直方圖

單維彰・2013 年 4 月

本節複習統計數據的長條圖，然後介紹直方圖。但是本文講解的直方圖與絕大多數國高中數學教材不同，是「面積的」直方圖。

我們以台灣人民的「年收入」當作範例，闡述統計資料與機率之間的基本關係。民國 100 年全台灣約有 2300 萬人，其中約 1100 萬人為工作人口（含失業人口，詳細的定義從略）。這些工作者的年收入（單位：新台幣萬元）在 40 以下者，約佔 13%，在 40—80 者約佔 24%。忽略年收入 200 萬元以上者不計，簡單的統計表格如下。

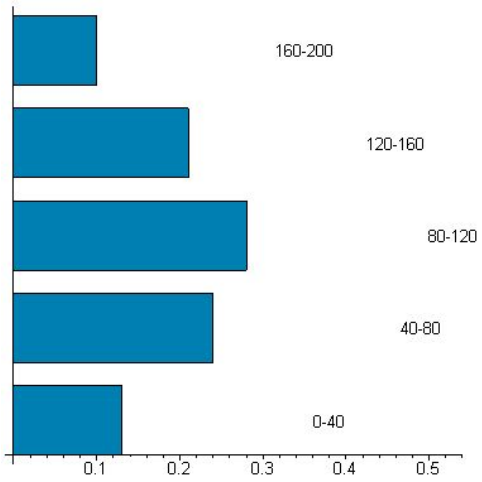
年收入（新台幣萬元）	0—40	40—80	80—120	120—160	160—200
佔工作人口比率	13%	24%	28%	21%	10%

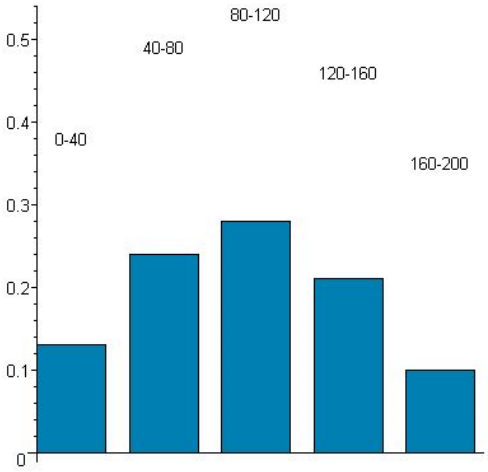
按照以上統計數據，我們「差不多」可以說，在台灣任抽一名工作者，其年收入在 40 萬元以下的機率是 0.13，其年收入在 40 至 80 萬元的機率是 0.24，…。我們還可以說，一名工作者的年收入介於 120 至 200 萬元的機率是 0.21+0.10=0.31，而年收入超過 200 萬元的機率是

$$1-(0.13+0.24+0.28+0.21+0.10)=0.04=4\%。$$

因為$1100\times 4\%=44$，我們可以推論大約有 44 萬人的年收入超過 200 萬。但是，我們無法推論收入在 90—100 萬元的機率，也不知道年收入不到 100 萬元的機率。因為上述表格以四十萬元為一組，就好像籮筐 (bin) 一般，把收入落在那四十萬範圍裡的相對次數全部丟在同一個籮筐裡；籮筐的寬度就稱為**筐寬** (binwidth)，以上表格的筐寬是 40。在筐寬為 40 的資料表格裡，我們只能每 40 讀取一段資料，無法讀取更詳細的資料。

為了瞭解直方圖，我們從長條圖說起。以上表格所顯示的數據資料，可以轉化為圖形。最常用的統計圖之一就是**長條圖** (bar chart)。長條圖可以畫成橫的或直的，如下（其實 bar 應該是橫的，直的稱為 column）。我們之後只談直的長條圖。





長條圖的特色如下：

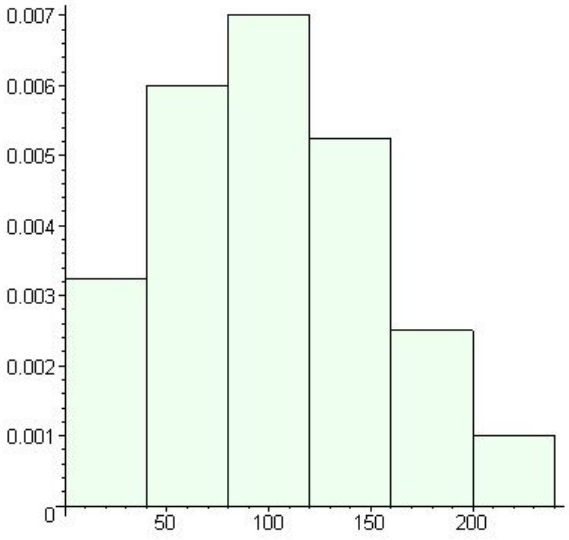
(1) 條條不相連，

(2) 橫坐標不是數線，通常在橫坐標下面以文字敘述每一條代表的意思，

(3) 每一條的寬度沒有意義，僅有其高度的坐標有意義。

在我們的例子中，長條圖的縱坐標就是表格內的比率（或機率）。畫圖並沒有增加資訊，只是比較容易比較相對的大小而已。在表現數據的藝術上，我們常看到不同狀態的同類資料畫成不同顏色的長條圖，以便交叉比較：例如將民國 80 年、90 年、100 年的收入比率畫在同一張長條圖上，以便跨年比較。我們不再細數長條圖的功能和作法，只用它來幫助理解直方圖。

根據同一份資料表格畫出來的**直方圖** (histogram) 如下。



相對於長條圖，可觀察其特色有

(1) 一定畫成直的，

(2) 條條相連，

(3) 橫坐標就表示做統計的數據（年收入，單位「新台幣萬元」），

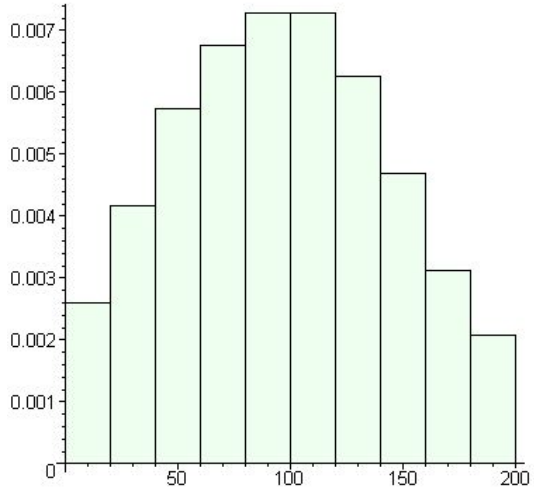
(4) 每一條的寬度有意義：其左右邊界的坐標就是數據分段點：0、40、80、120、

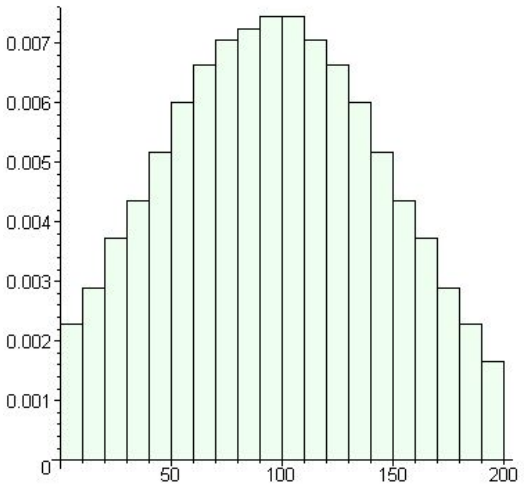
160、和 200，（為了製圖方便，我們把所有超過 200 萬的資料都併入 200—240 萬的區段內了）

(5) 每一條的高度，並不是表格中的比率（機率）。

相較之下，應該發現：長條圖比較「平易」，而直方圖比較「數學」。既然直方圖的高度不再是機率，那麼機率的資訊到哪裡去了呢？答案就在：面積。**面積代表機率**。直方圖上每一條長方形的面積，就是發生在那個區段內的相對次數，也就是機率。例如，最左邊跨越 0—40 區段的長方形高度是 0.0035，寬度是 40，相乘即為年收入低於 40 萬元的機率 0.13。因為機率的總和必須是 1，直方圖裡的長方形面積和為 1。

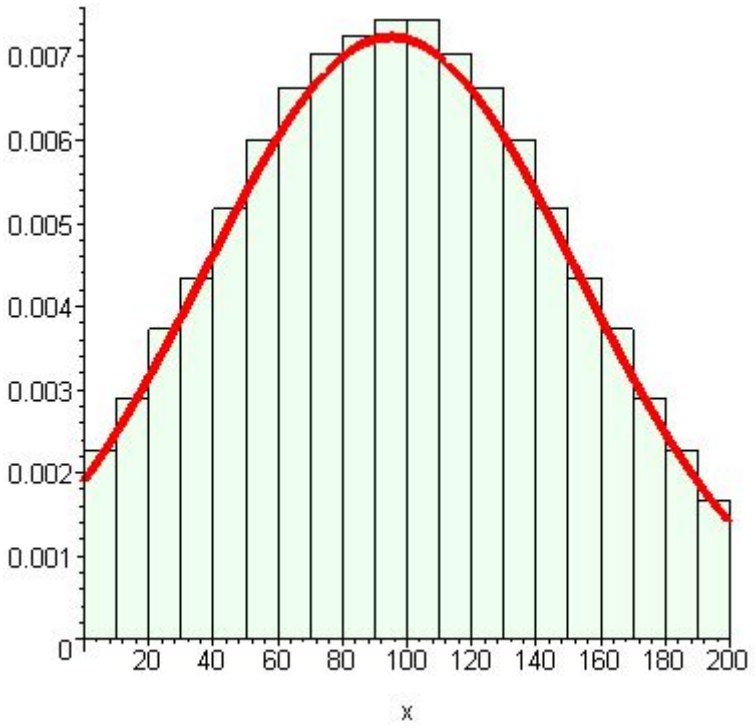
如果我們能獲得更詳細的個人收入資料，以便將年收入的區段分得更細，比如說每 20 萬元一個區段，則共分成十段，每段一條長方形，其直方圖如以下的左圖。我們看見每條長方形的寬度變窄了，但是總面積仍然維持是 1（下圖僅顯示 200 萬以下的部分，所以面積和略小於 1）。





同理，如果我們更詳細地將年收入的區隔切成每 10 萬元一段，共分成二十段，如以上的右圖。當我們將資料的分段切得越來越細，則長方形變得越來越多，但是它們的面積總和維持不變（是 1）。而且，每一條長方形的面積就是抽樣的數據（一名工作者的年收入）落在該區段內的機率。

用面積當機率的好處是：當筐寬變小，分割變細的時候，圖形的總面積不變，但每條長方形的高度略有增減。按照上述「切得越來越細」的趨勢，同學們可以看出來，那些長方形的頂部依稀形成一條連續的曲線。事實上，就像「插值多項式」一樣，我們的確可以為這些數據資料找到一個函數模型，如下圖中的紅色曲線。


0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
20
40
60
80
100
120
140
160
180
200
x