University of Waterloo W. H. Cherry

GRAPHICAL ATTRIBUTES: Histograms and Bar-Graphs – How to Construct and Use Them

1. Introduction

Histograms and bar-graphs are a widely-used method of data visualization. Their essential component is one or more (rectangular) bars which convey numerical information; other components include Cartesian axes and labels to provide context for the information being presented. What distinguishes a histogram from a bar-graph is that, in histograms, the bars are contiguous (allowing for bars of zero height if any are present) whereas, in a bar-graph, the bars have space between them. It is also easy to forget that, in a histogram, it is bar *area* (*not* height) that represents proportion or probability.

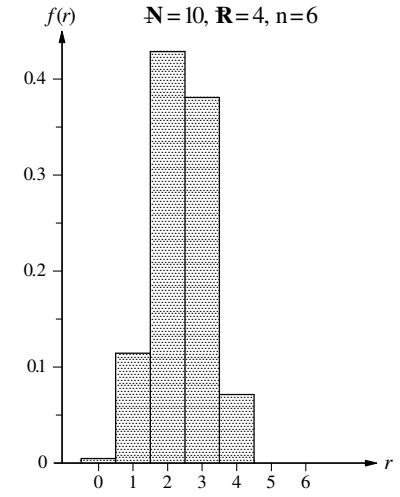
Choices to be made as the basis for constructing a histogram (*and* conveyed to users of the resulting histogram) include:

- the number of bars;
- the bar width(s) [including possible *un*equal bar widths] – bar widths are also called **intervals**;
- the starting value for the (main group of) bars;
- how to deal with observations on a bar edge/interval boundary;
- the labels (including a title) needed to achieve graphical excellence – giving the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space (see the bottom of page HL23.2 in Statistical Highlight #23).

We distinguish histograms for (discrete) probability distributions from histograms of data because the choices for consideration are usually more easily dealt with for probability distributions. The illustration below is for the *hypergeometric distribution* which (as discussed in Part 7 of the STAT 220 Course Materials) can be used to model the process of selecting a sample of n balls equiprobably with*out* replacement from an urn containing a total of $\mathbf{N}$ balls of which $\mathbf{R}$ are red and $\mathbf{N}-\mathbf{R}$ are black. If the random variable R represents the number of red balls in the sample, its probability function is as given in equation (HL26.1) at the right. When $\mathbf{N}=10$, $\mathbf{R}=4$ or $\mathbf{R}=6$, and $n=6$, we can (tabulate and) graph $f(r)$ as shown below – the numbers of bars and their labels follow immediately from the p.f. tables HL26.1 and HL26.2. [The bar *width* of 1 in these two histograms means that bar area is also its height.]

$$f(r) = \frac{\binom{\mathbf{R}}{r}\binom{\mathbf{N}-\mathbf{R}}{n-r}}{\binom{\mathbf{N}}{n}}, \quad r = 0, 1, 2, ..., n. \qquad \text{-----(HL26.1)}$$

Two Hypergeometric Probability Functions



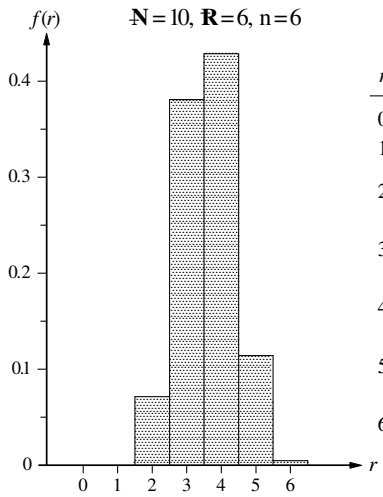


Table HL26.1

r	$f(r)$
0	$\frac{1}{210} \simeq 0.0048$
1	$\frac{24}{210} \simeq 0.1143$
2	$\frac{90}{210} \simeq 0.4286$
3	$\frac{80}{210} \simeq 0.3810$
4	$\frac{15}{210} \simeq 0.0714$
5	0
6	0

Table HL26.2

r	$f(r)$
0	0
1	0
2	$\frac{15}{210} \simeq 0.0714$
3	$\frac{80}{210} \simeq 0.3810$
4	$\frac{90}{210} \simeq 0.4286$
5	$\frac{24}{210} \simeq 0.1143$
6	$\frac{1}{210} \simeq 0.0048$

The two histograms above and those for other probability distributions are given in Statistical Highlight #4.

Another probability function, of interest here for its histogram, *not* its probabilistic background, is Table HL26.3 below, which arises in the context of a population of size $\mathbf{N}=8$ elements with response variate $\mathbf{Y}_i$ values of 1, 2, 4, 4, 7, 7, 7, 8; the population *total* response, denoted $_T\mathbf{Y}$ (the sum of these eight $\mathbf{Y}_i$ values), is 40, which equals $\mathbf{N}\overline{\mathbf{Y}}$, where $\overline{\mathbf{Y}}=5$ is the population *average* response. Because $\mathbf{N}\overline{\mathbf{Y}} = {}_T\mathbf{Y}$, as an illustration of the probabilistic basis for survey sampling theory (discussed in Statistical Highlight #77), we can *estimate* the population total by $_Ty = \mathbf{N}\overline{y}$, where $\overline{y}$ is the average of a *sample* selected equiprobably from the population. With $\mathbf{N}=8$ and if we choose a sample size of $n=4$ for this illustration, there are 70 possible samples with 15 possible (non-zero) $_Ty = \mathbf{N}\overline{y}$ values ranging from 22 to 58; we see from Table HL26.3 that, for example, four of the 70 samples have $_Ty = 36$.

To assist readers who wish to understand more fully the probabilistic background of this illustration, Table HL26.3 gives both the (neater) integer values of $70 \cdot f(_Ty)$ and the corresponding approximate (non-zero) decimal values of $f(_Ty)$; it also includes the 22 integers (like 23, for instance) between 22 and 58 that are *not* possible values of $_Ty$ and so have zero $f(_Ty)$ values.

Table HL26.3

$_Ty = \mathbf{N}\overline{y}$	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
$70 \cdot f(_Ty)$		1	0	0	0	0	0	6	0	2	0	3	0	7	0	4	0	6	0	12
$f(_Ty)$		0.0143						0.0857		0.0286		0.0429		0.1		0.0571		0.0857		0.1714

	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
$70 \cdot f(_Ty)$	0	6	0	4	0	7	0	3	0	2	0	6	0	0	0	0	0	1		
$f(_Ty)$		0.0857		0.0571		0.1		0.0429		0.0286		0.0857						0.0143		

A histogram of the Table HL26.3 probability function is given overleaf on page HL26.2 at the upper left; its notable features are:

(*continued overleaf*)

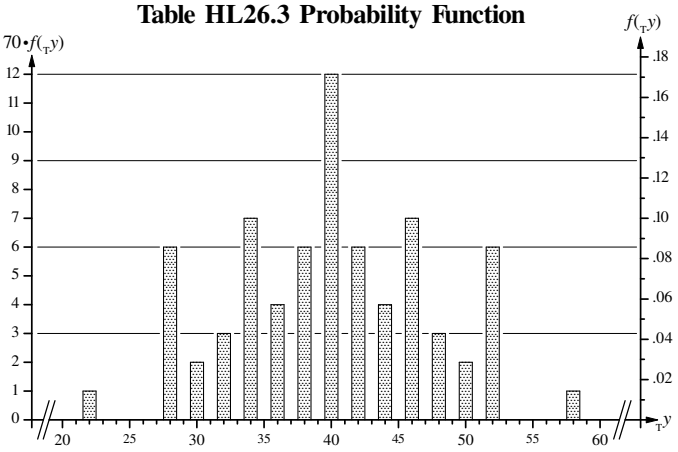
∗ it has two vertical axis scales – the integer $70 \bullet f(_{T}y)$ values on the left, the decimal $f(_{T}y)$ values on the right;

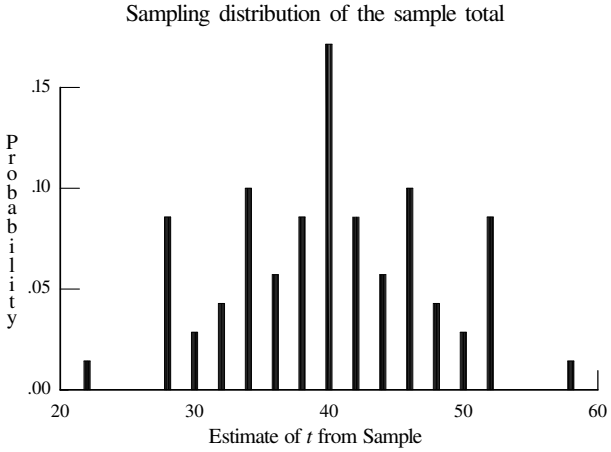
∗ it could be mistaken for a bar-graph on account of the (relatively) numerous gaps between bars – the 22 zero values of $f(_{T}y)$.

On the right is the original version of the histogram, where a carelessly-chosen bar width (too small) makes it an *actual* bar graph; its title seems to be confusing the *sample* total $n\overline{y}$, with $\mathbf{N}\overline{y}$, the *estimate* of the *population* total (denoted '*t*' in the x-axis label). Graphical malfeasance like this is not helpful to the image of statisticians as purveyors of important and useful information.

[The *sampling distribution* of an estimator (like $\overline{Y}$, the random variable representing the sample average $\overline{y}$ under equiprobable selecting, as an estimator of the population average $\overline{\mathbf{Y}}$) is the distribution of $\overline{Y}$ over the set of all possible samples of a given size from the population; in the illustration, this set contains 70 samples, with totals ranging from 11 to 29 in value. The sample total is usually *not* an estimator of interest.]

Table HL26.3 Probability Function





2. Histograms for Data

The following (hypothetical) data refer to water consumption (in millions of litres) in a small city on n = 100 consecutive days; two numerical data summaries are given at the right. $\sum_{j=1}^{n} y_j = 2{,}961$, $\sum_{j=1}^{n} y_j^2 = 88{,}883$

The data ordered from smallest to largest are shown at the right; this is simetimes (unhelpfully) referred to as an '**order statistic**'.

.....DATA SET.....									
25	32	28	26	30	21	36	27	32	32
33	23	31	29	33	31	29	33	28	31
27	29	29	33	27	28	32	23	30	25
29	32	21	24	30	31	28	32	31	34
31	26	30	30	34	26	26	29	32	22
24	31	28	32	28	37	30	35	27	33
28	30	33	30	31	30	32	28	31	28
29	26	32	24	29	32	30	22	34	26
32	27	27	34	29	31	33	39	30	33
30	28	31	30	37	31	25	32	33	28

.....ORDERED DATA SET.....									
21	21	22	22	23	23	24	24	24	25
25	25	26	26	26	26	26	26	27	27
27	27	27	27	28	28	28	28	28	28
28	28	28	28	28	29	29	29	29	29
29	29	29	29	30	30	30	30	30	30
30	30	30	30	30	30	30	31	31	31
31	31	31	31	31	31	31	31	31	32
32	32	32	32	32	32	32	32	32	32
32	32	33	33	33	33	33	33	33	33
33	34	34	34	34	35	36	37	37	39

To construct an (informative) histogram of these data, we make the following choices, likely with one or more reconsiderations (and refinements) being needed to achieve graphical excellence.

- **Number of bars and bar width:** the data set contains 100 observations with a range of 39 – 21 = 18 units; a choice of 3 units for bar width should yield a histogram with 5 to 10 bars with frequencies that are generally not too small.
- **Interval end points:** we choose to start the bars/intervals at 20 (an apppropriate 'round number' here), although 21 (the smallest observation) is also a possible choice (see page HL26.4). With integer data values and integer interval end-points, we choose the closed/open interval convention at the right.

[signifies a closed interval; *i.e.*, endpoint is *in*cluded.
) signifies an open interval; *i.e.*, endpoint is *ex*cluded.

- **Bar height calculation:** the calculations are shown in Table HL26.4 at the right.
  - + The frequency (the number of observations) for each interval is found by working through the ordered data set.
  - + Each proportion is found by dividing its frequency by the number of observations in the data set, here 100.
  - + Each bar height is found by dividing its proportion by its width – this calculation accommodates the less common histograms whose bars are *not* all the same width.

  The two rightmost columns of Table HL26.4 show the use of percentages rather than proportions; the former may be more approriate for some histogram user groups.

Table HL26.4

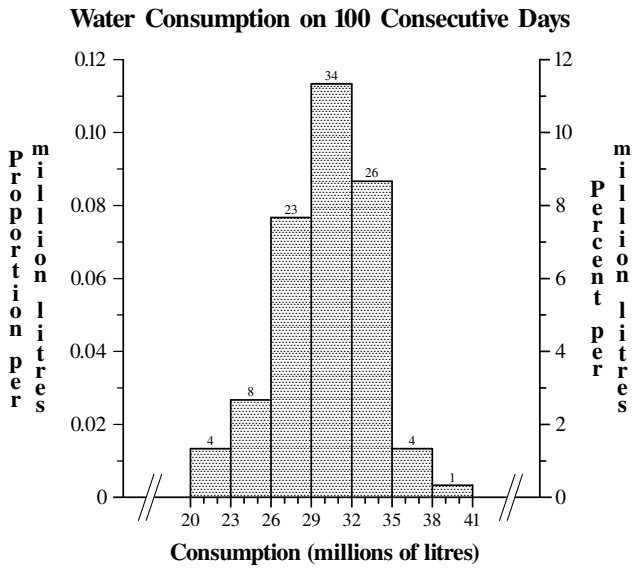
Bar width/ Interval	Frequency	Proportion	Bar Height	Percent	Bar Height
[20-23)	4	0.04	$0.01\dot{3}$	4	$1.\dot{3}$
[23-26)	8	0.08	$0.02\dot{6}$	8	$2.\dot{6}$
[26-29)	23	0.23	$0.07\dot{6}$	23	$7.\dot{6}$
[29-32)	34	0.34	$0.11\dot{3}$	34	$11.\dot{3}$
[32-35)	26	0.26	$0.08\dot{6}$	26	$8.\dot{6}$
[35-38)	4	0.04	$0.01\dot{3}$	4	$1.\dot{3}$
[38-41)	1	0.01	$0.00\dot{3}$	1	$0.\dot{3}$
	100	1.00		100	

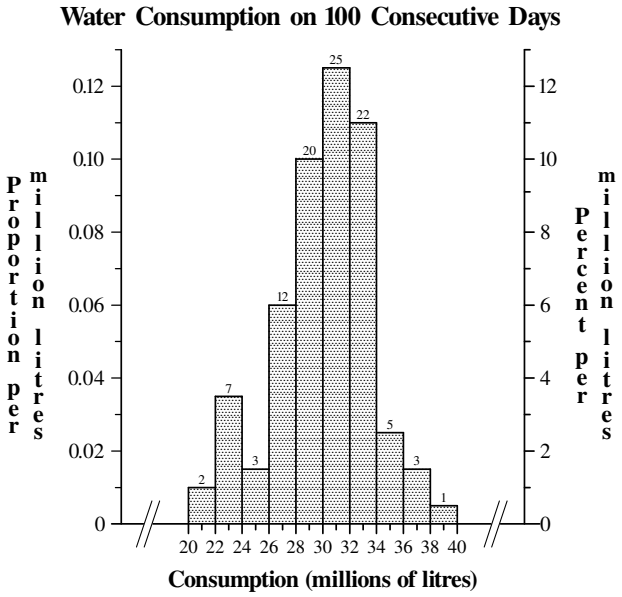
(*continued*)

University of Waterloo W. H. Cherry

GRAPHICAL ATTRIBUTES: Histograms and Bar-Graphs – How to Construct ... Them (continued 1)

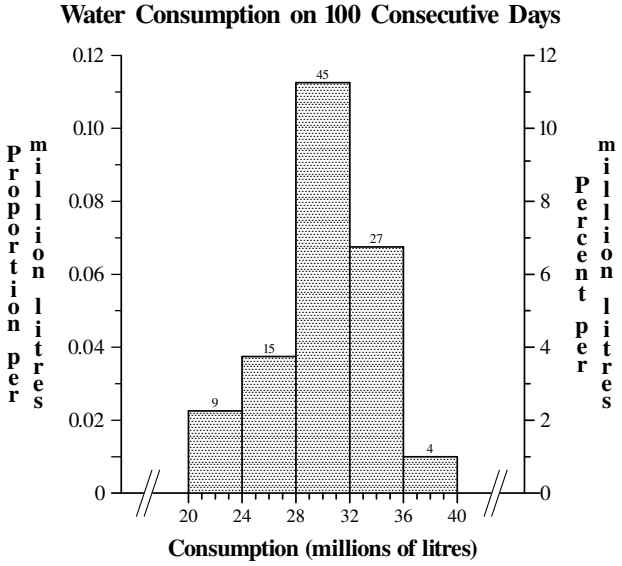
The histogram resulting from the choices described on the lower half of the facing page HL26.2 for the water consumption data is shown at the left below; for interest in this Highlight #26, the histograms for two other choices of bar width (2 and 4 units instead of 3) are shown at the right below. To illustrate Table HL26.4, these three histograms have (on their left and right) *both* proportion and percentage y-axis scales; in practice, usually only *one* (appropriately-chosen) scale would be shown. To assist the user, these histograms show bar frequencies at the tops of the bars.





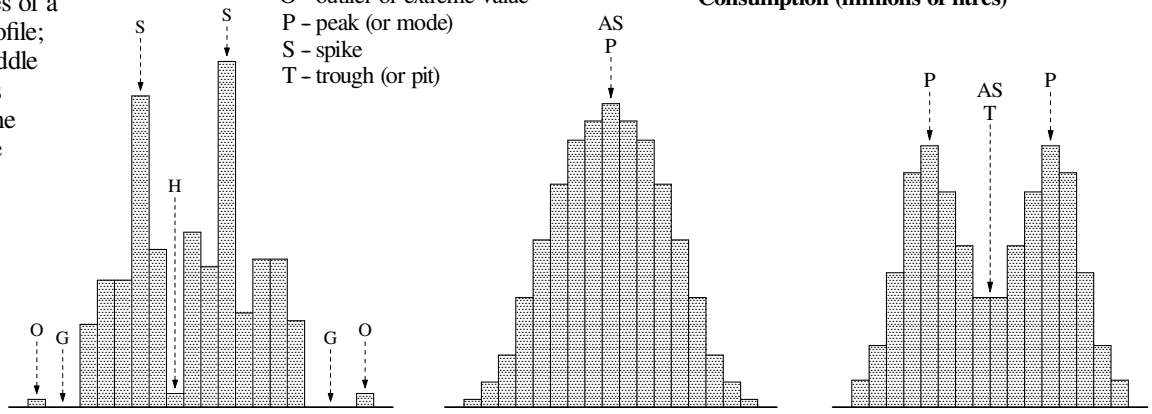
The left-hand histogram above shows a roughly symmetrical distribution of daily water consumption; a probability model for this distribution shape could be a normal distribution (discussed in Part 5 of the STAT 220 Course Materials). The other two histograms warn of dangers associated with an *in*appropriate choice of bar width.

- Too *narrow* a bar width risks showing (likely inconsequential) aberrations in distribution shape resulting from erratic small frequencies – the bar with frequency 7 at 22-24 at the upper right may be this phenomenon.
- Too *wide* a bar width risks distorting (*e.g.*, over-simplifying) the distribution shape – the *de*creased symmetry in the diagram at the right (with only five bars of width 4) may be this phenomenon.



The three diagrams below illustrate terminology at the right sometimes useful for describing notable features of a histogram's profile; we say the middle distribution is '**unimodal**', the right-hand one is '**bimodal**'. [Distributions and their features are seen more readily from a histogram than from a table.]

AS – axis of symmetry
G – gap
H – hole
O – outlier or extreme value
P – peak (or mode)
S – spike
T – trough (or pit)



(*continued overleaf*)

University of Waterloo W. H. Cherry

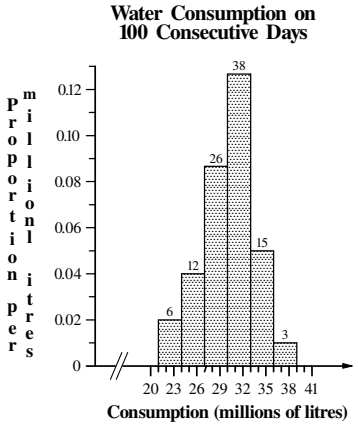
The distinction between uni- and bimodal distributions may be useful in, for example, a manufacturing context. If the same component is made on two different machines and a sample of components from the two machines shows a *bi*modal distribution for (say) a critical dimension, it may mean that the machines need to be adjusted to make their *average* for this dimension the same. More generally, a bimodal histogram suggests that the data come from (two) distinct groups of items that it may be useful to identify.

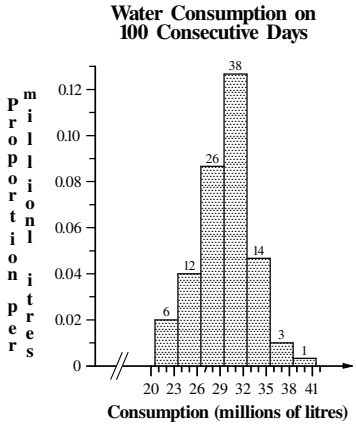
The ordered data set (of 66 proficiency scores) at the right, from question **A2 – 1** on STAT 220 Assignment 2, with a bar width of 50 yields a *uni*modal histogram when started at 300, a *bi*modal histogram when started at the smallest observation of 345. This reminds us why the choices made before constructing a histogram need care and possible reconsideration (statistical experience is also helpful).

345	354	384	395	402	406
417	417	420	431	439	439
444	446	461	464	468	468
468	472	475	479	486	490
490	490	494	494	505	505
505	515	516	523	523	523
523	527	530	541	545	549
549	556	560	563	574	574
575	578	582	585	585	585
593	596	603	604	604	607
611	624	629	629	691	730

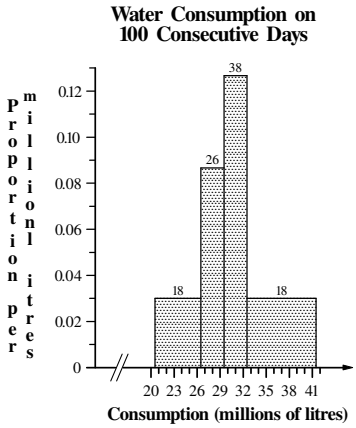
For the (integer) water consumption data, the three histograms overleaf on page HL26.3 all start at 20, a 'round number' just below the smallest observation of 21 million litres. At the right are two smaller versions of the left-hand one of the three histograms (with a bar width of 3) – these two illustrate the effect of choice of starting point for a histogram:

- in the left-hand histogram, the starting point is the smallest observation of 21 million litres;
- the right-hand histogram starts at 20½ million litres – the half-integer choice here avoids the need for the closed/open interval convention in the box at the lower right of page HL26.2.



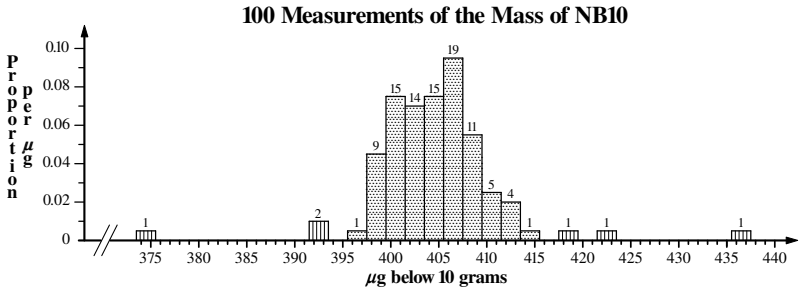


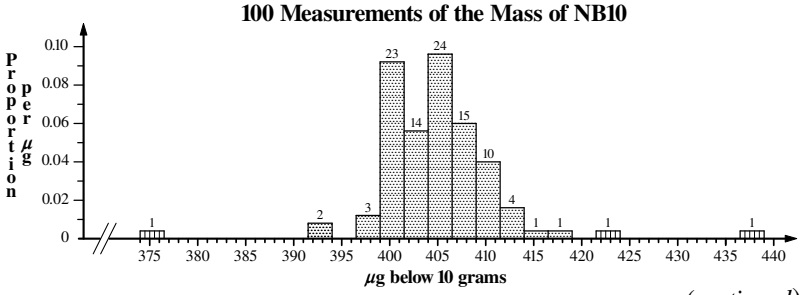
The two histograms have very similar bar frequencies (except that there is one fewer bar in the histogram on the left); both also show distributions generally similar to that of the left-hand histogram on page HL26.3. Thus, *un*like the proficiency scores above, the histogram shape for these water consumption data is *in*sensitive to choice of starting point.



The (lower) histogram at the right illustrates different bar *widths* – the bar of width 6 on the left replaces two bars of width 3 in the histogram above it and, similarly, a bar of width 9 on the right replaces three bars of width 3 above it; the frequencies for these wider bars are (of course) the total of the frequencies of the narrower bars they replace. Likewise, their *areas* are (of course) the same as the total areas of the bars they replace. Different bar widths in *this* histogram are **not** appropriate – they distort its shape. However, there are contexts where different bar widths in a histogram are essential for graphical excellence or are unavoidable for the data as available; examples are questions **A2 – 3** and **A2 – 4** in STAT 220 Assignment 2 – the former also raises the (unappealing) options for dealing with an open interval.

The two histograms at the right involve a data set of 100 measurements of the mass of the U.S. ten-gram standard NB10 – the values are micrograms (μg) below ten grams (more details are given on page HL2.1 in Statistical Highlight #2). The histograms have bar widths of 2 μg (upper version) and 2½ μg (lower version); the former suggests *six* outliers in the data set, the latter suggests only three – their bars are shaded with vertical lines instead of dots. While unequivocal identification of 'outliers' in a data set can be difficult, the difference in the matter of outliers between the two versions is another instance of the hazards of histogram construction, although a bar width of 2½ μg is a curious choice for this data set,





(*continued*)

University of Waterloo W. H. Cherry

GRAPHICAL ATTRIBUTES: Histograms and Bar-Graphs – How to Construct ... Them (continued 2)

and it necessitates a closed/open interval convention; the upper NB10 histogram avoids this by its half-integer starting point, in the same way that the right-hand smaller water consumption histogram on the upper half of the facing page HL26.4 does.

The eight data histograms on pages HL26.3 and HL26.4 show that there are *dependencies* among the choices (listed on the upper half of page HL26.1) to be made as the basis for constructing a histogram of a particular data set.

- Obviously, the number of bars and the bar width are dependent – choices here strive to balance reasonable bar frequencies with bars of such a width that the finished histogram properly conveys the data distribution.
- Less obviously, bar width, starting value and dealing with observations on bar edges/interval boundaries are also dependent – three options for the starting value:
  - a 'round number' just below the start of the (main group of) bars,
  - the smallest observation in the data set,
  - a starting point just below the smallest observation but with one more decimal place than the data,

  affect bar width and whether a closed/open interval convention is needed.

Dependencies like these make the task of achieving graphical excellence in a histogram usually an onerous (and iterative) process. Surprisingly for a statistical tool as common and as useful as histograms, (the usually brief) discussion in introductory texts may be marred by statements like ... *choose any convenient number of intervals but always choose intervals of equal width*; the term '**relative frequency**,' an obscure and unnecessary duplication of 'proportion,' may also be used.

The emphasis of the forgeoing discussion is histogram *distribution*, but a histogram conveys other useful information about a data set, most obviously about its attribute values for location and spread; the *actual* values, calculated from the two data summaries at the right just above the ordered data set on page HL26.2, are an average of 29.61, a median of 30 and a (data) standard deviation of 3.4928 million litres. Looking at the upper left-hand water consumption histogram on page HL26.3, we see that:

- the *average* daily water consumption should be around the middle of the tallest central bar of a roughly symmetrical distribution – about 30½ million litres,
- the *median* should also be about 30½ million litres, close to the average of a roughly *symmetrical* distribution,
- for reasons discussed in Figures 5.2 and 5.3 in Part 5 of the STAT 220 Course Materials (the so-called 68-95-99.7 rule for the normal distribution), the *standard deviation* should be around one-sixth of the histogram range of 21, about 3½ million litres.

These histogram values are close enough to the actual values to be useful – the standard deviation agreement is fortuitously close.

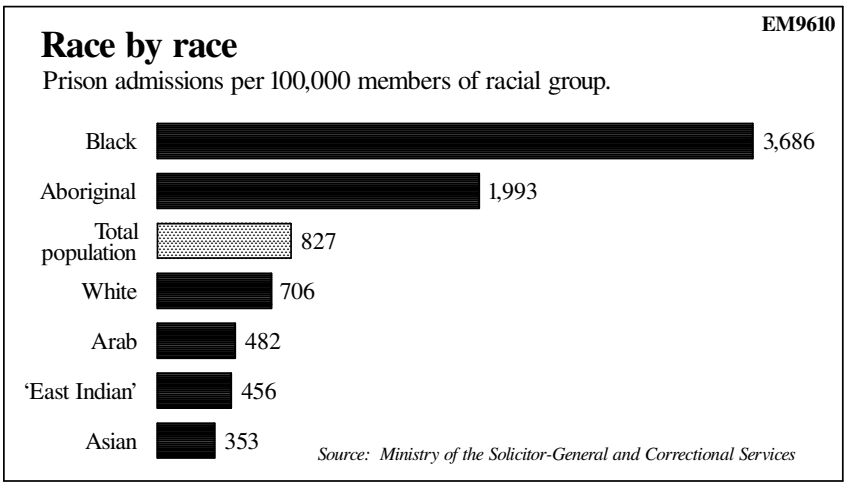
A histogram typically involves compromises associated with the choices made for its construction – for example, bar frequencies vary among the three water consumption histograms with bar width 3 on pages HL26.3 and HL26.4; for this reason, we should avoid trying to refine the approximate data-set attribute values by, for instance, arguing that, in aggregate, the frequencies of the three bars to the left of the tallest bar on page HL26.3 are somewhat higher than those of the three bars to its right, so the average should be a bit *less* than 30½; that this happens to be the case here is, again, fortuitous.

- Approximate data-set attribute values from a histogram can be (adversely) affected by a closed/open interval convention; this effect could be managed by changing the convention to half the values on an interval boundary being moved up and half moved down, but this has the (mildly-disconcerting) effect of making non-integer bar frequencies possible.

The effect of distribution *shape* on the relationships among values of the average, median and mode are pursued in Section 8 of Figure 4.1 of the STAT 220 Course Materials.

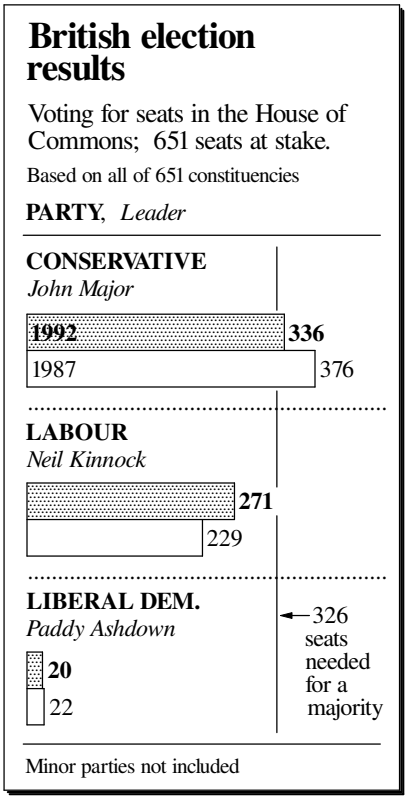
3. Using Bar-graphs for Comparisons and Trends

In contrast to the preoccupation of histograms with *distribution*, bar-graphs often visualize comparisons or trends; the example at the right and those on pages HL26.6 to HL26.8 are taken from the newspaper article referenced at the upper right of each display. [The reader may find it of interest to assess the degree to which each display meets the criteria for graphical excellence, as listed at the bottom of page HL23.2 in Statistical Highlight #23, although conflicts between theory and practice may arise from requiring ... *the*

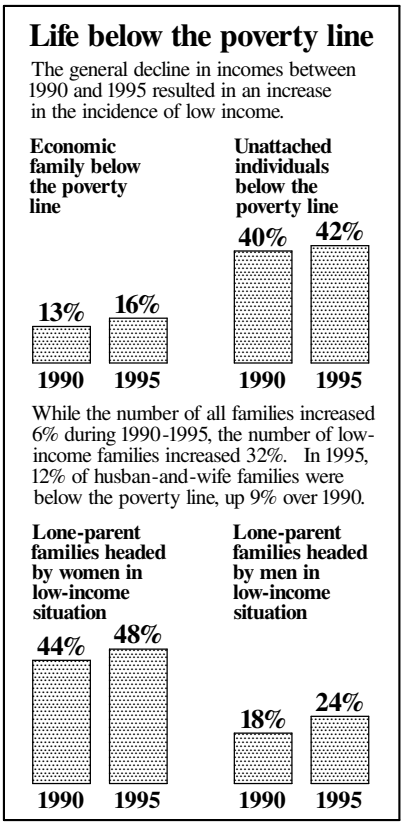


(*continued overleaf*)

EM9212

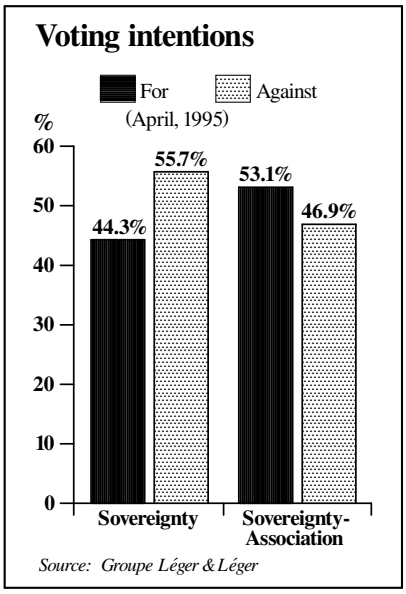


EM9807

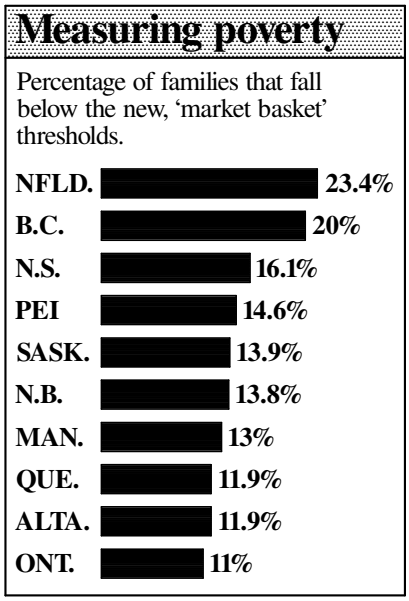


2020-08-20

EM9514



EM0302



EM9357

least ink in the smallest space.]

The bar-graphs reproduced on pages HL26.5 to HL26.8 are notable for three reasons.

- The wide variety of contexts in which they are used.
- The reminder that *orientation* of bars (of either bar-graphs or histograms) can be varied (usually as vertical or horizontal) to suit their context.
  - As with stemplots, back-to-back histograms or bar-graphs (usually with horizontal bars) can be useful to aid comparisons.
- Appropriate *labelling* of a display is essential for achieving graphical excellence.

4. Augmenting Bar Information for Histograms and Bar-graghs

The lower right display on the facing page HL26.7 from the newspaper article EM0422 illustrates using different (two, in this instance) shading patterns within bars to provide additional information to the viewer. The same idea (with five patterns) is used in the display overleaf on page HL26.8. In its original version, this display used five solid *colours* rather than (black and white) shading to show bar sub-categories; the text of the article EM0216, which accompanied the display of primary interest here, is included in this Highlight #26 to allow a reader to assess the degree to which the two components in combination commmunicate more effectively than either would do on its own.

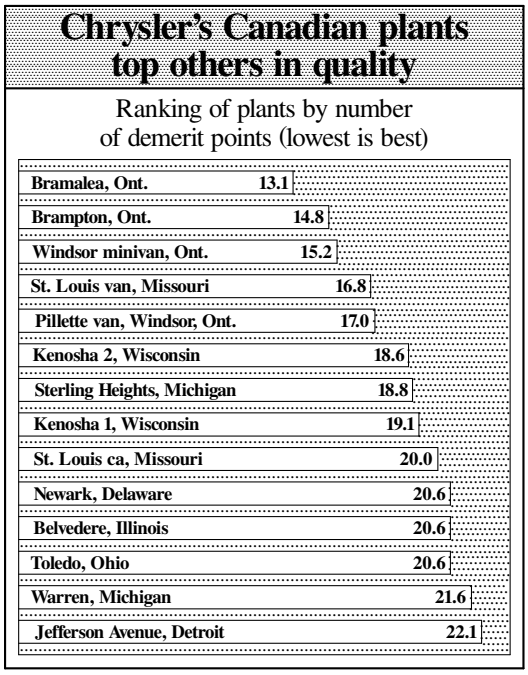
- Like comments on the overleaf side (page HL23.2) of Statistical Highlight #23, the useful additional bar information in the five subdivisions has the *dis*advantage for the viewer of making more difficult comparisons of bar sub-categories because their bases are no longer aligned horizontally.
  - Although it does not affect materially the illustrative use of the bar-graph

(*continued*)

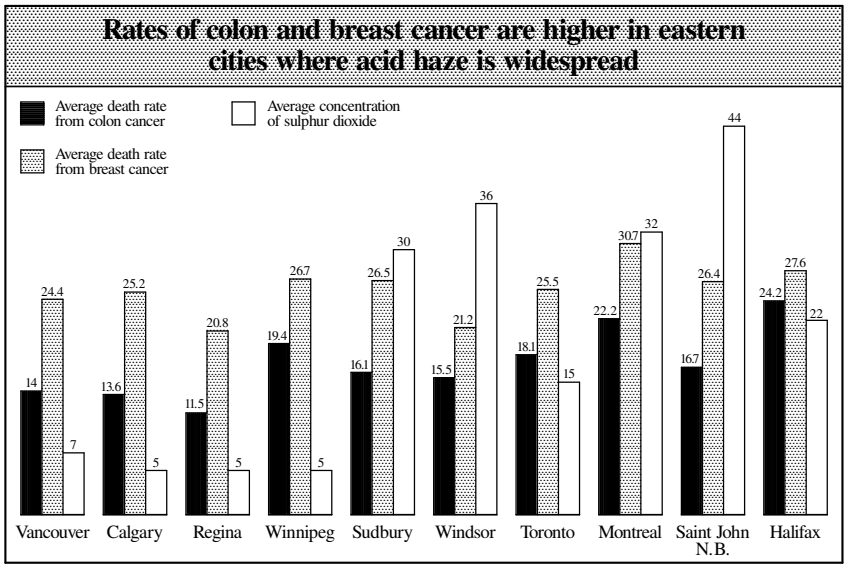
University of Waterloo W. H. Cherry

GRAPHICAL ATTRIBUTES: Histograms and Bar-Graphs – How to Construct ... Them (continued 3)

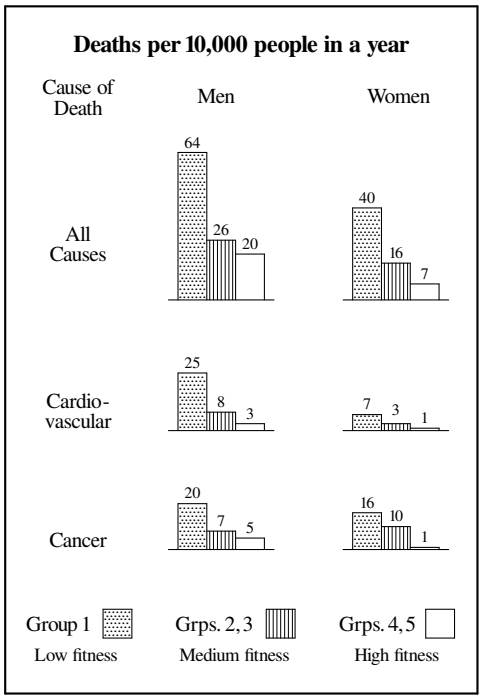
EM8903



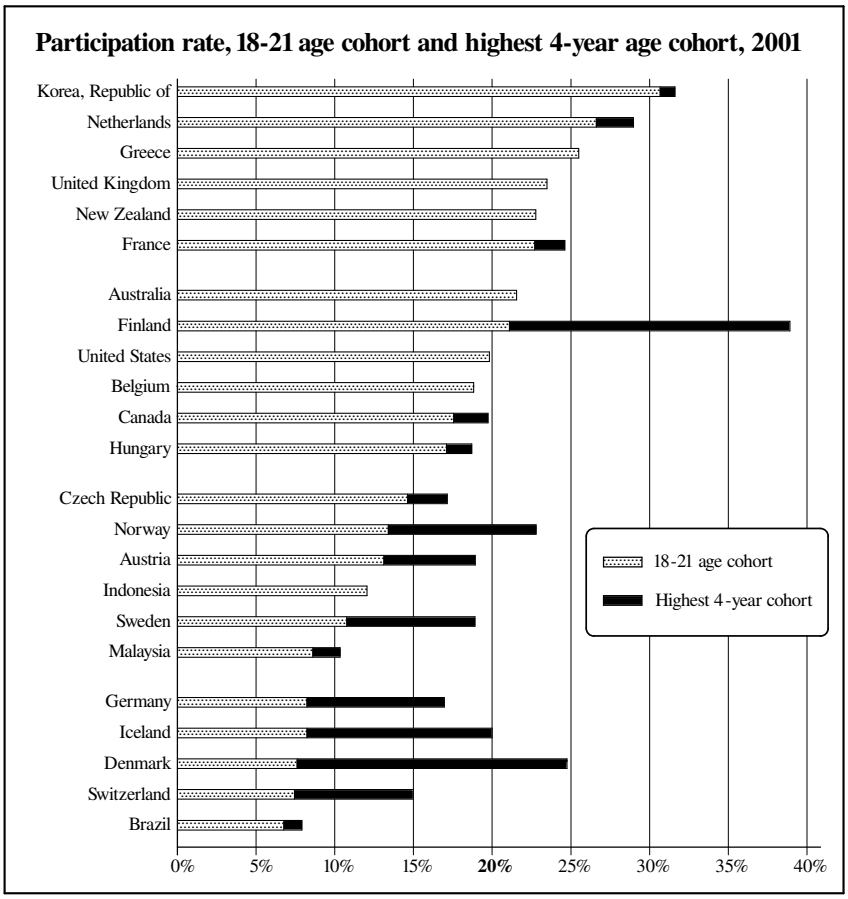
EM9102



EM9024



EM0422



from the article EM0216 in this Highlight #26, it is of interest to reconcile the disagreement of the article's title with the bar-graph. For example, the first box of the legend (the lightest vertical-line bar shading) is consistent in that this shading seems to occupy half of each bar's height, but the other shading patterns seem more in agreement with the article title (and content) if the order of the last four legend labels is reversed.

- Including the 'Total permits issued' category in the legend and bar-graph may *not* be graphical excellence because:
  - it is a category subsidiary to the topic of the article EM0216 and may distract the reader,

(*continued overleaf*)

○ the *same* information is provided by the remainder of each bar and with easier comparisons because the lower ends of the bars are aligned horizontally. [The (unnecessary) 'Total permits issued' category takes up about half of the display's space.]

EM0216: The Kitchener-Waterloo Record, April 8, 2002, page B3

Detached homes dominate Cambridge housing market

CAMBRIDGE

BY ROSE SIMONE
RECORD STAFF

Single detached homes are still dominating the housing market in Cambridge, according to the city's annual housing report released this week.

The report, with numbers compiled by the Canada Mortgage and Housing Corporation, shows the upward movement in house prices to an average of $167,000 for a single detached home last year, compared to $161,000 the previous year.

But even with that trend, 81.6 per cent of the housing built last year built in Cambridge were detached homes. People seemed to prefer the detached home over the row houses or the semi-detached units.

Meanwhile, less the one per cent of all units built last year were apartments.

Starting in the mid- to late-1990s, more first-time buyers purchased homes because of the lower mortgage interest rates and, over the past two years, many of those buyers were moving up into more expensive homes, the report shows.

Fewer homes were sold in the $130,000 to $149,000 range, but the number of homes sold in the $220,000 to $249,000 range more than doubled.

Robert Hemple, a Cambridge planner who prepared the annual housing report, said once people have a down payment and can afford to get into the housing market, they prefer to buy a detached house.

"They would rather have their own piece of real estate."

The main housing boom last year centred on three areas: Hespler East, the Shade's Mill neighbourhood and south of Myers Road.

But for all the people who still can't afford to buy because they don't have the down payment, the situation isn't so great.

For them, there was a bit of good news in that the number of vacant apartments went up slightly last year: Cambridge has had an extremely tight market for renters for the past three to four years.

The vacancy rate for apartments rose to 1.9 per cent of all units in 2001, which is a small improvement from the 1.1 per cent vacancy rate of the past two years.

But that's still lower than the three per cent vacancy rate considered an "ideal" for the market.

The problem for people with low incomes was that the vacant units were generally the ones with the higher rents, so more vacancies didn't make apartments more affordable.

The demand for two-bedroom apartments was the highest and the rents for those units increased an average of 3.9 per cent to $666.

A major issue for Cambridge is that fewer apartment buildings are being built compared to other cities such as Kitchener, Hemple said.

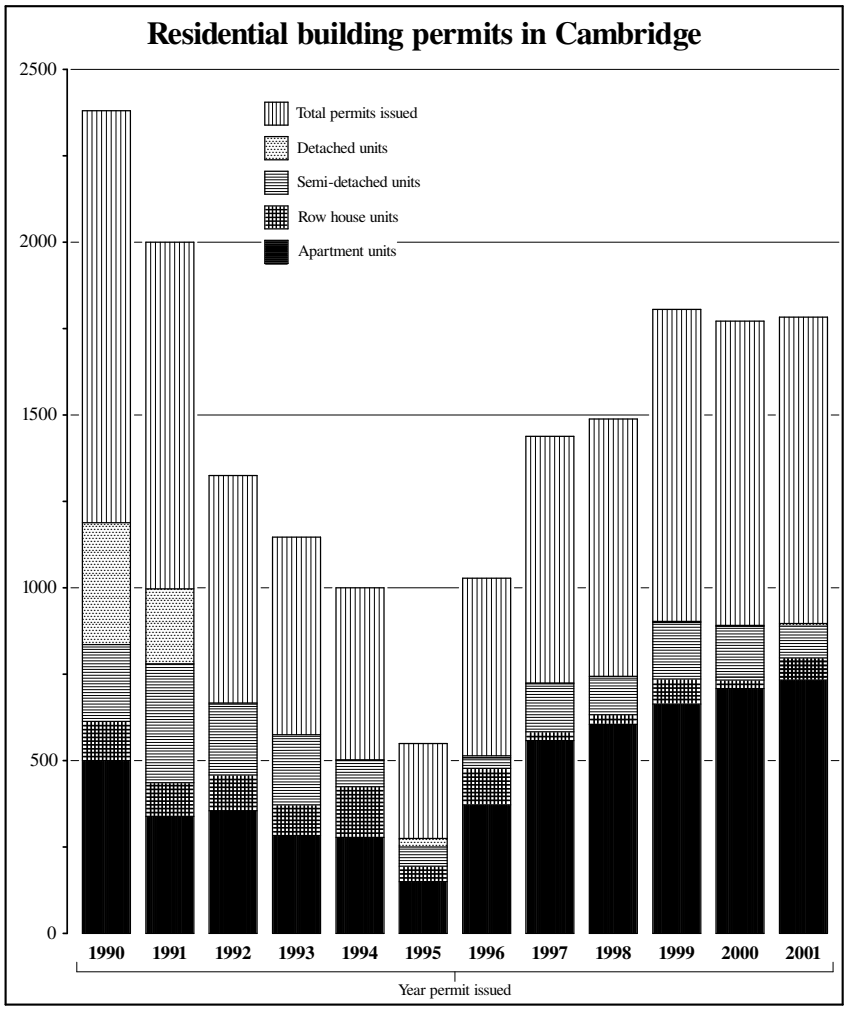
An interesting trend is that new homes have been going up on smaller lots.

In 1966, all of the single-family units built had frontages greater than 12 metres (39.4 feet), but in 1996, only 52 per cent of them had frontages of that size of more. That trend to smaller lots continued into 2001.

Even the "estate lots" being marketed now are much smaller than what people previously thought of as estates, Hemple said.

Nevertheless, that market strategy seems to be working and those lots are popular.

Cambridge councillors will be dealing with the housing report at a planning committee meeting on Monday.



Matters involving histograms and bar-graphs are also illustrated in Figures 2.5c, 2.7b and 3.8 of the STAT 220 Course Materials.