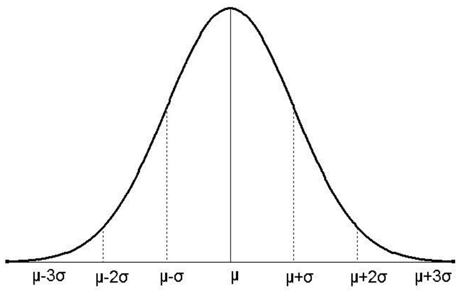
13 讓數字說話……統計



68-95-99.7 規則

高斯的常態分佈 [1]

統計談的就是如何蒐集、整理、分析和呈現數據……讓數字說話。統計是從數據中找出訊息，並且做成結論，而統計使用的工具是圖表和計算，並加上常識判斷。美國統計學會與美國數學學會的聯合課程委員會就建議，入門的統計課程都應該「強調如何作統計思考」，而且內容應該「多一些數據和觀念，少一點公式和推導過程。」統計可分成敘述統計與統計推論這兩大類。

敘述統計是以列舉一些指標或以繪圖、數學模型加以說明抽樣資料之特性；而統計推論是從抽樣資料中盡可能地得到全部資料的全貌，也就是從樣本資料來推論整個母體資料的特徵。無論是敘述統計或者是統計推論，首要工作都是如何取得可以反應母體特性的樣本資料。取得資料的過程就是抽樣調查，就讓我們從抽樣調查談起吧！

[1] 常態分佈都是對稱、單峰及像鐘形的曲線。因為對稱的關係，所以平均數、中位數及眾數落在曲線的中間位置，這也是尖峰點所在。常態分佈左右各有一個稱為「反曲點」的點，該點至對稱軸的距離剛好是一個標準差。常態分佈最迷人的地方就是所謂的 $68-95-99.7$ 規則，即約 68% 的資料會落在對稱軸正負 1 個標準差的範圍內，95% 的資料會落在對稱軸正負 2 個標準差的範圍內，99.7% 的資料會落在對稱軸正負 3 個標準差的範圍內。各種民意或抽樣調查幾乎都在使用 $68-95-99.7$ 這個規則。在歷史上，高斯是第一位運用常態分佈的人，有時又稱常態分佈為高斯分佈或鐘形曲線。

13.1 數據的蒐集與產生……抽樣調查

有句諺語說：「你不必吃完整頭牛，才知道肉是老的。」這就是抽樣的精髓：從檢查一部分來得知全體。

如何從母體中抽取一些具有代表性的個體（能充分反應母體的個體）是很不容易的一件事情。某叩應節目要求觀眾叩應，來表達它們是否贊成興建核能電廠。這類叩應通常要求贊成就打某支電話，不贊成就打另一支電話。他們是要求大家打電話進來，而不是主動抽取樣本。因為不贊成興建核能電廠的人對這個議題比較有強烈的感受，比較會將意見表達出來，甚至會多打電話。因此所得的調查結果是有所偏差的。又如美國專欄作家藍德絲有一次問她的讀者：「如果可以重來一次，你還要孩子嗎？」藍德絲接到接近 1 萬份答覆，其中 70% 說：「不要！」難道說 70% 的父母都後悔有了孩子嗎？當然不是，這是個自發性的回應樣本。通常對某個議題有強烈感受的人，尤其是負面感受的人，比較會不嫌麻煩地去回應，所以藍德絲的意見調查結果當然是有高度偏差的。在這兩個例子中，都是由於人為的因素而造成偏差，抽取的樣本沒辦法反應母體的實際狀況。統計學家的補救方法，是利用不牽涉人為選擇的「機率」來選取樣本。用機率選樣本，是藉由給每個個體有同樣的中選機會，來消除偏差。

抽樣調查比起普查的優點：假如你要測試鞭炮或保險絲的功能是否正常，測試過的產品就毀了，不能採取普查；要職員去清點庫存的所有 50 萬個備份零件，不如要他去抽取一個樣本好好清點，因為職員數煩的時候，就會愈數愈不準了。

13.1.1 普查

普查就是企圖把整個母體納入樣本的抽樣調查。大規模的普查是費錢又費時的工作，隨著科技的日新月異與大型抽樣調查的完善，廣泛普查工作已經漸漸不流行了。

13.1.2 簡單隨機抽樣

用機率選樣本最簡單的方法是，把全部個體的名字放進一個箱子（這就是母體)，然後從當中抽出一部份（也就是樣本)。當抽樣時不摻入人為因素，而且母體中每一個體被抽中的機會均等時，就稱這種抽樣方法為簡單隨機抽樣。

民調機構為了節省經費與時間，常經使用簡單隨機抽樣來進行他們的意見調查。當他們完成抽樣訪問或電話訪談之後，就是向社會大眾公布他們所做的結果。因為是抽查，難

免所得的數據與實際情況有所誤差，所以在公布數據的用語上，有一定的規範。例如：蓋洛普或一般的民意調查機構常用的信賴敘述說法如下：「對這次抽樣調查所得的結果來說，我們有 95% 的信心認為“過去 12 個月中，有 50% 的美國人曾經購買過樂透彩券”，而且調查的誤差範圍應該在正負 3 個百分點以內。」這意思是說：對這樣的抽樣調查重複進行 100 次的話，他們認為會有 95 次所得的數據（曾經購買過樂透彩券的百分比）介於

$$50\% - 3\% = 47\% \quad 與 \quad 50\% + 3\% = 53\%$$

之間。這當然也意味著另外 5 次所得的數據有可能不在這個區域內。至於“正負 3 個百分點”是根據什麼資料得來的呢？這牽涉到比較複雜的統計知識。不過若調查機構所使用的抽樣方法為簡單隨機抽樣的話，而且抽樣的數據是 n 個，那麼有一個誤差範圍的速算法為“正負 $\frac{100}{\sqrt{n}}$ 個百分點”。以本問題為例，因為是“正負 3 個百分點”，所以可以推敲抽樣的個數 n 會滿足

$$3 = \frac{100}{\sqrt{n}} \Rightarrow n = 1111.11\cdots.$$

故此次抽樣調查應該抽取 1100 個樣本左右。

“正負 $\frac{100}{\sqrt{n}}$ 個百分點”這則速算法是在 95% 的信心之下所決定的。若將信心百分比作改變，那麼誤差範圍的公式也會跟著修正。誤差範圍在統計學上有較精確的計算公式，不過這需要瞭解常態分佈之後才能談，就把他留到後頭再說吧！

例題 1　某報紙為了對某些女性議題作民調，從全國隨機抽取 1024 位婦女來訪問。調查發現，在 95% 的信心水準下，有 50% 的婦女說：「她們沒有足夠的個人時間。」求該次調查的誤差範圍是正負幾個百分點。

【解】由誤差範圍速算法知道：誤差範圍是正負

$$\frac{100}{\sqrt{1024}} = \frac{100}{32} = 3.125$$

個百分點。

13.1.3 分層隨機抽樣

如果母體具有某一性質，那麼先依該性質將母體分成若干層，再對每一層進行簡單隨機抽樣。這綜合各層的樣本，應更能反應母體的數據。像這樣的抽樣方法稱為分層隨機抽樣。例如，某大學有 6000 位男生，4000 位女生，想調查他們每天上網的平均時間。隨機抽取 60 位男生，40 位女生，再計算這一百位學生上網的平均時間。像這樣將學生分成男、女兩層，再針對每一層作抽樣的方法就是分層隨機抽樣。

分層隨機抽樣的弱點：分層隨機抽樣的樣本未必有給母體中每個個體同樣被抽中的機會，因為有些層在樣本中所佔比例有可能被刻意提高或壓低。

13.1.4 系統與部落抽樣

如果母體有 10 萬個個體，欲抽樣 1000 個個體當樣本，相當於每 100 個抽一個。於是先將母體編號，從 000000 號編到 99999 號，其次從 00 到 99 這一百個數字隨機抽取一個號碼（比如說抽到 38 號），那麼編碼的尾數是 38 的個體共有 1000 個，就抽樣這 1000 個個體當樣本。像這樣的抽樣方法稱為系統抽樣。比方說，為了調查市長辯論比賽的收視率，以台北市的電話簿為依據，每隔兩百個電話號碼就電話訪問，這就是系統抽樣。

值得一提的是：當母體的性質與其編號成規律性的變化時，不宜採取系統抽樣。

假設我們想要調查台北市大安區的家庭每月平均收入。因為大安區每一個里的條件相當，所以只需從大安區的幾十個里中，抽取一兩個里作普查即可得知平均收入。像這樣的抽查稱為部落抽樣。

13.2 數據的整合與呈現……圖形、表格與數字的呈現

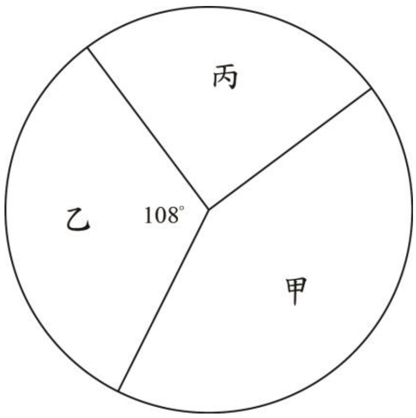
文字本身不構成故事，得由作者把文字組成句子，再把句子組成故事。如果文字組合得不好，故事可能讓人看不太懂。統計數據也一樣，如果要讓人看清楚統計數據所隱含的訊息，一樣需要經過整合。應該如何整理、綜合及呈現統計數據，就是本節的主題。

13.2.1 用圖形呈現分佈

無論是普查或者是抽樣調查，最終的結果是一堆數據表。統計分析的初步就是如何呈現這數據表，讓人看了之後賞心悅目，對想要知道的結果也瞭然於胸，清楚看得到。想要達到這樣的效果，就必須將數據表轉換成容易觀察的「圓形圖」、「長條圖」、「直方圖」、「象形圖」或者是「線圖」等的統計圖形。

要畫圓形圖，先得畫個圓，圓代表全體，圓裡面的扇形就代表各部分，各扇形的圓心角（或扇形面積）和各部分的大小成比例。

例題 2 (圓形圖)甲、乙、丙三人合資創立一家小型電子公司，甲出資金 45 萬，乙出資金 30 萬，並言明以後的盈餘或虧損，就按照每人所出資金的比例分配。右圖是甲、乙、丙三人所出資金分佈的圓形圖，其中乙所在的扇形之圓心角是108°。



(1) 丙出資金多少元？丙所在的扇形之圓心角是幾度？

(2) 公司經營一年之後，丙分得盈餘 5 萬，甲分得的盈餘是多少元？

【解】(1)：設甲所在的扇形之圓心角為x°，由「甲出資金 45 萬，乙出資金 30 萬及乙所在的扇形之圓心角是108°」知道：

$$\frac{x}{45}=\frac{108}{30}\Rightarrow x=162.$$

因此丙所在的扇形之圓心角為$360^\circ-(108^\circ+162^\circ)=90^\circ$。設丙出資金$m$萬，利用比例得到

$$\frac{90}{m}=\frac{108}{30}\Rightarrow m=25.$$

故丙出資金 25 萬，丙所在的扇形之圓心角為90°。

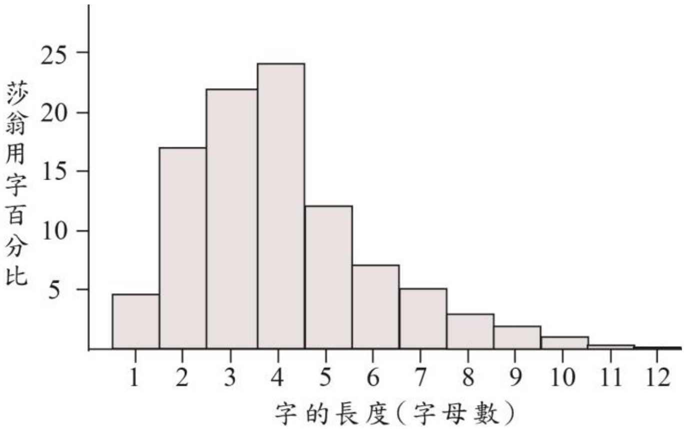
(2)：設甲分得的盈餘為r萬，利用比例得到

$$\frac{r}{45}=\frac{5}{25}\Rightarrow r=9.$$

故甲分得的盈餘為 9 萬。

〔註〕圓形圖的好處是讓我們看到：所有的部分合起來，的確是全體。但是扇形的圓心角比較不容易比較大小，所以圓形圖並不是比較各部分大小的好方法。

例題 3 (直方圖)下圖是莎士比亞的戲劇中用字長度的直方圖，圖中縱軸並不是數字，而是莎翁用字中各種長度所佔的百分比。



根據此直方圖，選出正確的選項：

(1) 莎翁最常用 4 個字母的字。

(2) 莎翁的戲劇中超過一半的字都是由 2 到 4 個字母所組成的字。

(3) 在莎翁的戲劇中，10 個字母以上（含 10 個字母）的字佔不到5%。

(4) 在莎翁的戲劇中，5 或 6 個字母的字超過四分之一。

(5) 在莎翁的戲劇中，4 個字母以上（含 4 個字母）的字，隨著字母數越長，出現的比例越低。

【解】分析如下：

(1) 因為 4 個字母的字在直方圖中所佔的比例最高，所以莎翁最常用 4 個字母的字。

(2) 由直方圖知道：2 個字母的字所佔的比例超過15%；3 個字母的字所佔的比例超過

20%；4 個字母的字所佔的比例超過 20%，所以 2 至 4 個字母的字所佔的比例超過 55%。因此莎翁的戲劇中超過一半的字都是由 2 到 4 個字母所組成的字。

(3) 由直方圖知道：10, 11, 12 個字母的字所佔的比例都沒超過 1.5%，所以 10 個字母以上（含 10 個字母）的字佔不到 5%。

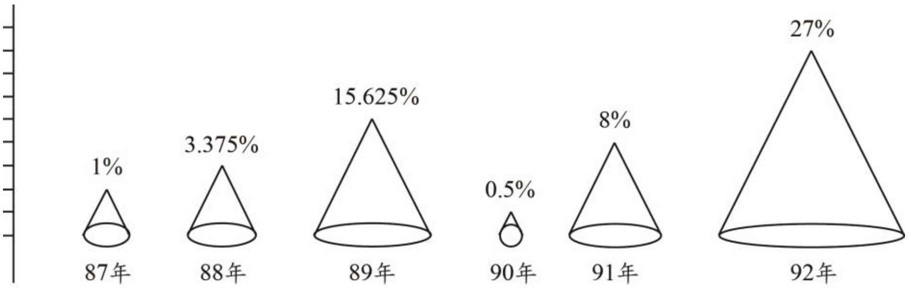
(4) 由直方圖知道：5 個字母的字不超過 15%；6 個字母的字不超過 10%，所以 5 或 6 個字母的字未超過四分之一。

(5) 由直方圖明顯看出遞減。

綜合得到正確答案為(1)(2)(3)(5)。

〔註〕從這則例題知道，用百分比或比例來當直方圖的縱軸要比用計數更方便。

例題 4 (象形圖)為了視覺上的美觀及宣傳效果，某公司員工將該公司過去六年（民國 87 年到 92 年）的投資報酬率，繪製成相似圓錐的形狀（即每個圓錐的半徑與高的比例都相同，如下圖所示），圓錐頂上的百分比就是該年度的投資報酬率。在統計學上，像這樣的圖稱為象形圖，而且要求象形圖中，數據（這裡是指投資報酬率）需與圖形的量（這裡是指圓錐的體積）成正比。為了瞭解這些象形圖是否符合統計學的圖表原理，在圖的左邊補上一條刻度尺供比對。如果以民國 87 年度的圖為標準，那麼民國 88 年到 92 年這五年度的象形圖中，哪幾個是正確的。



【解】觀察刻度尺，令每格刻度為 0.5 單位，民國 87 年度的圓錐高是 1 單位，令民國 88, 89, 90, 91, 92 年度的圓錐高分別是 h_1, h_2, h_3, h_4, h_5 單位。由圖形觀察得到

$$h_1 = 1.5, h_2 = 2.5, h_3 = 0.5, h_4 = 2, h_5 = 4.$$

由圓錐體積與投資報酬率的比例得到

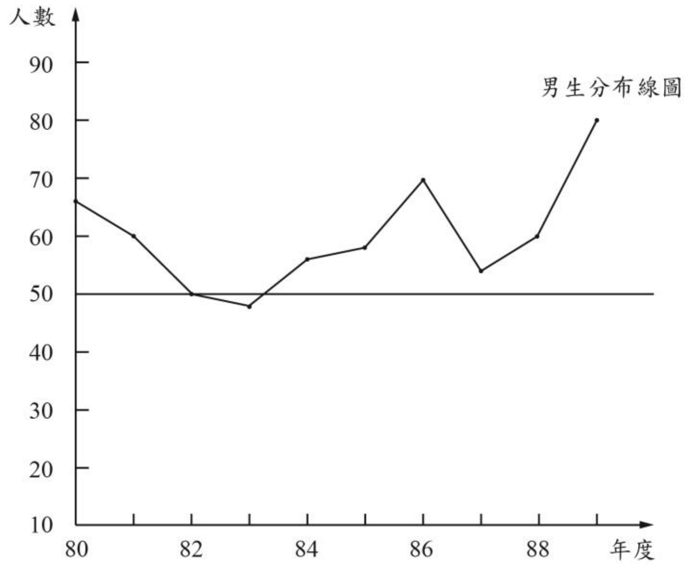
$$\frac{1^3}{1\%} = \frac{h_1^3}{3.375\%} \Rightarrow h_1 = 1.5;$$
$$\frac{1^3}{1\%} = \frac{h_2^3}{15.625\%} \Rightarrow h_2 = 2.5;$$
$$\frac{1^3}{1\%} = \frac{h_3^3}{0.5\%} \Rightarrow h_3 = \sqrt[3]{0.5} > 0.5;$$
$$\frac{1^3}{1\%} = \frac{h_4^3}{8\%} \Rightarrow h_4 = 2;$$
$$\frac{1^3}{1\%} = \frac{h_5^3}{27\%} \Rightarrow h_5 = 3 < 4.$$

故民國 88, 89, 91 年度的立體圖形之比例是正確的。

〔註〕象形圖是長條圖的變種。從這則例題得知：象形圖很吸引人，但容易被誤導。

許多數量變化都是隔一段時間量一次的。比方說，量成長中孩童的身高、體重，在每日紀錄某支股票的股價，或者每個月普通無鉛汽油的平均價格。在這類例子當中，我們只要感興趣的，是變數如何隨著時間變化。呈現這種統計數據最好的圖是所謂的線圖。線圖一定要把時間刻度放在你畫的圖的橫軸上，而把你正在度量的變數放在縱軸上。

例題 5　(線圖)某數學系從 80 至 89 學年度都錄取 100 位學生。下表是這十個年度的錄取人數中，男生人數的分布圖。



選出正確選項：

(1) 在這 10 年中，曾經有過女生人數低於 10 人。

(2) 有 2 年女生人數超過男生人數。

(3) 如果將女生分布線圖畫在同一個圖上，那麼女生分布線圖與男生分布線圖會對稱於高度是 50 人的那條水平線。

(4) 有 2 年女生人數與男生人數一樣多。

(5) 這 10 年的女生平均錄取人數超過 30 人，但不到 50 人。

(6) 如果令 $M(t), W(t)$ 分別代表第 $t(80 \le t \le 89)$ 學年度，男生與女生錄取的人數，那麼 $W(t) = 100 - M(t)$。

【解】分析如下：

(1) 因為男生與女生人數和是 100 人，又從男生分佈圖得知：男生至多是 80 人，所以女生至少有 20 人。故女生人數低於 10 人是不可能的。

(2) 男生人數必須小於 50 人，女生人數才會超過男生人數。從分佈圖得知：只有一年而已。

(3) 這是對的，因為男生與女生人數和是 100 人。

(4) 僅 82 年，男女一樣多。

(5) 這是正確的，因為男生人數（除 89 年外）都在 50 至 70 之間遊走，故女生人數介於 30 與 50 之間（加起來剛好 100）。

(6) 這是對的，因為因為男生與女生人數和是 100 人。

答案為(3)(5)(6)。

13.2.2 用數字描述分佈…眾數、中位數、四分位數、全距、四分位差、平均數與標準差

將抽樣調查或普查所得到的數據由小排到大作排序分佈圖：

$$x_1, x_2, x_3, \cdots, x_n,$$

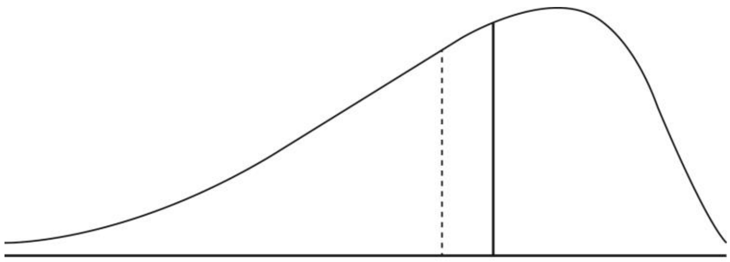
其中出現次數最多的數稱為眾數；最大數與最小數的差稱為全距（以符號 R 表示）；最中間的數或最中間兩數的平均數稱為中位數（以符號 M 表示）；第 1 四分位數（以符號 Q_1

表示)是前$\frac{n}{2}$個數據的中位數;第3四分位數(以符號Q_3表示)是後$\frac{n}{2}$個數據的中位數;四分位差(以符號Q表示)定為$Q=Q_3-Q_1$;所有數據的算術平均數稱為平均數(以符號μ表示);而標準差(以符號σ表示)則定為

$$\sigma=\begin{cases}\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}, & \text{(普查標準差公式);}\\ \sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}, & \text{(抽樣調查標準差公式).}\end{cases}$$

這些描述資料分佈的定義名稱可以分成兩大類:最小數,第1四分位數,中位數,平均數,第3四分位數,最大數是用來描述某些特定點的資料值(如中位數與平均數都是在描述資料的中心位置);而全距,四分位差,標準差則是在描述資料的離度(離散程度)。

例題 6 下圖是某優秀學區的國小學生智商分佈圖,鉛直黑線與鉛直虛線中,有一條是中位數所在的位置的鉛直線,而另一條是平均數所在的位置的鉛直線。運用你對於平均數與中位數的知識,判斷何者為中位數的鉛直線?哪條為平均數的鉛直線?



優秀學區的學生智商分佈圖

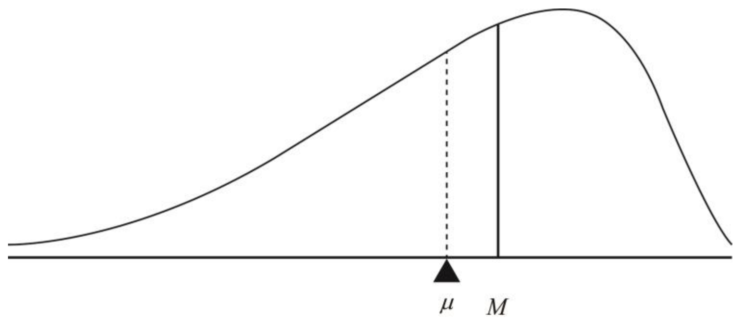
【解】首先分析一下這種曲線圖的中位數與平均數的位置:

(1) 中位數:曲線圖的中位數就是等面積點,也就是曲線圖底下的面積剛好被過該點的鉛直線平分。

(2) 平均數:因為平均數是它們的算術平均,所以當我們把數據想像成疊在蹺蹺板上的砝碼時,平均數就是蹺蹺板的平衡點。如果曲線圖對稱分佈,那麼平衡點剛好落在

正中央，也就是說平均數位在中點（此時中位數也在中點）。如果曲線圖偏斜分佈，那麼平衡點會落在被拉長尾巴方向（離蹺翹板越遠的數據會產生越大的力距，唯有平衡點向他們靠攏才能取得平衡），也就是說平均數位在被拉長尾巴方向。

利用上述說明不難發現實黑線是中位數的鉛直線，而虛線是平均數的鉛直線。



優秀學區的學生智商分佈圖

體育運動的裁判員都有一定的資歷和經驗，但無論如何，判斷上主觀的偏差還是很難避免的。因此有些體育競賽的計分法，就借用統計學上四分位數的觀念。

例題 7　某次溜冰比賽規定，在 7 名裁判員所給予的分數中，最高與最低的都會捨去，剩下的分數加起來取平均數，便是溜冰選手最後的成績。某溜冰選手所得到的 7 個分數，由小排到大依序為

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7.$$

選出正確的選項：

(1) 這 7 個分數的第 1 四分位數 Q_1 剛好等於 x_2。

(2) 這 7 個分數的第 3 四分位數 Q_3 剛好等於 x_6。

(3) 這 7 個分數的中位數 M 剛好等於 x_4。

(4) $Q_1 \leq$ 該溜冰選手最後的成績 $\leq Q_3$。

(5) 這 7 個分數的平均數 μ 與該溜冰選手最後的成績不相等。

【解】根據第 1 四分位數的定義，7 個數據的第 1 四分位數是前 $\frac{7}{2}=3.5$ 個（取前 3 個）數的中位數，所以第 1 四分位數應是 x_1,x_2,x_3 這三個數的中位數 $Q_1=x_2$。同理可得 $Q_3=x_6$。

中位數 $M=x_4$ 由定義直接得到。

根據分數的計算方法，該溜冰選手最後的成績為

$$\frac{x_2+x_3+x_4+x_5+x_6}{5}.$$

由不等式 $x_1\le x_2\le x_3\le x_4\le x_5\le x_6\le x_7$ 得到

$$Q_1=x_2\le\frac{x_2+x_3+x_4+x_5+x_6}{5}\le x_6=Q_3.$$

當被捨去的 x_1 與 x_7 搭配得很好時，這 7 個分數的平均數 μ 與該溜冰選手最後的成績是有可能相等的，故(5)是錯誤的。

綜合上述說明，正確答案為(1)(2)(3)(4)。

例題 8 根據主計處資料，台灣貧富差距倍數日趨懸殊，個人所得最高族群（前10%），其所得金額是最低族群（後10%）的 61 倍之多。一年多前，為了縮減貧富差距倍數及所得標準差，有學者建議政府採取底下幾個方案：

甲方案：每人每月減薪 5000 元；

乙方案：每人每月減薪10%；

丙方案：個人所得最低族群，每人每月發救濟金 3000 元。

下列有關這幾個方案的敘述，哪些是正確的：

(1) 實施甲方案，台灣貧富差距倍數會低於 61 倍。

(2) 實施乙方案，台灣貧富差距倍數還是 61 倍。

(3) 實施丙方案，台灣貧富差距倍數會低於 61 倍。

(4) 實施甲方案，台灣個人所得的標準差會縮小。

(5) 實施乙方案，台灣個人所得的標準差會縮小10%。

【解】設個人所得最低族群的人，每人每月所得m元，由題意知道：個人所得最高族群的人，每人每月所得為$61m$元。

(1) 根據甲方案，調薪之後的貧富差距倍數變為

$$\frac{61m-5000}{m-5000}>61.\ (\text{由假分數的性質得到})$$

(2) 根據乙方案，調薪之後的貧富差距倍數變為

$$\frac{61m\times 90\%}{m\times 90\%}=61.$$

(3) 根據丙方案，調薪之後的貧富差距倍數變為

$$\frac{61m}{m+3000}<61.$$

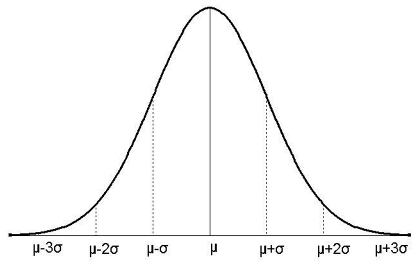
(4) 根據甲方案，只是將個人所得的數字平移一下而已，根據標準差的公式，「平移」會將平均數平移一樣的數，但不會改變標準差的值。故標準差應不變。

(5) 根據乙方案，將個人所得的數字「伸縮」為原來的九成，根據標準差的公式，平均數和標準差也都會變為原來的九成。

故正確答案為(2)(3)(5)。

13.2.3 用數學模型來描述分佈……常態分佈

常態分佈(一說<u>高斯</u>分佈)是由平均數與標準差這兩個量來決定的一類數學函數的圖形，圖形很像鐘形，故又名為鐘形曲線。右圖就是平均數為μ與標準差為σ的常態分佈。這種鐘形曲線有一很有用的$68-95-99.7$規則，大意是說：「鐘形曲線下方面積有68%集中在平均數正負 1 個標準差的範圍內（右圖中，$\mu-\sigma$與$\mu+\sigma$兩條虛線的區域面積佔整個鐘形曲線下方面積的68%）；有95%集中在平均數正負 2 個標準差的範圍內（右圖中，$\mu-2\sigma$與$\mu+2\sigma$ 兩條虛線的區域面積佔整個鐘形曲線下方面積的68%）；有99.7%集中在平均數正負 3 個標準差的範圍內（右圖中，$\mu-3\sigma$與$\mu+3\sigma$兩條虛線的區域面積佔整個鐘形曲線下方面積的99.7%）。」在統計學裡，很多資料的分佈都是常態分佈（如男性和女性的體重分佈，男性和女性的身高分佈，同一



年齡學童的 IQ 測驗分數分佈，懷孕期的天數分佈，傳染病的潛伏期天數分佈，輸卵管內精子數分佈等)，故鐘形曲線是很重要的統計模型，特別是它的$68-95-99.7$規則。

例題 9　某國中對該校 1000 位國一學生做 IQ 測驗，結果發現分數的分佈很接近平均數 111，標準差 11 的常態分佈。

(1) 分數超過 100 的約有幾人。

(2) 分數超過 111，而未達 133 的約有幾人。

(3) 某班 50 位學生中，沒人考超過 144 分，但是隔壁班有人超過 144 分，於是這班的一位家長打電話指責老師，認為是老師沒有教好學生的關係。你認為這種指責合理嗎？

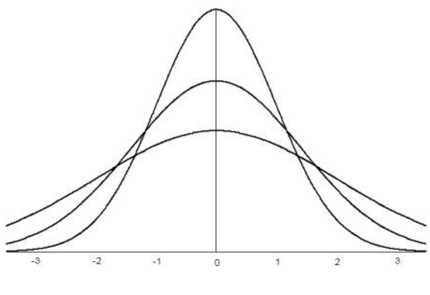
【解】分析如下：

(1) 100 剛好是負 1 個標準差。根據常態分佈知道：正負 1 個標準差以外的區域佔 $100\%-68\%=34\%$，所以 IQ 在$111+11=122$以上的佔$\dfrac{34}{2}\%=17\%$。故分數超過100的佔$(68+17)\%=85\%$，約$1000\times85\%=850$人。

(2) 分數超過 111，而未達 133 代表分數高於平均數，而不及正 2 個標準差，故此區域應佔$\dfrac{95}{2}\%=47.5\%$，約$1000\times47.5\%=475$人。

(3) 超過 144 分代表超過正 3 個標準差，根據法則，此區域佔全體學生的$\dfrac{100-99.7}{2}\%=0.15\%$，故人數為 50 人的班級應有$50\times0.15\%=0.075$人分數超過 144 分。0.075人離 1 人很遙遠，因此隔壁班是僥倖出現的，班上沒有人達到 144 分是正常的。故家長的指責比較沒道理(因為全校可能只有 1 到 2 個人可以超過 144 分)。

例題 10　右圖是三條平均數都是 0 的鐘形曲線，且每條鐘形曲線以下的面積都一樣。若它們的標準差由高度最高的鐘形曲線至最低的鐘形曲線依序為$\sigma_1,\sigma_2,\sigma_3$，則$\sigma_1,\sigma_2,\sigma_3$ 的大小關係為何？



【解】由常態分佈的特性，對稱軸左右一個標準差的面積佔全部面積（這裡都是 1）的 68%，知道三個分佈圖的對稱軸左右一個標準差內的面積都是 0.68。但是越高的分佈圖面積會越向對稱軸集中，故越高的分佈圖之標準差越小。也就是說

$$\sigma_1 < \sigma_2 < \sigma_3.$$

〔註〕也可以從圖形的反曲點之位置辨識，越高的分佈圖之反曲點離對稱軸越近。

13.2.4 散佈圖與相關係數

1965 年，考古學家於肯亞發現一塊手腕的化石骨頭，依其直觀（但專業）判斷，它若非黑猩猩，就一定是人類的骨頭。經用鈷六十照射鑑定後，確定它有 250 萬年之久。在此之前所發現最早的人類骨頭為 175 萬年，所以這塊骨頭若是人類的，那人類的存在史就往前推進數十萬年。要如何判別手腕骨頭的所屬物種呢？散佈圖與相關係數提供我們很好的方法。底下就以始祖鳥的化石來做說明！

始祖鳥是一種已滅絕的動物，牠有像鳥類一樣的羽毛，但是也有像爬蟲類的牙齒及長而多骨的尾巴。已知的化石標本只有 6 件，因為這些標本的大小差異很大，有些科學家認為這些標本可能是不同的種類，而不是同一種類的不同個體。在 5 件仍同時保有股骨（一種腿骨）及肱骨（上臂的骨頭）的標本中，我們檢查股骨及肱骨的長度，以下就是這組資料（單位：公分）：

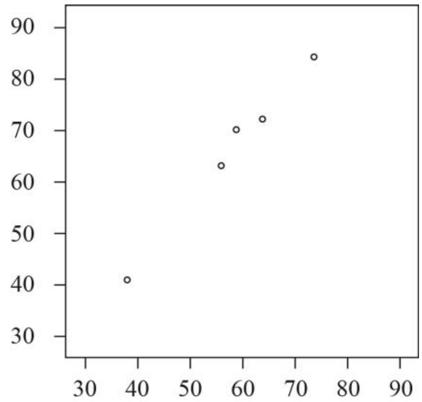
股骨	38	56	59	64	74
肱骨	41	63	70	72	84

統計學裡的散佈圖與相關係數剛好提供我們，認定這些始祖鳥化石是否出自同一種類的好方法！散佈圖是將兩個變量的變化數對繪入坐標圖中，以清晰地表明其分佈情形。細察散佈圖上各點的分佈情形，或許會對兩個變量間的相關情形有初步的瞭解。

例題 11　繪製 5 隻始祖鳥化石的股骨及肱骨散佈圖，並計算他們的相關係數。

【解】繪製散佈圖：圖中的 x 坐標表示「股骨長度（公分）」，y 坐標表示「肱骨長

度（公分）」



計算相關係數：散佈圖中，五個點的座標為

$$(38,41), (56,63), (59,70), (64,72), (74,84).$$

這五個點的 x 與 y 座標之平均值 $\overline{x},\overline{y}$ 分別為

$$\overline{x}=\frac{38+56+59+64+74}{5}=58.2;\ \ \overline{y}=\frac{41+63+70+72+84}{5}=66.$$

根據相關係數的公式得到相關係數 r 為

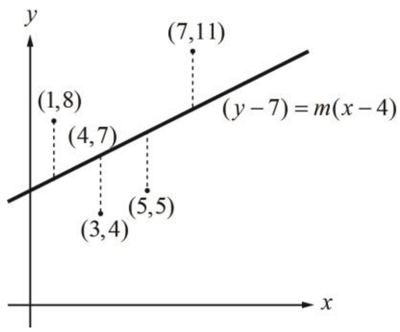
$$\begin{aligned} r &= \frac{\sum_{i=1}^{5}(\overline{x}-x_i)(\overline{y}-y_i)}{\sqrt{\sum_{i=1}^{5}(\overline{x}-x_i)^2}\sqrt{\sum_{i=1}^{5}(\overline{y}-y_i)^2}} \\ &= \frac{(20.2)(25)+(2.2)(3)+(-0.8)(-4)+(-5.8)(-6)+(-15.8)(-18)}{\sqrt{696.8}\sqrt{1010}} \\ &= \frac{834.6}{\sqrt{696.8}\sqrt{1010}} \\ &\doteqdot 0.995. \end{aligned}$$

13.2.5 迴歸直線

四塊鳥類的化石的兩個變量 x 與 y 的測量值如下：

x	1	3	5	7
y	8	4	5	11

其中 x 變量的平均值 $\overline{x}=4$；y 變量的平均值 $\overline{y}=7$。將這兩個變量的散佈圖繪製如下：



為了進一步的瞭解散佈圖，希望在散佈圖上畫一條通過數據點的中心$(\overline{x},\overline{y})=(4,7)$之直線

$$(y-7)=m(x-4),$$

並要求四個數據點(1,8), (3,4), (5,5), (7,11)至此直線的鉛直方向上的距離之平方和最小。滿足這樣的直線稱為此散佈圖的迴歸直線，迴歸直線的方程式稱為迴歸方程式。

現在讓我們來算這個散佈圖的迴歸直線：令$d(m)$是四個數據點

$$(1,8),\ (3,4),\ (5,5),\ (7,11)$$

至直線$(y-7)=m(x-4)$的鉛直方向上的距離之平方和，由距離公式得到

$$\begin{aligned}d(m)&=\big(m(1-4)+7-8\big)^2+\big(m(3-4)+7-4\big)^2\\&\quad+\big(m(5-4)+7-5\big)^2+\big(m(7-4)+7-11\big)^2\\&=(-3m-1)^2+(-m+3)^2+(m+2)^2+(3m-4)^2\\&=20m^2-20m+30\\&=20\left(m-\frac{1}{2}\right)^2+25.\end{aligned}$$

因此該散佈圖的迴歸方程式為

$$(y-7)=\frac{1}{2}(x-4).$$

例題 12　承上文，科學家經過仔細的鑑定之後，發現這四塊化石是始於同一時代，同種鳥類的化石。試問這種鳥類的x值為 12 時，其y值約為多少？

【解】將$x=12$代入散佈圖的迴歸方程式得到

$$(y-7)=\frac{1}{2}(12-4)\Rightarrow y=13.$$

因此y值為 13。

13.3 統計推論之一：機率與期望值

13.3.1 統計與機率

例題 13　電腦工程師設計一種玩猜拳的軟體，並要求電腦出剪刀、石頭與布的機率各為$\frac{1}{3}$。

(1) 如果你是習慣上出剪刀、石頭與布的分佈為2:3:5的人，那麼跟電腦軟體玩猜拳遊戲，第一拳你就獲勝的機率有多高？

(2) 如果你是習慣上出剪刀、石頭與布的機率為$\alpha,\beta,$ 與γ的人，那麼跟電腦軟體玩猜拳遊戲，第一拳你就獲勝的機率有多高？（以α,β,γ表示，並盡可能的化簡）

【解】(1)：第一拳出剪刀贏的機率為

$$\frac{2}{2+3+5}\times\frac{1}{3}=\frac{2}{10}\cdot\frac{1}{3};$$

第一拳出石頭贏的機率為

$$\frac{3}{2+3+5}\times\frac{1}{3}=\frac{3}{10}\cdot\frac{1}{3};$$

第一拳出布贏的機率為

$$\frac{5}{2+3+5}\times\frac{1}{3}=\frac{5}{10}\cdot\frac{1}{3}.$$

故第一拳贏的機率為

$$\left(\frac{2}{10}+\frac{3}{10}+\frac{5}{10}\right)\times\frac{1}{3}=\frac{1}{3}.$$

(2)：由$\alpha+\beta+\gamma=1$得到：第一拳出剪刀贏的機率為

$$\frac{\alpha}{\alpha+\beta+\gamma}\times\frac{1}{3}=\frac{\alpha}{3};$$

第一拳出石頭贏的機率為

$$\frac{\beta}{\alpha+\beta+\gamma}\times\frac{1}{3}=\frac{\beta}{3};$$

第一拳出布贏的機率為

$$\frac{\gamma}{\alpha+\beta+\gamma}\times\frac{1}{3}=\frac{\gamma}{3}.$$

故第一拳贏的機率為

$$\frac{\alpha}{3}+\frac{\beta}{3}+\frac{\gamma}{3}=\frac{1}{3}.$$

13.3.2 期望值

期望值是所有可能結果的加權平均，每個結果所對應的權數是該結果的機率。如果你不知道各結果的機率，你可以利用模擬來估計期望值（以及各結果的機率）。歷史上最有名的例子就是求圓周率π的近似值。在邊長為 1 的正方形內，畫它的內切圓，此內切圓的圓心剛好是正方形的中心，而且圓直徑恰為 1。統計學家將這圖形貼在牆壁上，並拿飛鏢對這圖形隨意瞄準。因為正方形的面積為 1，而內切圓的面積為$\pi\left(\frac{1}{2}\right)^2=\frac{\pi}{4}$，所以直觀上告訴我們

$$\frac{\text{飛鏢射中內切圓區域的次數}}{\text{飛鏢射中正方形區域的次數}}=\frac{\text{內切圓面積}}{\text{正方形面積}}=\frac{\pi}{4}.$$

有了上述等式，現在是進行統計實驗的時候了，如果你對正方形隨意發射 1000 次的射擊，統計得知有 786 次射中內切圓區域，則圓周率π的近似值為

$$\frac{786}{1000}\doteqdot\frac{\pi}{4}\Rightarrow\pi\doteqdot 3.144.$$

例題 14　十五、六世紀的加勒比海是海盜的天堂。因為海盜必須經年征戰，四肢容易殘缺，所以想一圓海盜夢的人，常常需付出慘痛的代價。

當時無征戰能力的海盜成立了醫療保險制度。只要是繳交 12 披索保險費的海盜，就終生享有如下的殘缺給付：傷殘成獨眼海盜給付 100 披索；傷殘成單臂海盜給付 500 披索；傷殘成跛腳海盜給付 400 披索，死亡或是其它殘缺不給予給付，而且每人至多僅給付一次。

根據調查，海盜終其一生，傷殘成獨眼海盜、單臂海盜及跛腳海盜的機會是千分之十八、十及八。試問：這種醫療保險制度預期可以從每位投保的海盜身上賺取____披索。

【解】由期望值的定義知道：這種醫療保險制度預期可以從每位投保的海盜身上賺取

$$
\begin{aligned}
&12-\left(100\times\frac{18}{1000}+500\times\frac{10}{1000}+400\times\frac{8}{1000}\right)\\
&=12-\frac{1800+5000+3200}{1000}\\
&=2
\end{aligned}
$$

披索。

13.4 統計推論之二：民意調查與平均數的統計推論

（中央社舊金山二００三年四月二十八日專電）根據美國廣播公司新聞部的一項最新民意調查顯示，僅有三成三的美國民眾擔心自己或家人會染上嚴重急性呼吸道症候群（SARS）。美國廣播公司新聞部是在四月二十三日到二十七日之間，以隨機抽樣的方式對全美各地一千零一十八名成人有效樣本進行電話調查，並宣稱調查結果的誤差在正負三個百分點以下。像這樣的統計抽樣調查報導，在報紙或電視上經常看見或聽到，這樣的文字報導是否給了讀者充分的資訊，又我們該如何解讀這樣的報導呢？

統計是面對不確定的情形下,能夠幫助我們做出明智抉擇的一種科學方法。既然是抉擇，那代表統計所得出的推論也有不準確的風險存在。為了規避風險或者避免統計推論被誤解或濫用，統計推論的講法受到一定的約束的，而這約束是根據一些數學定律得來的。雖然沒辦法在這本書裡講解這些數學定律，但是處在資訊發達，天天發佈許多民調，接收統計數據的世界裡，身為世界公民的一份子，瞭解民調與解讀統計推論的意義是相當重要的。在此，我們把常見的兩種統計推論（民意調查與平均數的統計推論）的數學內涵做個簡單的描述，供讀者參考。常態分佈，大數法則與中央極限定理是形成這兩個統計推論的數學推手。

最後引用愛因斯坦的名言：「數學定律不能百分之百確實的用在現實生活裡：能百分之百確實的用數學定律描述的，就不是現實生活。」既然統計推論是由一些數學定律建立起來的，這推論當然就有失算的時候！

13.4.1 有關民意調查的統計推論

報紙上報導一項民意調查的結果說:「蓋洛普抽查n個人的意見後,調查出$p\%$的美國人,在過去一年當中,曾購買樂透彩券。對於用這樣大小的樣本所得到的結果來說,我們可以有95%的信心,由於抽樣或其他隨機因素所導致之誤差,應在正負q個百分點以內。」像這樣的民意調查報告幾乎天天都有,有的是調查候選人的支持度,對政府施政的滿意度,估計失業率等不一而足。以蓋洛普的這則統計結論來說,它是指:「如果對這樣的樣本(抽查n個人)進行100次,應該有95次所得到的比率會落在

$$(p-q)\%\sim(p+q)\%$$

的範圍內,但並不代表全美國人,在過去一年當中,曾購買樂透彩券的比率一定在這個範圍內。」

既然統計是一們嚴謹的科學,那麼民意調查中的數字n, p, q是否應該滿足數學上的方程式呢?事實上,他們經常滿足如下的方程式

$$2\frac{\sqrt{p\%(1-p\%)}}{\sqrt{n}}=q\%,$$

也就是滿足

$$2\frac{\sqrt{p(100-p)}}{\sqrt{n}}=q.$$

這就是一般民意調查所必須滿足的數學通式。

例題 15　根據「蓋洛普調查出60%的美國人,在過去一年當中,曾購買樂透彩券。對於這次的調查,蓋洛普有95%的信心認為,其誤差應在正負4個百分點以內。」這則民調,估計蓋洛普抽查的樣本是多大。

【解】設抽查的樣本為n人,由民意調查的數學通式得到

$$2\frac{\sqrt{60(100-60)}}{\sqrt{n}}=4\Rightarrow n=600.$$

故蓋洛普抽查的樣本是600人。

如果信賴指數95%改成其它的百分比,那麼民意調查的數學通式也會跟著改變。例如下

表所列：

信賴指數	數學通式
80%	$1.28\frac{\sqrt{p(100-p)}}{\sqrt{n}}=q$
90%	$1.64\frac{\sqrt{p(100-p)}}{\sqrt{n}}=q$
95%	$1.96\frac{\sqrt{p(100-p)}}{\sqrt{n}}=q$
99%	$2.58\frac{\sqrt{p(100-p)}}{\sqrt{n}}=q$

從這個表格知道四件事情：

(1) 信賴指數為95%時，參數是 1.96，而非 2，文中取 2 是為了方便計算與記憶。

(2) 如果讓信賴指數越高，則需付出誤差範圍擴大的危險。

(3) 在同樣的信賴指數之下，抽樣樣本放大 4 倍，誤差範圍減半。這是因為大樣本越能反應實際情況的緣故。

(4) 因為 n, p, q 需滿足方程式，所以民調報告時，經常將抽取樣本數 n 或誤差範圍 q% 擇一省略不說（讀者需懂得民意調查的數學通式才能估出省略的數字是多少）。

例題 16 統計學教授為了驗證一枚羅馬時代的銅板是否為均勻的銅板，做了若干次的投擲實驗。實驗後，教授做了如下的統計推論：「對這次抽投擲實驗的結果來說，我有 95% 的信心認為“投擲這枚銅板有 42% 至 48% 的機率會出現正面。”」

試推測在那次實驗中，教授投擲銅板幾次，出現幾次正面。

【解】設投擲銅板 n 次，因為 42% 至 48% 的機率可以表為 45%±3%，所以機率為 45%，而誤差範圍是 3 個百分點。由數學通式得到

$$2\frac{\sqrt{45(100-45)}}{\sqrt{n}}=3\Rightarrow n=1100.$$

故在那次實驗中，教授投擲銅板 1100 次，出現

$$1100\times 45\%=495$$

次正面。

13.4.2 有關平均數的統計推論

在前一小節裡，我們談到有關民意調查的統計推論，特別是民意調查的數學通式。如果是調查每戶有多少輛機車，國人平均壽命，大學生每天平均花幾個小時在課本上，人體的平均溫度，國人年平均所得等這些有關平均數的調查，那麼統計推論該如何下呢？是否也存在著平均數的數學通式呢？這答案是肯定的，底下就是有關平均數的一則統計推論：衛生單位隨機抽取 n 位 40 到 50 歲的男性，測量他們的血壓，得收縮壓的平均數 μ，標準差 σ，於是提出報告說：「我們有 95% 的信心認為“國人男性的平均收縮壓為 μ”，而且誤差範圍應該在正負 q 個百分點以內。」有關平均數的數學通式為

$$2\frac{\sigma}{\sqrt{n}}=q\%.$$

例題 17　衛生單提出報告說：「在抽樣 400 位 40 到 50 歲的男性，測量他們的血壓後，我們有 95% 的信心認為：國人男性的收縮壓介於 110 與 140 之間。」試問：這 400 位民眾的收縮壓之平均數與標準差是多少？

【解】由 110 與 140 的平均得到

$$\mu=\frac{110+140}{2}=125.$$

又

$$q\%=\frac{140-125}{125}=6\%\Rightarrow 2\frac{\sigma}{\sqrt{400}}=6\%,$$

即 $\sigma=12$。故這 400 位民眾的收縮壓之平均數為 125，而標準差為 15。

有關平均數的信賴指數與數學通式的關係列表如下：

信賴指數	數學通式
80%	$1.28\frac{\sigma}{\sqrt{n}}=q\%$
90%	$1.64\frac{\sigma}{\sqrt{n}}=q\%$
95%	$1.96\frac{\sigma}{\sqrt{n}}=q\%$
99%	$2.58\frac{\sigma}{\sqrt{n}}=q\%$

習題 1 隨機號碼表（一說亂數表）是一連串的 0,1,2,3,4,5,6,7,8,9 這些數字，且滿足以下兩個條件：

① 表中任一位置的數字，其為 0 至 9 中任何一個數字的機率相同。

② 不同位置的數字之間是獨立的。也就是說，知道表中某一部份是些什麼數字，不會提供你任何關於其他部分是些什麼數字的訊息。

請選出正確的選項：

(1) 像 0123456789 這樣的連續數字不可能出現在隨機號碼表上，因為容易被猜中。

(2) 如果把 0, 2, 4, 6, 8 當正面，1, 3, 5, 7, 9 當反面，那麼隨機號碼表可以用來模擬連續投擲一枚公正硬幣的情形。

(3) 隨機號碼表上不可能出現連續 99 個 0，因為容易被猜中。

(4) 隨機號碼表上應該奇偶數字相間出現，這樣才容易均勻，不被猜中。

(5) 像 119 或 110 這樣易被猜中的連續數字不會在隨機號碼表上出現。

習題 2 國際面板行銷研究與諮詢機構指出，2003 年液晶電視（LCD）的市佔率 2.2%（400 萬台），而電漿電視（PDP）市佔率為 3.3%。2003 年全球液晶電視市場龍頭為日本夏普，市場佔有率高達 46%；其次為南韓三星電子，市佔率 16%，至於第三、四名則分別為日本國際的 8%、日本新力索尼的 7%。試利用這則報導的數據，選出正確的選項：

(1) 2003 年電漿電視銷售量約為 600 萬台。

(2) 2003 年各種電視的銷售量約為 1 億 8182 萬台。

(3) 2003 年南韓三星電子售出的液晶電視數量約為日本國際的兩倍。

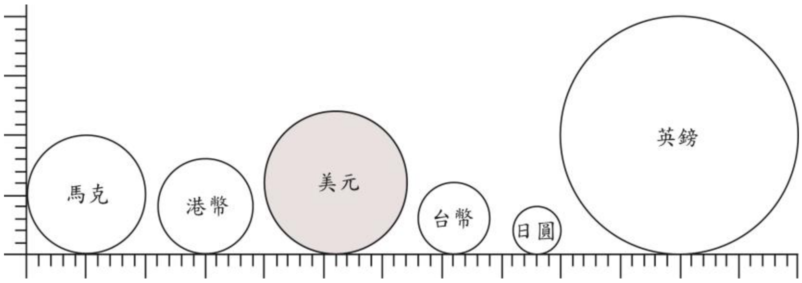
(4) 2003 年日本夏普售出約 184 萬台液晶電視。

(5) 2003 年全球液晶電視佔有率排名第五名（含）以後的總銷售量超過 100 萬台。

習題 3 銀行的匯率表是提供各國貨幣價值的一種比較簡表，最常用的是以美元為標準，下表列出每 1 美元可以兌換英鎊、馬克、港幣、台幣與日圓五種幣值的兌換表：

匯率表	英鎊	馬克	港幣	台幣	日圓
美元	0.64	1.44	9	36	144

某銀行為了凸顯不同貨幣的價值，改採如下的圓盤象形圖來彰顯幣值的大小。為了合理的反應幣值大小，幣值所代表的圓盤面積與該種幣值一塊錢的實際價值成正比。



如果以美元的圓盤當基準，試問英鎊、馬克、港幣、台幣與日圓這五種幣值所呈現的圓盤，哪幾個是合理的？

習題 4 鄭教授的統計學課有 40 位學生選修，期中考成績不是很理想。鄭教授提出兩個調整分數的方案供學生表決，在表決之前，鄭教授補充說：「這次考試，每個人的成績都是整數，而且無論選那個方案，每個人調整後的分數都不會低於原來的分數。」

$$甲方案：調整分數=\frac{考試分數}{2}+40;$$

$$乙方案：調整分數=\frac{考試分數}{4}+55.$$

在每人至多選一個方案的規定下，表決結果是 18 位選甲方案；19 位選乙方案。已知每個人都選對他最有利的調整分數方案，至於未舉手的 3 人，因為兩個方案對他們來說，最後分數都一樣，因此沒舉手。

問題一：已知大學裡的考試，60 分（含）以上算及格，請選出正確的選項：

(1) 這次考試，沒有人超過 80 分。

(2) 這次考試，沒有人超過 73 分。

(3) 這次考試，剛好及格的有 3 位。

(4) 這次考試，及格的有 22 位。

(5) 這次考試，不及格的有 18 位。

向來快人快語的鄭教授接著說：「那我們就用乙方案來調整各位的成績，不過調整之後，全班的平均分數 66，標準差 2.5，且仍有 5 位同學不及格，希望不及格的下次加油。」

問題二：請選出正確的選項：

(1) 學生實際成績的平均數為 44 分。

(2) 學生實際成績的標準差為 10 分。

(3) 有 5 位學生的實際成績低於 20 分。

(4) 學生實際成績的中位數恰為 60 分。

(5) 有 14 位學生的實際成績不及格，但是超過 19 分。

習題 5 體操比賽規定，在 7 名裁判員所給予的分數中，最高與最低的都會捨去，剩下的分數加起來取平均數，便是體操選手最後的成績。體操選手涅莫夫在完成體操動作之後，電視螢幕上出現了 7 個分數，因為畫面停留時間不夠長，只記得頭 6 位裁判員給的分數為

$$9.8, 9.6, 9.7, 9.8, 9.9, 9.6.$$

但是下一個電視畫面顯示涅莫夫在這項體操的最後成績為 9.74。求第 7 位裁判員給涅莫夫的分數是幾分。

習題 6 某公司將個人薪資所得高低分成五等分，最高 20% 的平均薪資所得是最低 20% 的平均薪資所得的 3 倍。為了因應不景氣及縮小所得的倍數差距，將薪資最高 20% 的員工，每人每月減薪 1.6 萬元（經此調整後，他們的薪資仍是前 20%），薪資最低 20% 的員工，每人每月減薪 0.4 萬元。經此調薪之後，薪資最高 20% 的平均薪資所得是薪資最低 20% 的平均薪資所得的 2.8 倍。求原來薪資最高 20% 及最低 20% 的員工之平均月薪資是多少錢？

習題 7 一份調查報告是這樣寫的：「如將家庭依可支配所得高低分成五等分，去年最高 20% 家庭平均可支配所得是最低 20% 家庭平均可支配所得的 6.4 倍，足足比他們多出 151 萬 2000 元。」求最高 20% 與最低 20% 家庭平均可支配所得分別是多少錢？

習題 8 教授統計學的蔡教授擔任數學系二年級的導師，該班有 35 位學生。照往例，蔡教授都在剛開學及學期末請學生喝汽水。學期快結束了，蔡教授翻閱筆記本上的紀錄得知，剛開學時，班上 35 位學生喝的汽水罐數之中位數 3 罐，平均數 4 罐，標準差 1.6 罐。

(1) 蔡教授的太太認為可以買 35×3 罐汽水，而蔡教授打算買 35×4 罐汽水。問誰的想法較妥當？又為什麼？

(2) 在剛開學的那次活動中，是否有學生喝超過 15 罐汽水。

習題 9 某針對國中一年級學生實施的智力測驗成績分佈為平均數 100，標準差 15 之常態分配，小明就讀國中一年級接受此測驗，成績為 115，請問他大約贏過多少比例的同年級學生？

習題 10 人類從受孕到分娩的懷孕期，長短各有不同，但大至遵循平均數 266 天，標準差 16 天的常態分佈。問：約有多少比例的人，會在 234 天內就分娩。

習題 11 根據數學 SAT 考試規定，該項測驗的總分有可能超過 800 分，但是超過的人的成績單一律只記錄為 800 分而已（也就是說學生拿到 800 分的成績單，不代表他只考 800 分而已）。某一年，男性在數學 SAT 考試的實際分數呈常態分佈，平均數 560，標準差 120。問：約有多少比例的男性會收到 800 分的成績單。

習題 12 汽車每加侖汽油跑的英里數在速度增加時先會上升再下降。下表就是測試的速

度（每小時英里數）和汽油里程（每加侖汽油跑的英里數）的資料：

速度	20	30	40	50	60
汽油里程	22	28	30	28	22

(1) 求速度和汽油里程之間的相關係數。

(2) 將速度設為 x 坐標，汽油里程為 y 坐標。求通過資料中，那五個點的二次函數。

(3) 利用(2)的二次函數推論一下，當汽車速度為每小時 70 英里時，每加侖汽油能跑多少英里？

習題 13 在蜀地(今四川)發現五個陪葬的銅製人體模型，他們分別在不同的地點發現，而且明顯的身體長度遠大於手臂長度。科學家為了判定這五個銅製人體模型是否屬於同一文化，分別測量紀錄他們的身體長度（單位：公分）和手臂長度（單位：公分），如下表所示：

身體長度	37	43	51	59	60
手臂長度	5	6	9	13	17

(1) 繪製這五個銅製人體模型的身體長度和手臂長度(以公分為單位)的散佈圖。

(2) 求這五個銅製人體模型的身體長度和手臂長度的相關係數。

(3) 你認為這五個銅製人體模型出自同一文化嗎？

(4) 有一奇怪的科學家，堅持身體長度和手臂長度皆要以公尺為單位。問：這科學家得到的相關係數為何？

習題 14 人壽保險公司的內部資料顯示，21 歲的男性在次年會死亡的機率大約是 0.0015。於是公司開發一種給 21 歲的男性投保的壽險，契約明訂：「在一年內死亡可後理賠 100000 美元，而保費僅需 200 元。」試問：每張這種保單的期望獲利是多少？

習題 15 莫理斯的腎臟不行了，他在等待腎臟移植。莫理斯希望知道，他能活過五年的

機率是多少？於是上網查詢和他狀況類似的病人資料。他發現：撐過移植手術的占 90%，另外 10% 會死亡。在手術後存活的人中有 60% 移植成功，另外的 40% 還是得回去洗腎。五年後存活率對於有新腎的人來說是 70%，對於回去洗腎的人來說是 50%。

問：莫理斯活過五年的機率是多少？

習題 16 某民調機構發佈一則民調是這樣說的：「我們有 95% 的信心認為：會有 62.5% 到 67.5% 的選民在下週的選舉中，投票給甲候選人。」問：他們約抽樣幾個選民當樣本。

習題 17 雖然智商測試並非完美，但是這種工具已經發展了九十九年，經過無數校訂及調節，是評估個人工作能力的指標。在一項全球智商研究發現：我們有 95% 的信心認為環太平洋地區國家具有最高智商，平均智商大約介於 103.8 點與 106.2 點之間，這是抽查 900 個樣本所得到的結果。問：這 900 個樣本的標準差約為多少？

習題 18 二次大戰之後的德國，經濟蕭條，民生凋弊，糧食實行配給制度，其中最重要的麵包每人每天只能分到 200 公克。麵包師都得用特製的模型烘製 200 公克重的麵包，以供應附近居民之所需。

一位物理學的老教授，每天早上都到麵包店領取他的配額。有一天，他對麵包師說：「你這個壞蛋，你的模型烤出來的麵包比法定的少 3 公克，你欺騙政府、顧客，把省下來麵粉都賣給黑市賺錢。」麵包師立刻為自己辯解說：「但是本來就不可能烘製出同樣大小的麵包呀！總是會有些小一點，有些大一點。」教授回答：「是沒錯，過去 25 天來，我把你給我的麵包都用我實驗室的天平秤過了，它們確實有自然的變異。但你看這張計算表，過去 25 天，平均起來卻是 197 公克，而不是規定的 200 公克，而且這 25 個數據的標準差高達 10，顯然

太離譜了。你要是再不改用正確的模型，我就要向當局告發你。」

你會為物理學教授鼓掌叫好？還是替麵包師傅叫屈呢？(本題修改自白啟光教授的數學嘉年華網站)

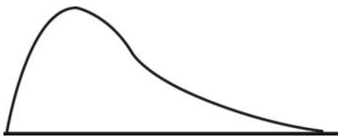
統計思考

美國高中會宣布他們入學新生的「平均」分數，而通常每所學校都希望這個「平均」越高越好。根據統計：「用獎學金來大量收買頂尖學生的私立學校喜歡用平均數；而誰都可以申請入學的公立學校喜歡用中位數。」請根據下列五個曲線圖(水平代表學生分數，曲線高度是學生人數)，選出最能代表私立學校的圖及最能反應公立學校的分佈圖。

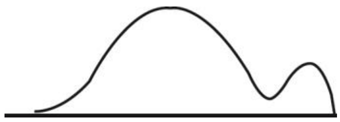
(1) 像「鐘形」的對稱分佈：

(2) 向左（往負向）偏斜分佈：

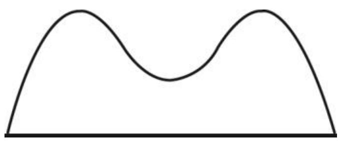
(3) 向右（往正向）偏斜分佈：



(4) 不對稱的雙峰分佈：



(5) 對稱的雙峰分佈：



運用你對於平均數和中位數的知識，來說明為什麼私校與公立學校會各自有如此之偏好。

參考文獻

[1] 黃提源，《機率與統計入門》，協進圖書有限公司。

[2] 墨爾著（鄭惟厚譯），《統計，讓數字說話》，天下文化。

[3] 墨爾著（鄭惟厚譯），《統計學的世界》，天下文化。

[4] 薩爾斯伯格著（葉偉文譯），《統計改變了世界》，天下文化。

[5] 羅浩源，《生活的數學》，九章出版社。

讓數字說話……統計的習題解答

習題 1

(1)(3)(4)(5)是不正確的，因為隨機號碼表上的每個數字是獨立不相關的，所以這些情形都有可能發生。而(2)是正確的，隨機號碼表就是設計來當公正的抽樣機器。

答案為(2)。

習題 2

分析如下：

(1) 因為電漿電視市佔率為3.3%是液晶電視市佔率2.2%的1.5倍，所以電漿電視銷售量約為$400\times1.5=600$萬台。

(2) 因為液晶電視400萬台佔各種電視的2.2%，所以各種電視的銷售量應為 $400\div(2.2\%)=1$億8182萬台。

(3) 因為南韓三星電子市佔率16%，而日本國際8%，所以南韓三星電子售出的液晶電視數量約為日本國際的兩倍。

(4) 因為日本夏普液晶電視市場佔有率為46%，所以日本夏普售出約$400\times46\%=184$萬台液晶電視。

(5) 因為全球液晶電視佔有率排名第五名（含）以後的總佔有率為 $100\%-(46+16+8+7)\%=23\%$，所以其銷售總量約為$400\times23\%=92$萬台。

綜合這些說明，正確答案為(1)(2)(3)(4)。

習題 3

利用輔助刻度得到：美元、英鎊、馬克、港幣、台幣與日圓的半徑分別為

$$6, 10, 5, 4, 3, 2.$$

故美元、英鎊、馬克、港幣、台幣與日圓每一塊錢的幣值比應為

$$6^2:10^2:5^2:4^2:3^2:2^2=36:100:25:16:9:4.$$

即每一美元可以兌換英鎊、馬克、港幣、台幣與日圓的錢數分別為

$$\frac{36}{100}=0.36,\frac{36}{25}=1.44,\frac{36}{16}=2.25,\frac{36}{9}=4,\frac{36}{4}=9.$$

這與匯率表相比較，僅與兌換馬克的數值相符。故僅有馬克所代表的圓盤是正確的，其餘都是不正確的。

習題 4

問題一：設 x 是考試的實際分數，x_1 代表經甲方案調整後的分數；x_2 代表經乙方案調整後的分數，由定義知道

$$x_1=\frac{x}{2}+40;\ \ x_2=\frac{x}{4}+55.$$

因為甲、乙兩方案都使得調整後的分數不會低於實際分數，所以

$$\begin{cases} x_1 \ge x; \\ x_2 \ge x \end{cases} \Rightarrow \begin{cases} \dfrac{x}{2}+40 \ge x; \\ \dfrac{x}{4}+55 \ge x \end{cases}$$
$$\Rightarrow \begin{cases} x \le 80; \\ x \le \dfrac{220}{3}=73\dfrac{1}{3} \end{cases}$$
$$\Rightarrow x \le 73 \ (\text{因為 } x \text{ 是整數}).$$

故選項(1)與(2)是正確的。

由題意知道：有 3 位學生的實際成績 x 滿足

$$\frac{x}{2}+40=\frac{x}{4}+55 \Rightarrow x=60.$$

由此得知：未舉手的 3 位學生的實際成績都是 60 分，故選項(3)是正確的。

再由題意知道：有 19 位學生的實際成績 x 滿足

$$x_1=\frac{x}{2}+40<\frac{x}{4}+55=x_2 \Rightarrow x<60.$$

同法得到：有 18 位學生的實際成績 x 滿足 $x>60$。故不及格的有 19 位，及格的有

$18+3=22$ 位，選項(4)(5)是不正確的。

綜合這些討論，正確的選項是(1)(2)(3)。

問題二：設學生實際成績的平均數 μ，標準差 σ。由平均數與標準差的公式知道：採乙方案調整之後的平均數為 $\dfrac{\mu}{4}+55$ 與標準差 $\dfrac{\sigma}{4}$。依題意得到

$$\begin{cases}\dfrac{\mu}{4}+55=66; \\ \dfrac{\sigma}{4}=2.5\end{cases} \Rightarrow \begin{cases}\mu=44; \\ \sigma=10.\end{cases}$$

故選項(1)與(2)是正確的。依題意，有 5 位學生的實際分數 x 滿足 $\dfrac{x}{4}+55<60$，即 $x<20$，故選項(3)是正確的。將 40 位學生的實際成績由低到高排列，第 20, 21, 22 位學生的成績都是 60 分，故學生實際成績的中位數為 60 分，即選項(4)是正確的。因為實際分數不及格的有 19 位，低於 20 分的有 5 位，所以有 14 位學生的實際成績不及格，但是超過 19 分。

綜合這些討論，正確的選項是(1)(2)(3)(4)(5)。

習題 5

設第 7 位裁判員給涅莫夫的分數是 x 分。將已知的 6 個分數依小到大排列為

$$9.6,\ 9.6,\ 9.7,\ 9.8,\ 9.8,\ 9.9.$$

(1) 若 x 是七個分數中最小的，則此時 x 與 9.9 是被捨去的分數，那麼涅莫夫的最後成績應為

$$\frac{9.6+9.6+9.7+9.8+9.8}{5}=9.7\neq 9.74 \text{ (不合)}$$

(2) 若 x 是七個分數中最大的，則此時 9.6 與 x 是被捨去的分數，那麼涅莫夫的最後成績應為

$$\frac{9.6+9.7+9.8+9.8+9.9}{5}=9.76\neq 9.74 \text{ (不合)}$$

(3) 若 x 在七個分數中不是最大，也不是最小的，則此時 9.6 與 9.9 是被捨去的分數，那麼涅莫夫的最後成績應為

$$\frac{x+9.6+9.7+9.8+9.8}{5}=9.74 \Rightarrow x=9.8.$$

故第 7 位裁判員給涅莫夫的分數是 9.8 分。

習題 6

設薪資最高20%及最低20%的員工之平均月薪資分別是 x 與 y 萬元。

依題意得到

$$\begin{cases} x=3y \\ (x-1.6)=2.8(y-0.4) \end{cases} \Rightarrow \begin{cases} x=3y \\ 3y-1.6=2.8y-1.12 \end{cases}$$

$$\Rightarrow \begin{cases} x=7.2; \\ y=2.4. \end{cases}$$

因此薪資最高20%及最低20%的員工之平均月薪資分別是 7 萬 2000 元與 2 萬 4000 元。

習題 7

設最高20%與最低20%家庭平均可支配所得分別是 x 與 y 萬元。

依題意得到

$$\begin{cases} x=6.4y \\ x-y=151.2 \end{cases} \Rightarrow \begin{cases} x=179.2; \\ y=28. \end{cases}$$

因此最高 20%與最低20%家庭平均可支配所得分別是 179 萬 2000 元與 28 萬元。

習題 8

(1) 平均數代表的是真正的算術平均值，中位數只是正中間那個數據而已。在此問題上，取算術平均值是較確實的算法。故蔡教授的估算較妥當。

(2) 利用反證法，如果有學生喝超過 15 罐汽水。由標準差的公式得到

$$1.6=\text{標準差}\geq\sqrt{\frac{(16-4)^2}{35}}=\sqrt{\frac{144}{35}}>2$$

矛盾。故沒有學生喝超過 15 罐汽水。

習題 9

因為平均數 100，標準差 15，所以 115 剛好比平均數多 1 個標準差。贏過的人都在平均數正 1 個標準差以下，其百分比為

$$68\% + \frac{100\% - 68\%}{2} = 84\%.$$

習題 10

因為平均數 266，標準差 16，所以在 234 天內分娩代表低於平均數 2 個標準差。所佔的百分比為

$$\frac{100\% - 95\%}{2} = 2.5\%.$$

習題 11

因為平均數 560，標準差 120，所以 800 分以上代表高過平均數 2 個標準差。所以收到 800 分成績單的人數百分比為

$$\frac{100\% - 95\%}{2} = 2.5\%.$$

習題 12

(1) 設 x 軸代表速度，y 軸代表對應的汽油里程，得到的五組數據為

$$(x, y) = (20, 22), (30, 28), (40, 30), (50, 28), (60, 22).$$

對應的平均數分別為

$$\bar{x} = \frac{20+30+40+50+60}{5} = 40; \ \bar{y} = \frac{22+28+30+28+22}{5} = 26.$$

根據相關係數的公式得到相關係數 r 為

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^{5}(\bar{x}-x_i)(\bar{y}-y_i)}{\sqrt{\sum_{i=1}^{5}(\bar{x}-x_i)^2}\sqrt{\sum_{i=1}^{5}(\bar{y}-y_i)^2}} \\
&= \frac{(20)(4)+(10)(-2)+(0)(-4)+(-10)(-2)+(-20)(4)}{\sqrt{1000}\sqrt{56}} \\
&= \frac{0}{\sqrt{56000}} \\
&= 0.
\end{aligned}
$$

(2) 由拋物線的對稱性質知道：此二次函數的頂點坐標為(40,30)，而且開口向下。令二次函數為

$$y = a(x-40)^2 + 30.$$

將(50,28)代入得到 $a = -\dfrac{1}{50}$。故二次函數為

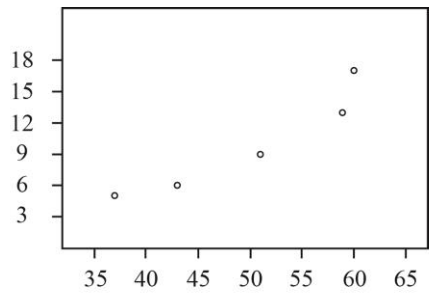
$$y = -\frac{(x-40)^2}{50} + 30.$$

(3) 將 $x = 70$ 代入(2)中的二次函數得到 $y = 12$。故當汽車速度為每小時 70 英里時，每加侖汽油能跑 12 英里。

〔註〕這問題告訴我們，當數據近似一條直線時，相關係數的絕對值才會高；當數據落在非直線的圖形上時，相關係數不高。

習題 13

(1) 繪製散佈圖：圖中的 x 坐標表示銅製人體模型的「身體長度（公分）」，而 y 坐標表示該模型的「手臂長度（公分）」



(2) 計算相關係數：散佈圖中，五個點的座標為

$$(37,5), (43,6), (51,9), (59,13), (60,17).$$

x 與 y 座標的平均值 $\overline{x},\overline{y}$ 分別為

$$\overline{x}=\frac{37+43+51+59+60}{5}=50;\ \ \overline{y}=\frac{5+6+9+13+17}{5}=10.$$

根據相關係數的公式得到相關係數 r 為

$$\begin{aligned} r &= \frac{\sum_{i=1}^{5}(\overline{x}-x_i)(\overline{y}-y_i)}{\sqrt{\sum_{i=1}^{5}(\overline{x}-x_i)^2}\sqrt{\sum_{i=1}^{5}(\overline{y}-y_i)^2}} \\ &= \frac{(13)(5)+(7)(4)+(-1)(1)+(-9)(-3)+(-10)(-7)}{\sqrt{400}\sqrt{100}} \\ &= \frac{189}{200} \\ &= 0.945. \end{aligned}$$

(3) 因為相關係數高達 0.945，所以這五個銅製人體模型出自同一文化的可能性很高。

(4) 求出的相關係數仍為 0.945（註：相關係數的公式是由單位向量作內積而得，所以改變刻度並不會改變相關係數的計算值）。

習題 14

由期望值的定義知道：每張保單的期望獲利為

$$200-(0.0015\times 100000)=200-150=50$$

美元。

習題 15

莫理斯活過五年的機率為

$$0.9\times 0.6\times 0.7+0.9\times 0.4\times 0.5=0.558.$$

習題 16

設抽樣的選民數為 n 人。62.5% 與 67.5% 的中間值為

$$\frac{62.5\% + 67.5\%}{2} = 65\%.$$

因為此範圍亦可表為 $65\% \pm 2.5\%$，所以誤差範圍在 2.5 個百分點。由誤差範圍的數學公式得到

$$2\frac{\sqrt{65(100-65)}}{\sqrt{n}} = 2.5 \Rightarrow n = 1456.$$

答案為約抽樣 1456 個選民當樣本。

習題 17

設標準差為 σ。因為平均智商的範圍可表為

$$105\% \pm 1.2\%,$$

所以誤差範圍為 1.2 點。由誤差範圍的數學公式得到

$$2\frac{\sigma}{\sqrt{900}} = 1.2 \Rightarrow \sigma = 18.$$

故標準差為 18。

習題 18

我們以 95% 的信賴指數來計算麵包的平均重量賄賂在那個區間：由公式知道平均重量的上、下界分別是

$$\mu \pm 2\frac{\sigma}{\sqrt{n}}.$$

代入數字得到

$$197 \pm 2 \times \frac{10}{\sqrt{25}} \Rightarrow 197 \pm 4.$$

故物理教授如果懂統計語言的話，他應該說：「我有 95% 的信心相信，你的麵包的平均重量介於 $197-4=193$ 公克與 $197+4=201$ 公克之間。」因此，政府公定的 200 公克仍在這誤差範圍內。

統計思考

私立學校的分佈圖是不對稱的雙峰分佈，較小的高峰是人為力量產生的。此時的中位數位於最高峰的右側一點的地方，而平均數也偏右，但由於第二高峰離中位數較遠，產生的平均數較最左邊的大，故平均數會比中位數偏右。因此私立學校公布平均數是較有力的選擇。

公立學校的分佈圖是向右偏斜分佈，此時的中位數位於高峰的右側一點的地方，而平均數更偏右，這是因為中位數左側的數據離中位數進，產生的加權平均數較最右邊的小。故中位數會比平均數偏右。因此公立學校公布中位數是較有力的選擇。