈

所有20至74歲的膽固醇水準

美國男性，平均 =211 和 =46。

• 請記住，這些被稱為人口

均值和總體標準

偏差，與平均值和

從

特定樣本的尺寸有限。

示例 3

29

• 考慮 n=25 的採樣大小。（我們

可能有許多這樣的樣本，每個

包含25個觀測值。

• 問題 - 什麼比例

這許多樣本（每個

n=25） 的平均值為

230 或更高？（回想一下

人口平均值為 211。

30

請注意，我們不是在問“什麼比例”

的人群將有膽固醇水準

230 或更高」。

求解前分析

回想一下，對於

整個人口。

如果我們的答案很小（例如，只有10%的

n=25 的樣本可能具有均值

值為 230 或更高），則許多 \*

這些樣本（n=25 中的每一個）將有一個

平均值為 230 或更小，即

更接近實際平均值 211..

\* 根據 中心極限定理，分佈

這些樣本均值近似正常。 所以 會有

樣本的 40% 均值介於 211 和 230 之間。31

50 100 150 200 250 300 350 400

0

1

2

3

4

5

6

7

8

9

x 10-3

根據血清分佈

人群的膽固醇水準~~

平均值

25人

平均值

25人

平均值

25人

平均值

25人

平均值

25人

平均值

25 人平均值=211

230 或更高

我們無法回答這個問題

提出問題，因為

水平軸不

表示正確

隨機變數我們是

感興趣。

32

180 190 200 210 220 230 240 250

0

0.01

0.02

0.03

0.04

0.05

0.06

平均值

25人

平均值

25人

平均值

25人

平均值

25人

平均值

25人

平均值

25 人平均值=211

230 或更高

抽樣均值

血清分佈

膽固醇水準

25 20-74歲

美國的男性

給定

n=25~~ 的樣本均值

33

所以對於這個特別的人來說會是值得懷疑的

樣本（n=25），平均值=230

代表整個人口。[因為

有許多較小的手段

更接近 211 的實際平均值，

這些顯然是更好的 樣本。]

34

想像一下，我們的25個觀察結果大多是

取自老年人，平均值

可能會高於

人口平均數為 211，並且可能

高於規定的 230 閾值。

在此特定中採樣量為25

案例不會太可靠給我們

準確估計  用於

代表整個人口（20至

74歲）。

35

如果特定樣本均值太遠

關閉（已經足夠遠）實際平均值

211 在抽樣均值分佈中，它是

不容易找到其他採樣手段

這是更遠的。

這就是我們提出這個問題的原因

對於“我們的概率

採樣均值為 230 或

更大“。[回想一下，採樣均值為

也是 一個隨機變數，它遵循

正態分佈。

36

溶液

• 根據中心極限定理，我們將

具有 正態採樣分佈

均值，均值  = 211 且

標準差  = 46/251/2 = 46/5 = 9.2。

• 因此，以下轉換將

將此採樣分佈提高到標準

=0 且 =1 的正態分佈。

9.2

Z  X 211

請注意，這是示例

樣本均值的分佈，

不是膽固醇水準

分配。

使用 9.2，而不是 46!!!37

>> x=[70：0.1：360];

>> Y1=1/（sqrt（2\*pi）\*46）\*exp（-0.5\*（（x-211）/46）.^2）;

>> Y2=1/（sqrt（2\*pi）\*9.2）\*exp（-0.5\*（（x-211）/9.2）.^2）

) )

46

211

(

2

1

exp（

2 46

1

1（ ）   2



Y x  x



) )

9.2

211

(

2

1

exp（

2 9.2

1

2（ ）   2



Y x  x



樣品分佈

的平均值

樣本數量 n=25

膽固醇分佈

230

38

>> F='1/（sqrt（2\*pi））\*exp（-0.5\*x^2）';

>> z=（230-211）/9.2

z = 2.0652

>> int（F，z，inf）

年份 =.19451217852135832736501724576047e-1

>>

• 現在，為了獲得解決方案，我們可以轉換

X=230 為 Z 得分，並找到

標準法線的右尾

分發，就像我們在上一講中所做的那樣。

• 因此，大約只有1.9%的幾率

對於此，平均值將超過 230

抽樣過程。（這是由

從上一張幻燈片中看到的窄鐘形。

標準正常

分配

從 2.0652 到  的積分以獲得區域

在右尾下方。

39

40

) )

9.2

211

(

2

1

exp（

2 9.2

1

2（ ）   2



Y x  x



樣品分佈

的平均值

樣本數量 n=25

膽固醇

分配

230

40

\*

\*\*

\*\*\*\*

1.9%

>> F2 ='1/（sqrt（2\*pi）\*9.2）\*exp（-0.5\*（（x-211）/9.2）^2）';

>>整數（F2，230，inf）

年 =.19451217852135832736501724576025e-1

>>

• 或者，方便使用時

MATLAB如下所示，我們甚至不必

將 X 轉換為 Z 分數。我們可以直接

在原始版本上執行集成

採樣分佈，從 230 到 inf。這

答案是完全一樣的。

請注意使用 9.2（受樣本量 25 的限制），而不是

總體標準差為 46。

41

解決方案摘要

• 只有1.9%的機會

其他樣本的平均值為

230 或更大。

• 我們的樣本均值 =230 與

實際人口 平均值。

•很難找到其他樣品有

更不準確的平均膽固醇值

比這個。

• 結論：這不是一個

良好的採樣。42

應用

• 同樣，我們可以計算，例如，

平均值範圍為 2.5% 以下

兩端尾巴。

• 這涵蓋了95%的平均值，所有可以

被認為是“足夠好”（對於這些

樣本）來表示整個總體。

43

• 可以證明 Z 值介於 -1.96 之間

和 +1.96，或平均膽固醇值

在 193.0 和 229.0 之間將呈現

這95%的比例，在抽樣量下

的 n=25。

• 嘗試取得這些間隔

使用 MATLAB！

44

討論 - 1

• 為什麼我們要說“覆蓋95%的

考慮 平均值

足夠好“（對於這些

樣本）來表示整個

人口？

•不會90%，甚至50%很多

更好？

45

討論 - 2

• 事實上，在這裡我們看到的是這個“好”

夠了“從另一個角度思考

周圍。

• 任何“不是”的東西

極端“被認為是

“好”。

• 在統計學世界中，5%是一般

被認為是「極端」。。

46

討論 - 3

• 兩條尾巴與一條尾巴？

• 非常低和極高

兩者都考慮值

極端？（例如，血壓也

低和太高都是不好的）

• 或僅非常低（例如，英年早逝）

或僅非常高（例如，空氣污染）

被認為是極端的嗎？

47

多大才算足夠好

對於採樣大小 n？

n /sqrt（n） 間隔覆蓋

95%的手段

長度

間隔

1 46.0 120.8 – 301.2 180.4

10 14.5 182.5 – 239.5 57.0

25 9.2 193.0 – 229.0 36.0

50 6.5 198.2 – 223.8 25.6

100 4.6 202.0 – 220.0 18.0

首選樣本量以獲得滿意

表示實際值的樣本均值

人口 均值。 間隔越窄，

採樣效果越好。48

此圖表示美國死亡時的年齡分佈

在1979-1981年。它有 =73.9 和 =18.1。請注意，我們不

這裡有一個正態分佈。我們 期望什麼

當我們從這個年齡人群中抽樣時，會發生什麼？

示例 4

49

4個隨機選擇的樣本

形成，具有其平均值和性病

在左側顯示。相應

顯示概率直方圖

下面。您可以看到可變性

從這樣的抽樣。

=73.9 和

=18.1

人口。

50

大小為 100 的 4 個樣本的直方圖

51

大小為 500 的 4 個樣本的直方圖

52

結論

• 這個例子向我們展示了這一點，即使

基礎分佈不正常，

具有較大 n 的採樣仍將具有

獲得樣品的機會非常好

與真實相似的統計數據

人口統計。

• 如前所述，n 可以更小，如果

基礎分佈已經是正態的，

並且必須更大，如果底層

分佈不正常。

53

示例 5

• 假設年齡人口在

死亡時間。平均值 =73.9

年和 =18.1 年。

• Q1：在大小為 25 的樣本中，有

從這個人群中抽取，什麼

平均年齡

死亡在70到78歲之間？

•Q2：如果出現以下情況，則Q1中的比例是多少

樣本數量是100？

54

3.62

73.9

18.1/ 25

Z  X 73.9  X 

• 根據中心極限定理，

這些樣本的平均值 X 將具有

以

人口平均數 73.9，標準

偏差 18.1 除以

樣本數量 25。

• 此正規分佈（平均值來自

許多 n=25 個樣本）可以轉換為

標準正態分佈為：

55

（ 1.08 1.13）

)

3.62

78 73.9

3.62

73.9

3.62

70 73.9

(

(70 78)

  



   

 

P Z

X

P

P X

>> F='1/（sqrt（2\*pi））\*exp（-0.5\*x^2）';

>>國際（F，-1.08，1.13）

年 = 0.73069079767121312631047998477208

56

1.81

73.9

18.1/ 100

Z  X 73.9  X 

（ 2.15 2.27）

)

1.81

78 73.9

1.81

73.9

1.81

70 73.9

(

(70 78)

  



   

 

P Z

X

P

P X

>> F='1/（sqrt（2\*pi））\*exp（-0.5\*x^2）';

>> 國際（F，-2.15， 2.27）

年 = 0. 97261860108700595588800257362372

57

對於 n = 100：

百分比遠高於 n=25

因為鐘形變窄了。