# 機率密度函數

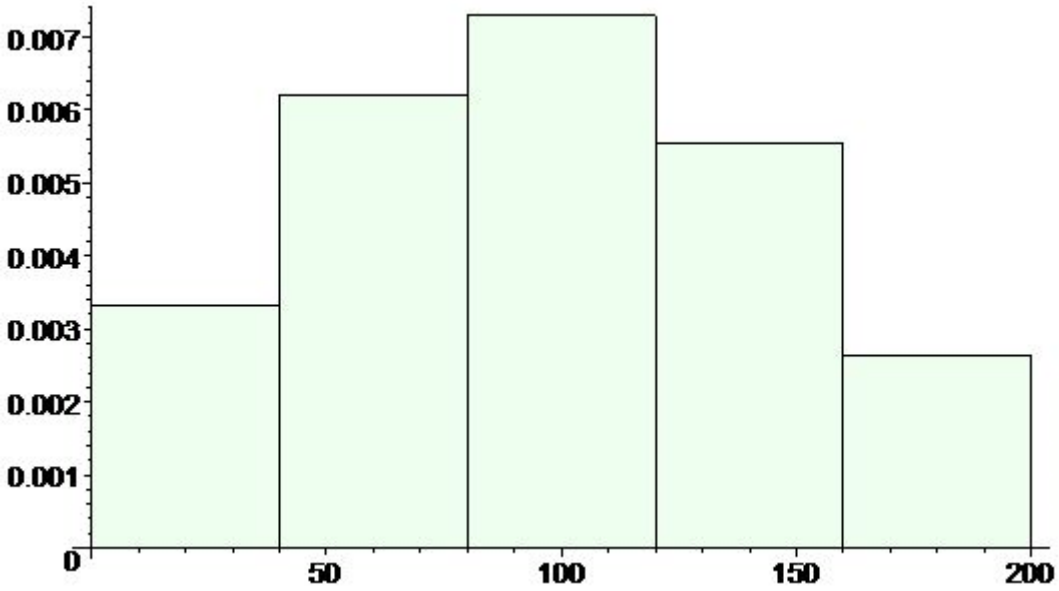
**單維彰・2015 年 5 月**

本節從「面積的」直方圖引伸出連續型隨機變數的機率密度函數。

民國 100 年全台灣約有 1100 萬工作人口（含失業人口，詳細的定義從略）。從這些工作者當中隨機抽選一名，令其年收入（單位：新台幣萬元）為隨機變數 $x$。雖然通常年收入的台幣數值是整數，但因為處理大量數據而可以將 $x$ 視為實數，而其理論值域為 $[0,\infty)$。下圖是以 40 為筐寬的（面積）直方圖，由左至右每條長方形的面積依序為

0.13　0.24　0.29　0.21　0.10

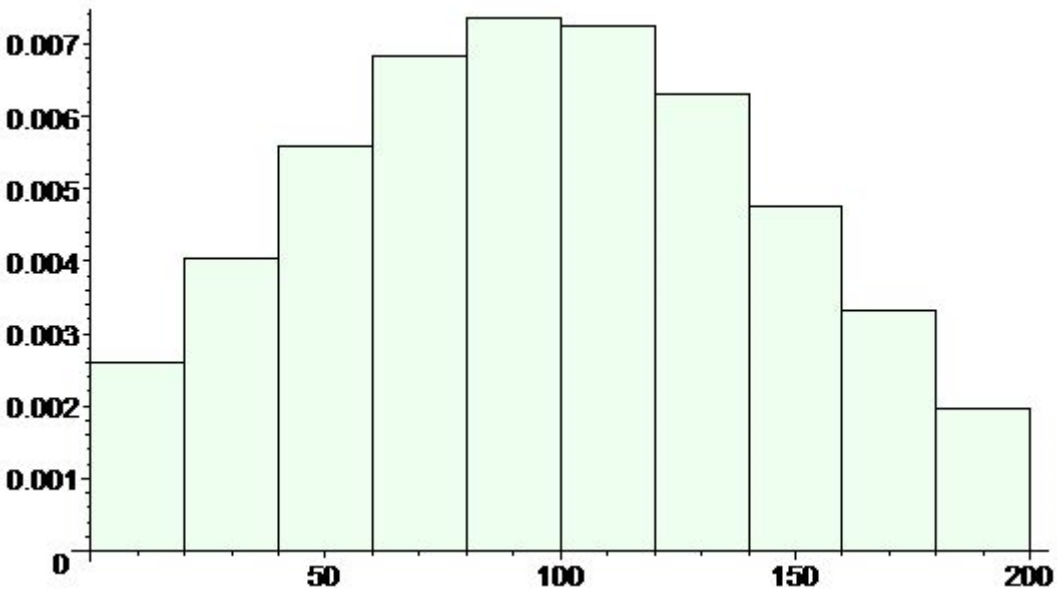
而其意義為 $x$ 落在該範圍內的機率。



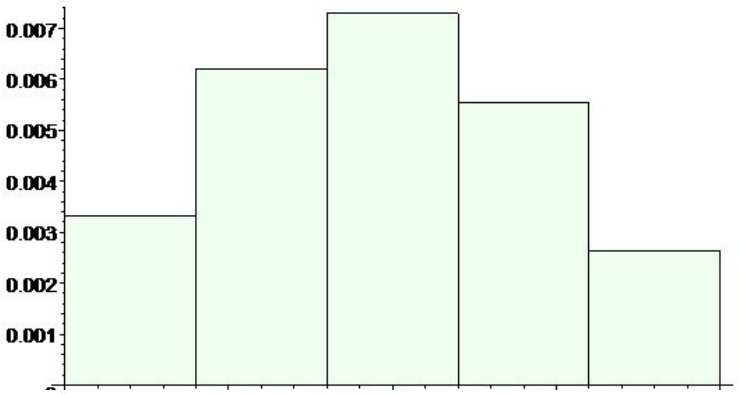

例如 $P(0 \le x \le 40) = 0.13$、$P(40 < x \le 80) = 0.24$、$P(80 < x \le 120) = 0.29$ 等等。為節省空間，我們忽略了 $x > 200$ 的數據，但可以計算年收入超過 200 萬元的機率是

$$P(x > 200) = 1 - (0.13 + 0.24 + 0.28 + 0.21 + 0.10) = 0.04 = 4\% \text{。}$$

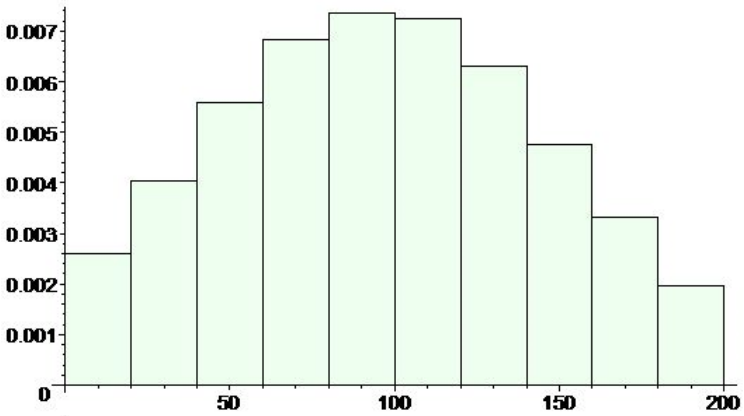
因為以上直方圖的筐寬是 40，我們無法獲得更細節的機率。譬如，若想知道 $p(40 < x \le 60)$，則只能使用內插估計而不能從直方圖的數據查得。但如果我們能夠查到原始的年收入數據，則可以用比較窄的筐寬（例如 binwidth 為 20）重新製作直方圖，如下，而從圖中讀出第三條長方形的高度，從而算出面積（也就是機率），獲知 $p(40 < x \le 60) = 0.11$。



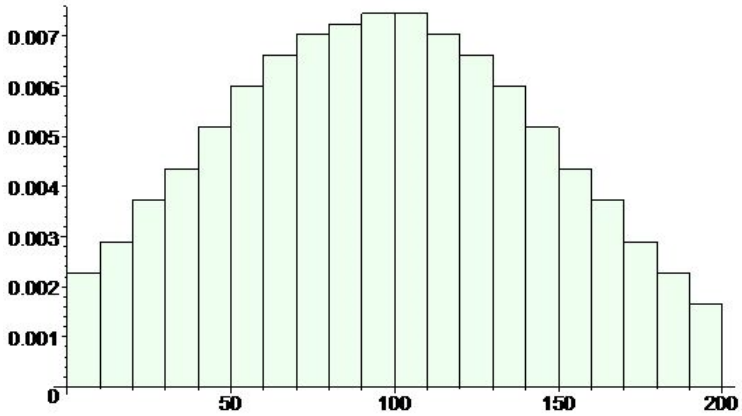

為方便比較，我們將筐寬為 40 和 20 的圖分置上下，如右。用面積當機率的好處是：當筐寬變小，分割變細的時候，圖形的總寬度不變，但因為每條長方形的面積改變了，所以高度也略有改變。



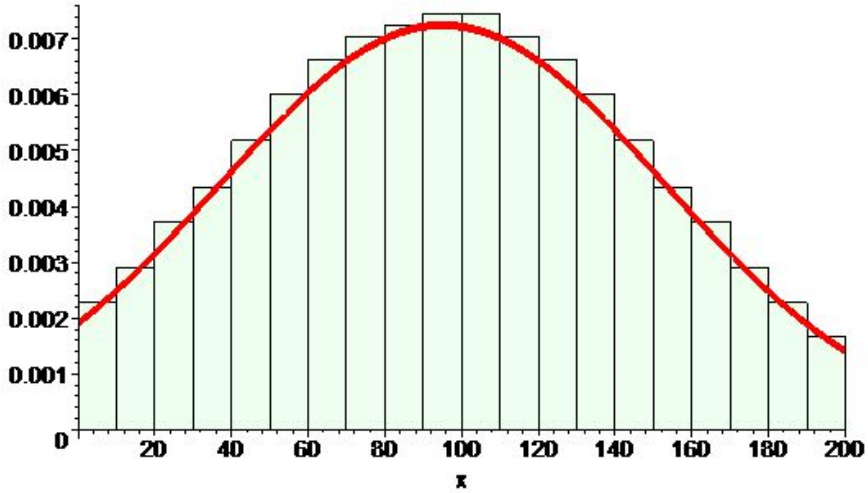

把筐寬切半之後，原來的每條長方形被切為兩條；兩條新長方形的寬度為舊長方形的一半，但兩條新長方形的面積和等於舊長方形。因此，兩條新長方形的平均高度應該等於舊長方形的高度，也就是兩條新長方形的高度一升一降而維持面積和為常數。例如右圖中，上面的 $40 < x \le 80$ 拆成 $40 < x \le 60$ 和 $60 < x \le 80$ 兩個新長方形，舊長方形的面積是 0.24，兩條新長方形的高度略微一升一降，面積依序為 0.11 和 0.13，面積和等於舊長方形。



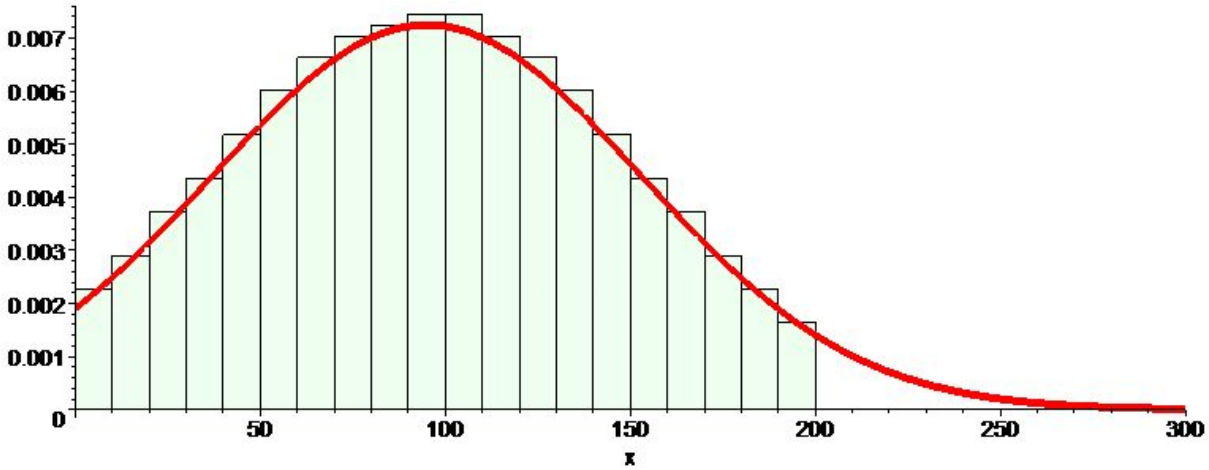

如此這般，（如果資料充足）當筐寬越來越細，直方圖就變成越來越窄的長方形，而這些長方形的頂端，會越來越接近一條曲



線。確實是的，如下圖的紅色曲線。




這條紅色曲線暗示年收入在 200 萬以上的分布，我們可以拉長顯示曲線的 $x$ 範圍，如下圖。




橫軸的單位是「萬元」，面積是機率，縱軸是大約 0.007 以下的正數，那麼縱軸的單位是什麼呢？它的單位是 $\frac{\text{機率}}{\text{萬元}}$。回顧質量密度的單位是 $\frac{\text{Kg}}{\text{m}^3}$，人口密度的單位是 $\frac{\text{人數}}{\text{km}^2}$，我們稱 $\frac{\text{機率}}{\text{隨機變數的單位}}$ 為**機率密度**。上面那條紅色曲線，就是所謂**機率密度函數**的圖形。以上範例中的機率密度函數，是右斜 (right skew) 正規分布，超出本課程目前的進度，所以我們不再深入討論。重點是：

**機率密度函數的曲線下面積是機率**