MAKING SENSE OF DATA

Essentials series

CHARTS AND GRAPHS

The main purpose of visualising data is to organize and display data to make your point clearly, effectively, and correctly. In this topic the most common data displays are given used to summarize categorical and numerical data, cautions on their interpretation, and tips for evaluating them.

Pie charts

A pie chart takes categorical data and shows the percentage of individuals that fall into each category. The sum of all the slices of the pie should be 100% or close to it (due to round-off error). Because a pie chart is a circle, categories can easily be compared and contrasted to one another.

The 2001 Census outputs can be displayed using a pie chart. Chart 1 shows religious affiliation in England & Wales. As can been seen almost ¾ of the population (42 million) view themselves as Christian.

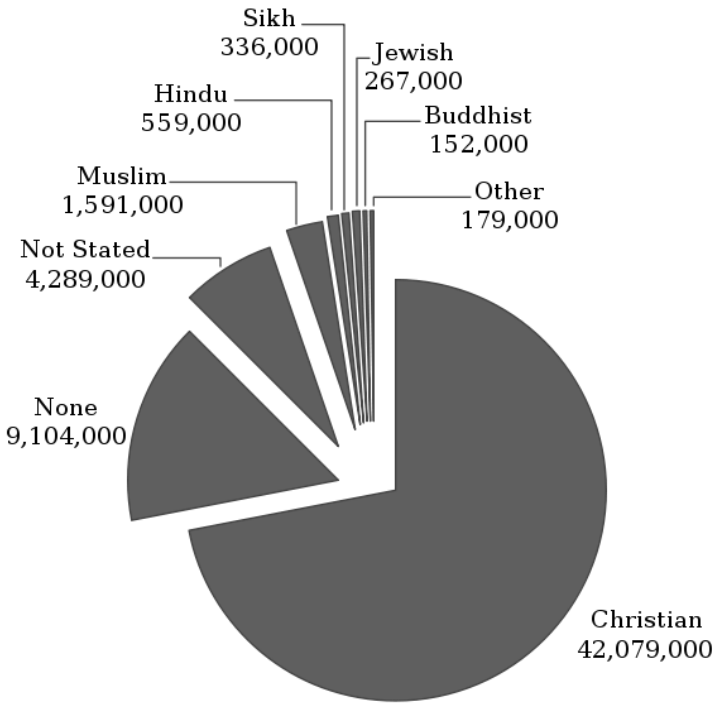

Chart 1: Religion in England & Wales (Census 2001)

To evaluate a pie chart for statistical correctness:

- ✓ Check to be sure the percentages add up to 100% or close to it (any round-off error should be very small).
- ✓ Beware of slices of the pie called “other” that are larger than many of the other slices. This shows a lack of detail in the information gathered.
- ✓ A pie chart may only show the percentage in each group, not the number in each group. Always ask for or look for a report of the total size of the data set.

Bar graphs

A bar graph is another means for summarizing categorical data. Like a pie chart, a bar graph breaks categorical data down by group, showing how many individuals lie in each group, or what percentage lies in each group.

Bar graphs are often used to compare groups by breaking down the categories for each and showing them as side-by-side bars. For example, has the percentage of mothers in the U.S. workforce changed over time? Chart 2 says "yes" and shows that the overall percentage of mothers in the workforce climbed from 47% to 72% between 1975 and 1998. Taking the age of the child into account, fewer mothers work while their children are younger and not yet in school, but the difference from 1975 to 1998 is still around 25% in each case.

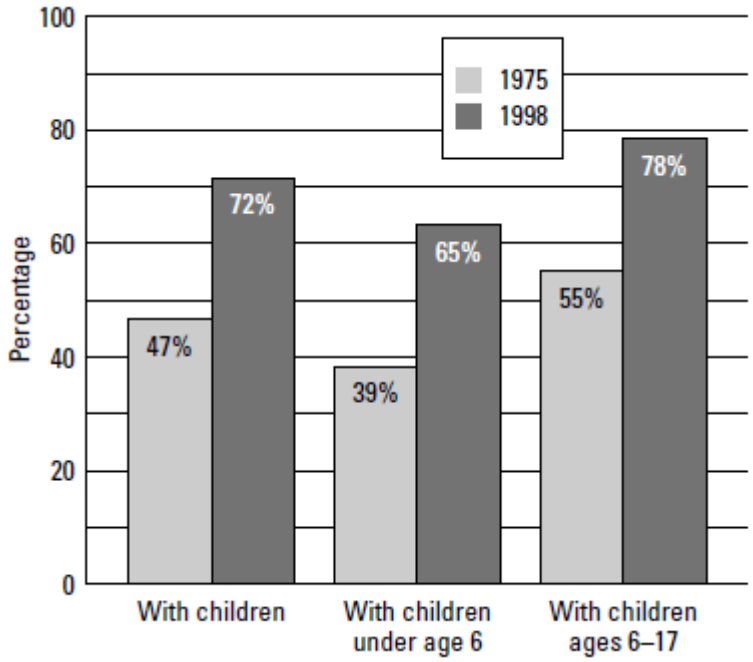

Chart 2: Percentage of mothers in workforce, by age of child (1975 and 1998 data from the U.S. Census)

Here is a checklist for evaluating bar graphs:

- ✓ Check the units on the y-axis. Make sure they are evenly spaced.
- ✓ Be aware of the scale of the bar graph (the units in which bar heights are represented). Using a smaller scale (for example, each half inch of height representing 10 units versus 50) you can make differences look more dramatic.
- ✓ In the case where the bars represent percents and not counts, make sure to ask for the total number of individuals summarised by the bar graph if it is not listed.

Time charts

A *time chart* is a data display whose main point is to examine trends over time. Typically this chart has some unit of time on the horizontal axis (year, month) and a measured quantity on the vertical axis (average household income, birth rate). At each time period, the amount is shown as a dot, and the dots connect to form the time chart.

The Office for National Statistics produces national population projections. Chart 3 displays actual and projected births and deaths in the U.K. (using a 2010-base). The largest projected gap between births and deaths will occur around 2016, with ~300 thousand more births than deaths.

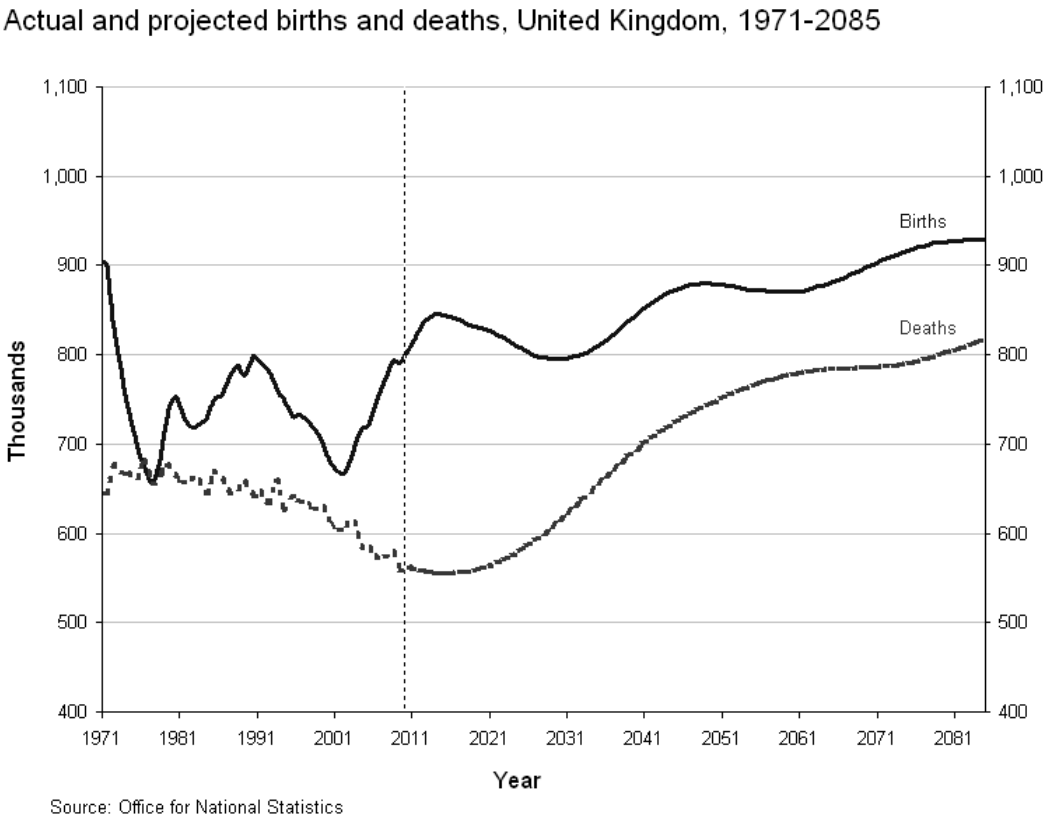

Chart 3

Here is a checklist for evaluating time charts:

- ✓ Examine the scale on the vertical (quantity) axis as well as the horizontal (timeline) axis; results can be made to look more or less dramatic than they actually are simply by changing the scale.
- ✓ Take into account the units used in the chart and be sure they're appropriate for comparison over time.
- ✓ Watch for gaps in the timeline on a time chart. Connecting the dots across a short period of time is better than connecting across a long time.

Bubble plot

A bubble plot is a type of chart that displays three dimensions of data. Each entity with its triplet (v_1, v_2, v_3) of associated data is plotted as a disk that expresses two of the v_i values through the disk's xy location and the third through its size. Bubble charts can be considered a variation of the *scatter plot*, in which the data points are replaced with bubbles.

The circumference of disk's need to be scaled to the corresponding data value v_3 by using this linear relationship between the circumference and the radius:

$$c = \pi . 2r$$

Where

c = the circumference
r = the radius
π (Pi) = the ratio of c/2r

Example: Gapminder World Map

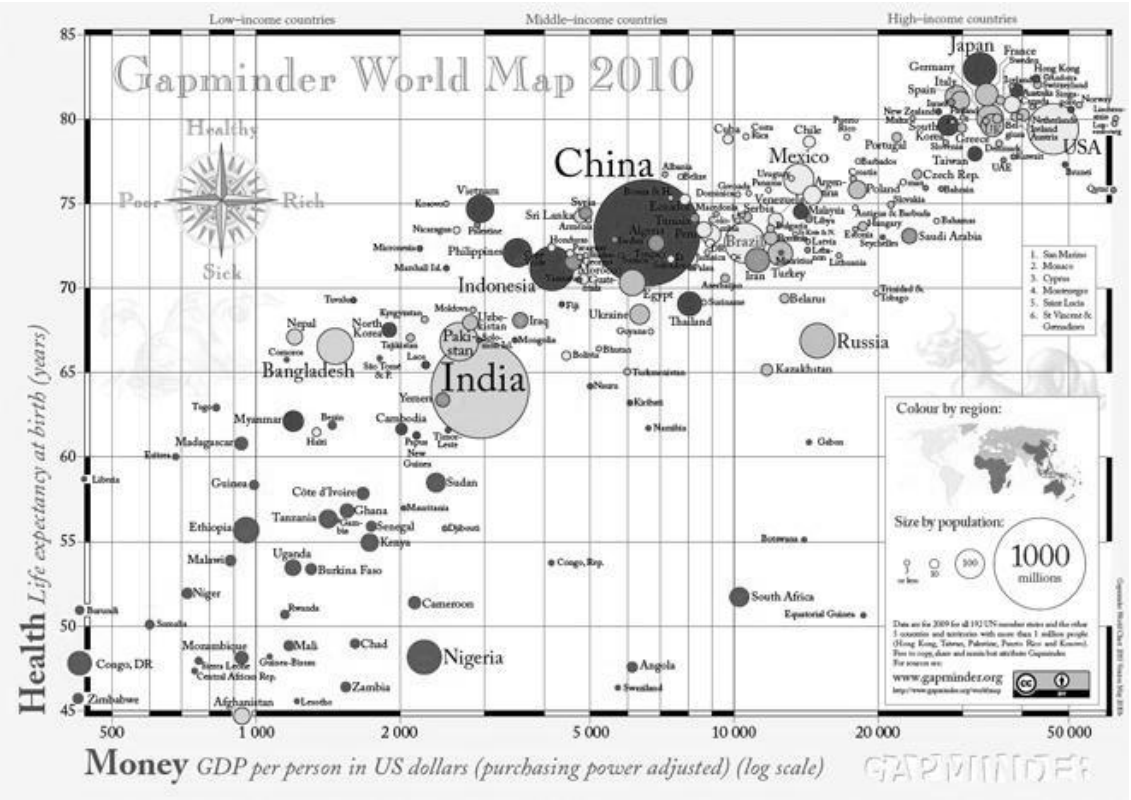

Chart 4 http://www.gapminder.org/

Histograms

For numerical data, a histogram is the statistician's graph of choice. It provides a snapshot of all the data broken down into numerically ordered groups. Histograms provide a quick way to get the big idea about a numerical data set.

Making a histogram

A histogram is basically a bar graph that applies to numerical data. Because the data are numerical, the categories are ordered from smallest to largest. To be sure each number falls into exactly one group, the bars on a histogram touch each other but don't overlap. Each bar is marked on the x-axis (horizontal) by the values representing its beginning and endpoints. The height of each bar represents the number of individuals in each group (the *frequency* of each group).

Table 1 shows the number of live births in Colorado State (U.S.) by age of mother for selected years from 1975–2000. The numerical variable age is broken down into categories of 5-year groupings. Relative frequency histograms comparing 1975 and 2000 are shown in Chart 5. You can see more older mothers in 2000 than in 1975.

Year	*Total births*	*10-14*	*15-19*	*20-24*	*25-29*	*30-34*	*35-39*	*40-44*	*45-49*
1975	40148	88	6627	14533	12565	4885	1211	222	16
1980	49716	57	6530	16642	16081	8349	1842	198	12
1985	55115	90	5634	16242	18065	11231	3464	370	13
1990	53491	91	5975	13118	16352	12444	4772	717	15
1995	54310	134	6462	12935	14286	13186	6184	1071	38
2000	65429	117	7546	15865	17408	15275	7546	1545	93

Table 1: Colorado state (U.S.) live births by mother's age

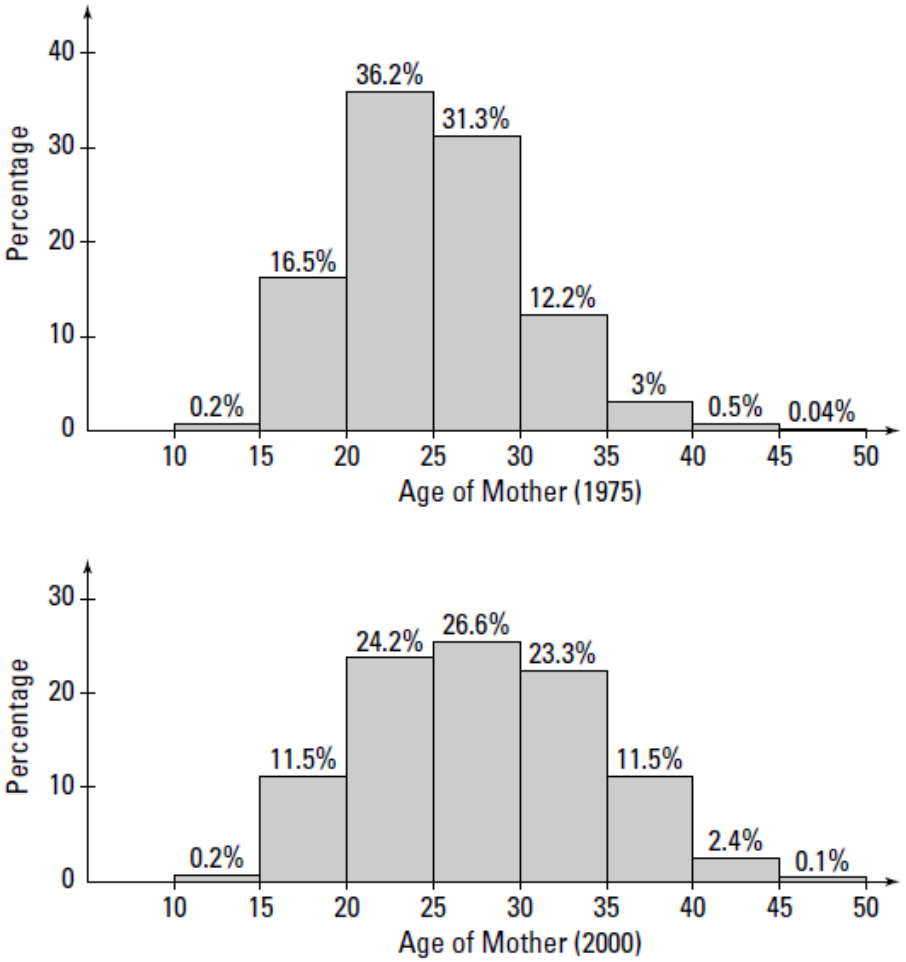

Chart 5: Colorado live births, by age of mother for 1975 and 2000

Interpreting a histogram

A histogram tells you three main features of numerical data:

- ✓ How the data are distributed (symmetric, skewed right, skewed left, bell-shaped, and so on)
- ✓ The amount of variability in the data
- ✓ Where the center of the data is (approximately)

The distribution of the data in a histogram

One of the features that a histogram can show you is the *shape* of the data (how the data are distributed among the groups). Many shapes exist, and many data sets show a combination of shapes, but there are three major shapes to look for:

1. *Symmetric*, meaning that the left-hand side of the histogram is a mirror image of the right-hand side
2. *Skewed right*, meaning that it looks like a lopsided mound with one long tail going off to the right
3. *Skewed left*, meaning that it looks like a lopsided mound with one long tail going off to the left

Mothers' ages in Chart 5 for years 1975 and 2000 appear to be mostly mound-shaped, although the data for 1975 are slightly skewed to the right, indicating that as women got older, fewer had babies relative to the situation in 2000. In other words, in 2000 a higher proportion of older women were having babies compared to 1975.

Variability in the data from a histogram

You can also get a sense of variability in the data by looking at a histogram. If a histogram is quite flat with the bars close to the same height, you may think it indicates less variability, but in fact the opposite is true. That's because you have an equal number in each bar, but the bars themselves represent different ranges of values, so the entire data set is actually quite spread out. A histogram with a big lump in the middle and tails on the sides indicates more data in the middle bars than the outer bars, so the data are actually closer together.

Comparing 1975 to 2000, there's more variability in 2000. This, again, indicates changing times; more women are waiting to have children (in 1975 most women had their children by age 30), and the length of time waiting varies.

Variability in a histogram should not be confused with variability in a *time chart*. If values change over time, they're shown on a time chart as highs and lows, and many changes from high to low (over time) indicate lots of variability. So, a flat line on a time chart indicates no change and no variability in the values across time.

Center of the data from a histogram

A histogram can also give you a rough idea of where the center of the data lies. To visualize the mean, picture the data as people on a see-saw; the mean is the point where the fulcrum has to be in order to balance the weight on each side.

Note in Chart 5 that the mean appears to be around 25 years for 1975 and around 27.5 years for 2000. This suggests that in 2000, Colorado women were having children at older ages, on average, than they did in 1975.

Evaluating a histogram

Here is a checklist for evaluating a histogram:

- ✓ Examine the scale used for the vertical (frequency or relative frequency) axis and beware of results that appear exaggerated or played down through the use of inappropriate scales.
- ✓ Check out the units on the vertical axis to see whether the histogram reports frequencies (numbers) or relative frequencies (percentages), and then take this into account when evaluating the information.
- ✓ Look at the scale used for the groupings of the numerical variable (on the horizontal axis). If the range for each group is very small, the data may look overly volatile. If the ranges are very large, the data may appear to be smoother than they really are.

Boxplots

A *boxplot* is a one-dimensional graph of numerical data based on a five-number summary, which includes the minimum (smallest) value, the 25th percentile (known as Q_1), the median (or 50th percentile), the 75th percentile (Q_3), and the maximum (largest) value in a data set.

Making a boxplot

Consider the following 25 exam scores: 43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, and 99. The five-number summary for these exam scores is 43, 68, 77, 89, and 99, respectively. The vertical version of the boxplot for these exam scores is shown in Chart 6.

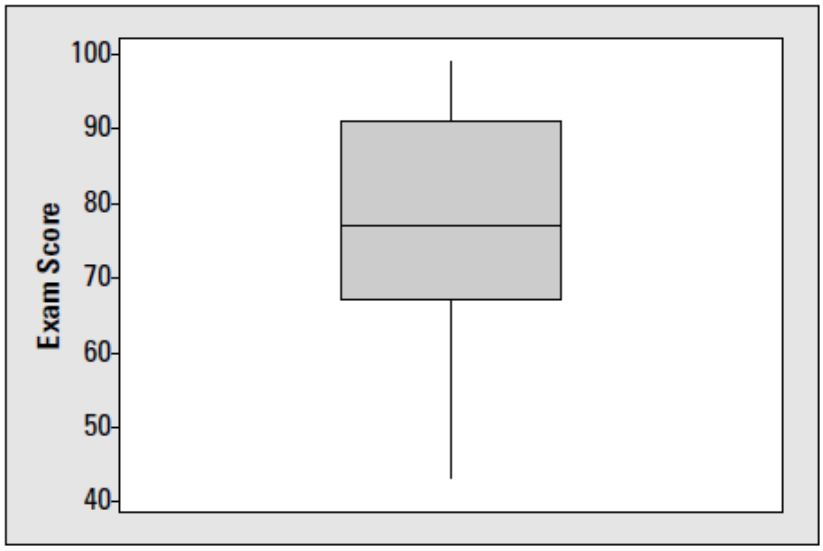

Chart 6: Boxplot of 25 exam scores

Some statistical software adds asterisk signs (*) to show numbers in the data set that are considered to be *outliers* — numbers determined to be far enough away from the rest of the data to be noteworthy.

Interpreting a boxplot

A boxplot can show information about the distribution, variability, and center of a data set.

Distribution of data in a boxplot

A boxplot can show whether a data set is symmetric (roughly the same on each side when cut down the middle), or skewed (lopsided). Symmetric data shows a symmetric boxplot; skewed data show a lopsided boxplot, where the median cuts the box into two unequal pieces. If the longer part of the box is to the right (or above) the median, the data is said to be *skewed right*. If the longer part is to the left (or below) the median, the data is skewed left. However, no data set falls perfectly into one category or the other.

Variability in a data set from a boxplot

Variability in a data set that is described by the five-number summary is measured by the interquartile range (IQR). The interquartile range is equal to $Q_3 - Q_1$. A large distance from the 25th percentile to the 75th indicates the data are more variable. Note that the IQR ignores data below the 25th percentile or above the 75th, which may contain outliers that could inflate the measure of variability of the entire data set. In the exam score data, the IQR is 89 – 68 = 21, compared to the *range* of the entire data set (max – min = 56). This indicates a fairly large spread within the innermost 50% of the exam scores.

Center of the data from a boxplot

The median is part of the five-number summary, and is shown by the line that cuts through the box in the boxplot. This makes it very easy to identify. The mean, however, is not part of the boxplot, and couldn't be determined accurately from a boxplot. In the exam score data, the median is 77. Separate calculations show the mean to be 76.96. These are extremely close, and the reasoning is because the skewness to the right within the middle 50% of the data offsets the skewness to the left of the outer part of the data set.

It's easy to misinterpret a boxplot by thinking the bigger the box, the more data. Remember each of the four sections shown in the boxplot contains an equal percentage (25%) of the data. A bigger part of the box means there is more *variability* (a wider range of values) in that part of the box, not more data.