Informational Graphics

GC

Motivation: effective information transmission

- begin by looking carefully
  - weather maps
  - tables
  - exercise: stock quotations, grade lists,
- numeric graphs
  - stem 'n leaf diagrams
  - histograms
    - relative, absolute
  - time series
  - bar charts
    - segmented bar charts
  - pie charts
  - box 'n whisker diagrams
- tables vs. graphs

Distributions

GC

DISTRIBUTION is the pattern of variation of values

Example: 'Old Farmers Almanac' reports "growth seasons"—the number of (consecutive) frost-free days. Here is data from $1901 - 1957$ for a southern US city.

279	244	318	262	335	321	165	180	201	252
145	192	217	179	182	210	271	302	169	192
156	181	156	125	166	248	198	220	134	189
141	142	211	196	169	237	136	203	184	224
178	279	201	173	252	149	229	300	217	203
148	220	175	188	160	176	128			

we can categorize these and note frequency

frost-free days	years
101 - 150	9
151 - 200	22
201 - 250	15
251 - 300	6
301 - 355	5

rough, and choice of categories may be important

Stem 'n Leaf

GC

stem 'n leaf diagram organizes data around most significant (leftmost) digit. Same data, ignoring rightmost digit.

```
1  |  2233444445566666777778888889999
2  |  00001111222344556777
3  |  00123
```

frost-free days	years
101 - 150	9
151 - 200	22
201 - 250	15
251 - 300	6
301 - 355	5

stem n' leaf

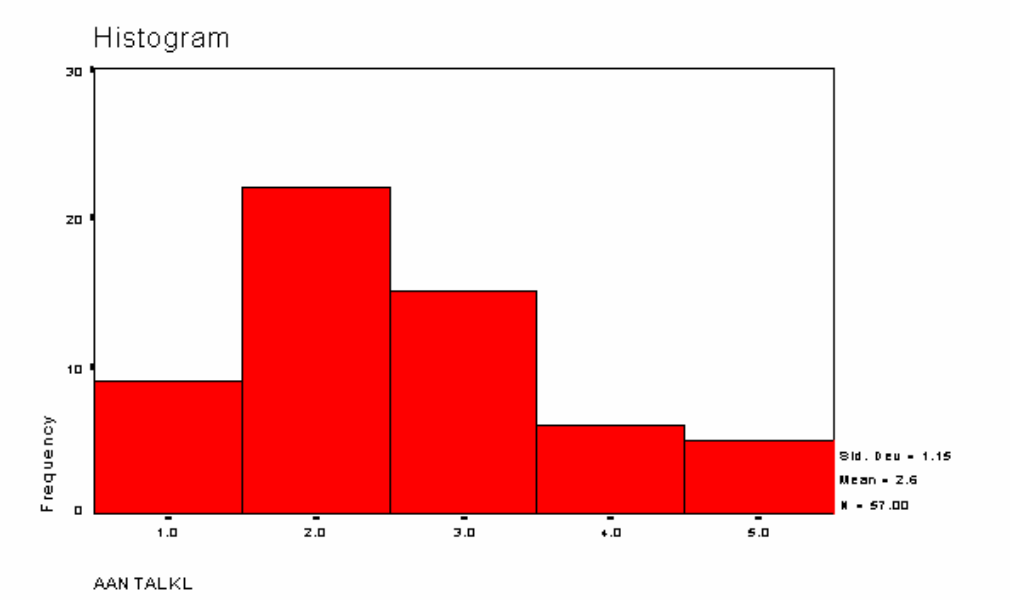
- allows finer discrimination
- less sensitive to category choice (more categories may be used)
- distribution *reflected visually* in shape of diagram
- clumsy if too many datapoints

Histogram

GC

Histogram divides a distribution into a small number of ranges and shows how frequently values in the different ranges appear

Histogram of length of growth seasons (classes as above); labeling here *terrible*.



care needed!

- show entire distribution
- don't use more than about 10 categories
- *absolute* histogram shows numbers
  relative histogram shows percentages

RuG

Histogram vs. Stem n' Leaf

both

- show entire distribution
- allow view of outliers, symmetry/skewness, general patterns

histogram

- ineffective w. fine categorization
- provides exact numbers directly

Stem n' Leaf

- effective w. fine categorization
- less directly interpretable (need to count!)

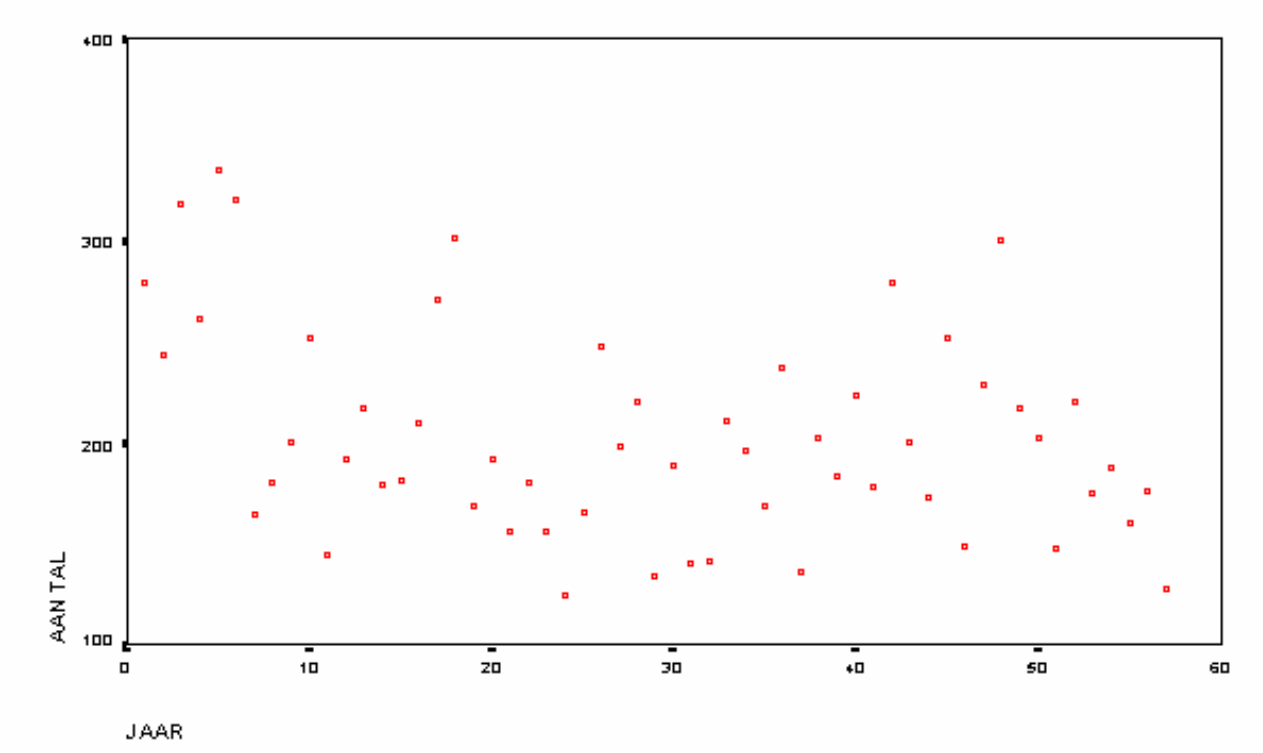
Time Series

GC

time series measurement of same variable at regular intervals e.g., indices, unemployment rate, orders placed,...

change often focus of attention

Growth seasons shown as time series.



Excellent investigative technique!

Review ("Cross-Tables")

GC

CROSS-TABLES show frequencies of nonnumeric variables by value—values of one var. horizontally against the values of another vertically.

```
NL_CLASS  class of ability  by  SEX
       Count
       %          SEX  male    female

                                      Row
                         1        2       Total
NL_CLASS   ---------+--------+--------+
                 0   |    3   |   3    |   6
  beginner           |        |        |  15.0
                     +--------|--------+
                 1   |   12   |   6    |  18
  advanced beginner  |        |        |  45.0
                     +--------|--------+
                 2   |    6   |   7    |  13
  intermediate       |        |        |  32.5
                     +--------|--------+
                 3   |    2   |   1    |   3
  advanced           |        |        |   7.5
                     +--------+--------+
            Column       23       17       40
             Total       57.5     42.5    100.0
```

Each cell shows the number of cases with the combination of values. Upper left shows that 3 beginners are male, etc.

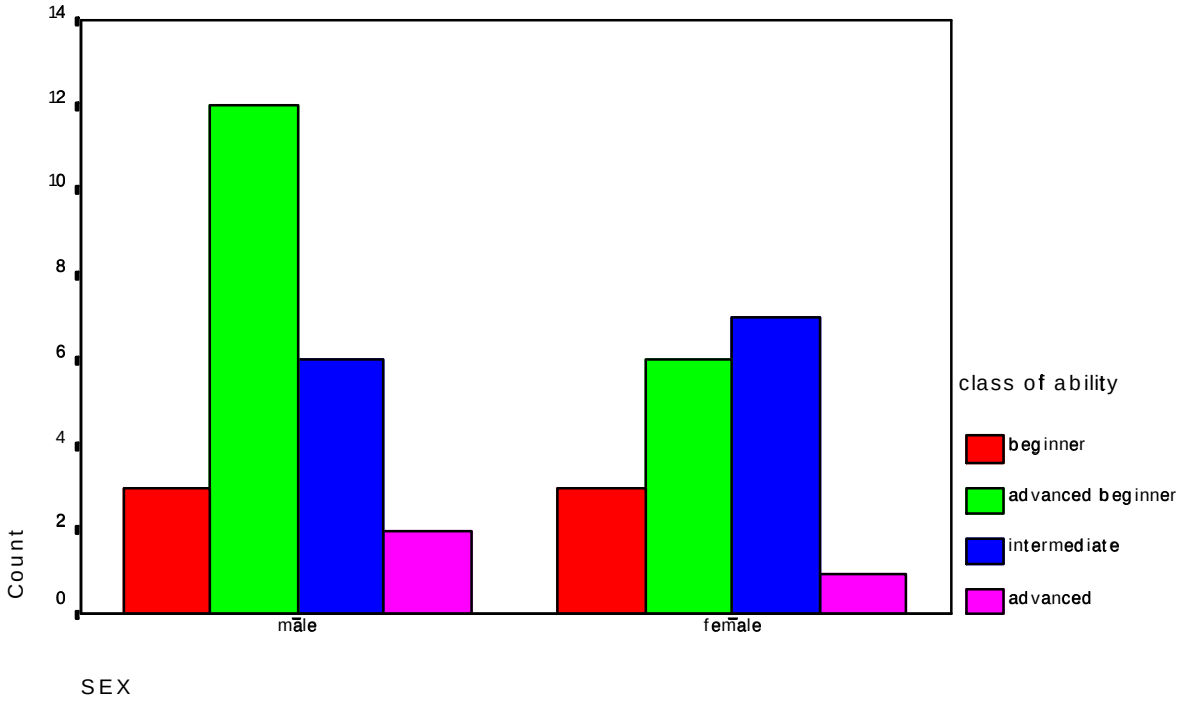
Presentation restructured and summed, but original data not lost.

Numeric alternatives?

Side-by-Side Histograms

GC

We can visualize the information in the cross table by creating two histograms, one restricted to male, one to female.

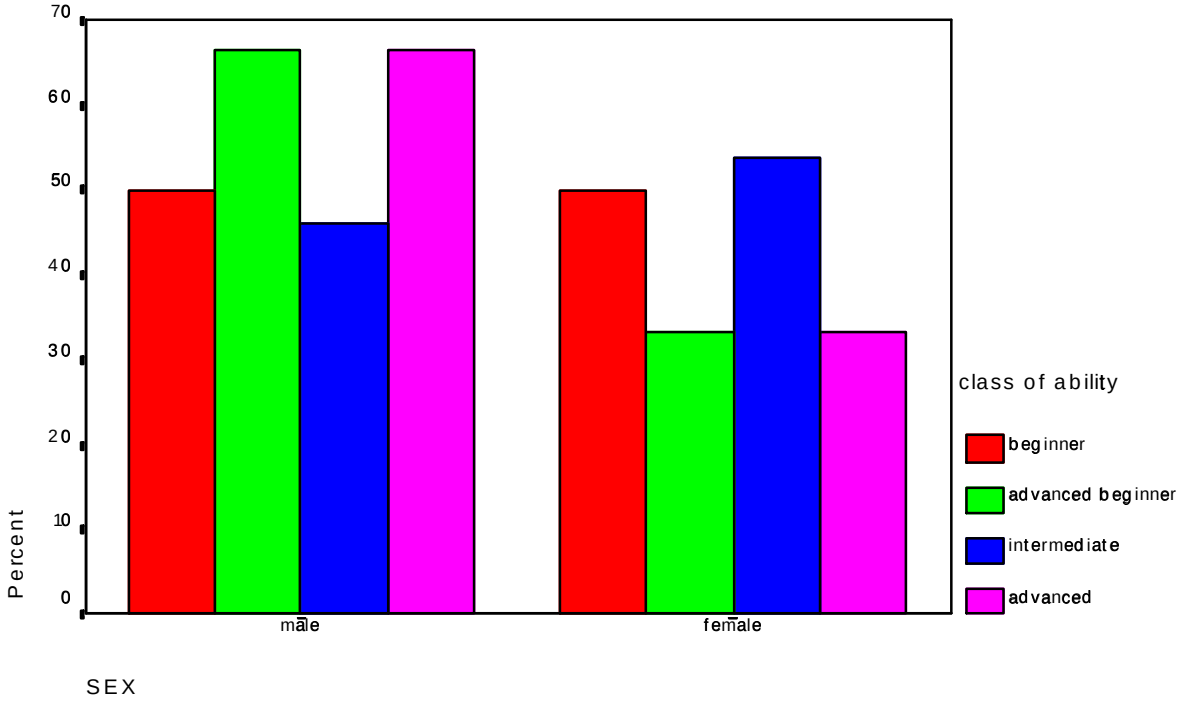


Note that this retains all original information as well. In particular, frequencies are shown (graphically).

Relative Histograms

GC

Relative histograms, showing percentages **hide** some data (the frequencies).

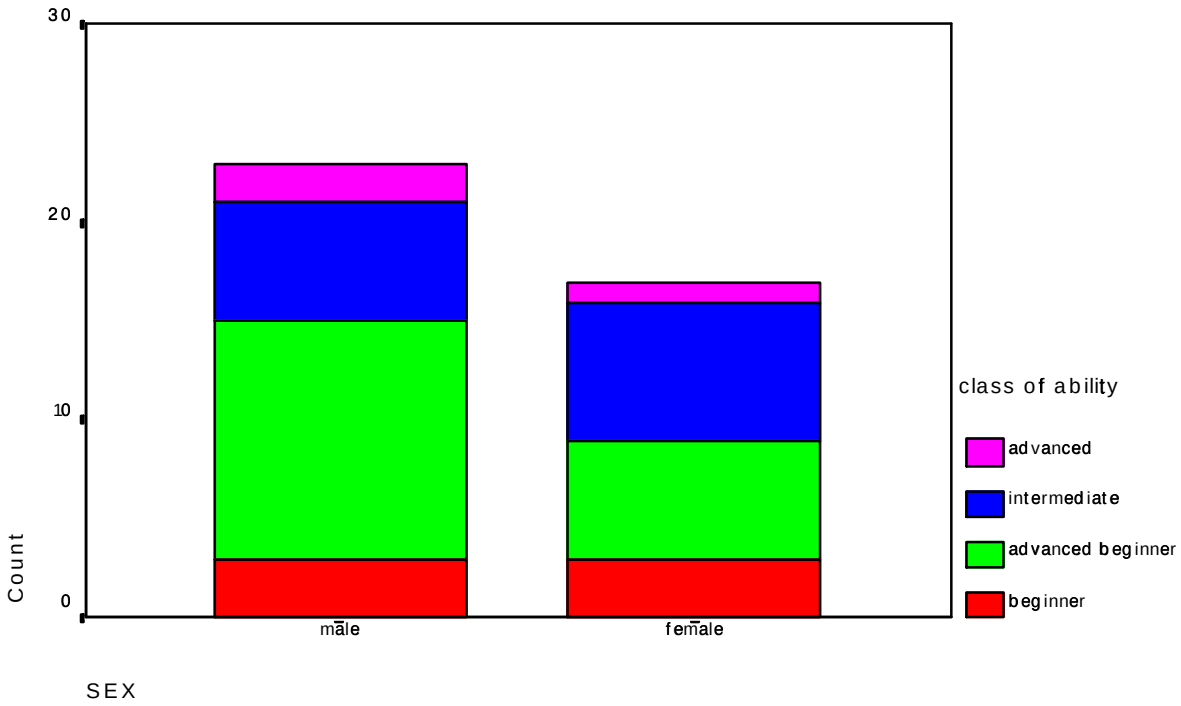


Note the "advanced" column, showing that ca. $2/3$ were male, $1/3$ female—but hiding that 2 were male, and 1 was female.

Relative histograms appropriate when *rates* are significant.

Segmented Bar Charts

SEGMENTED BAR CHARTS show the same information as side-by-side histograms, but show more directly overall differences in the classes compared (male and female).



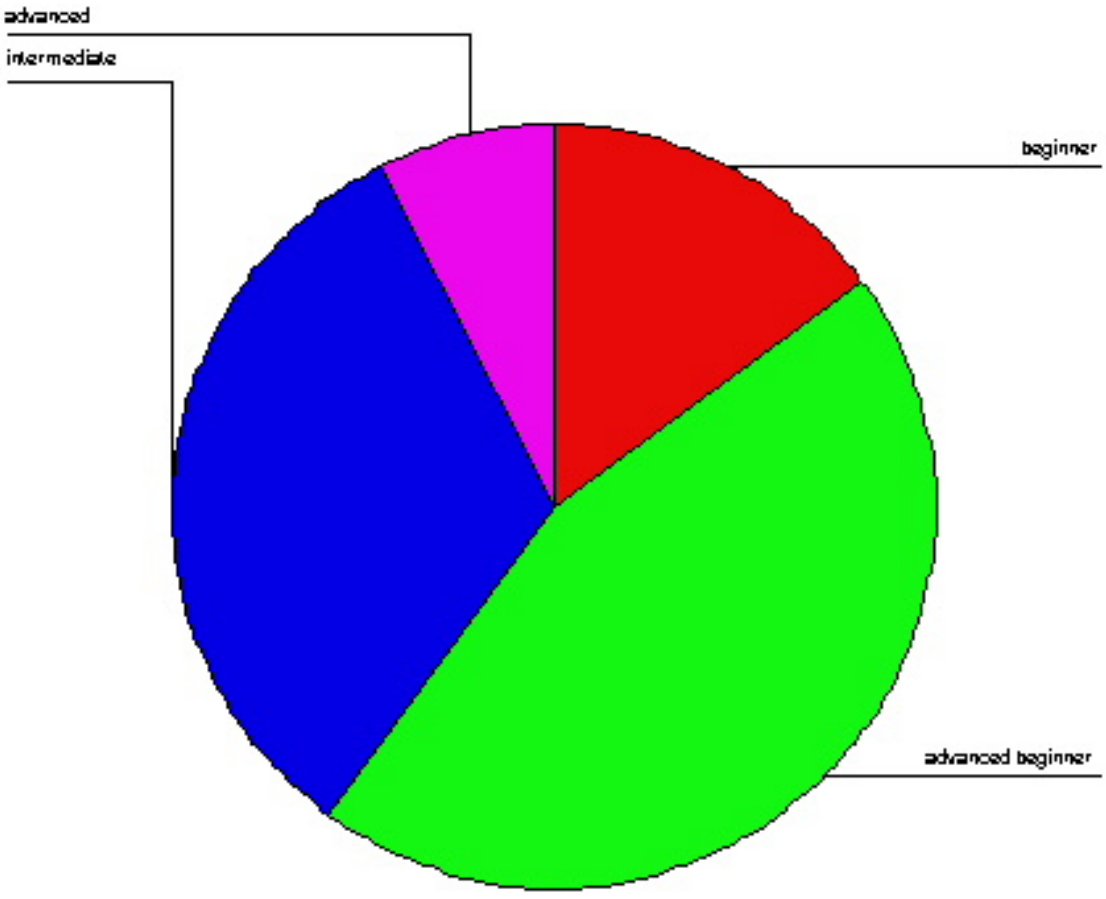
It is immediately clear here that *more* men than women were involved, and also that women outperformed men in the largest classes (advanced beginner and intermediate).

Segmented bar charts most recommended means of visualizing relation between nominal variables.

Pie Charts

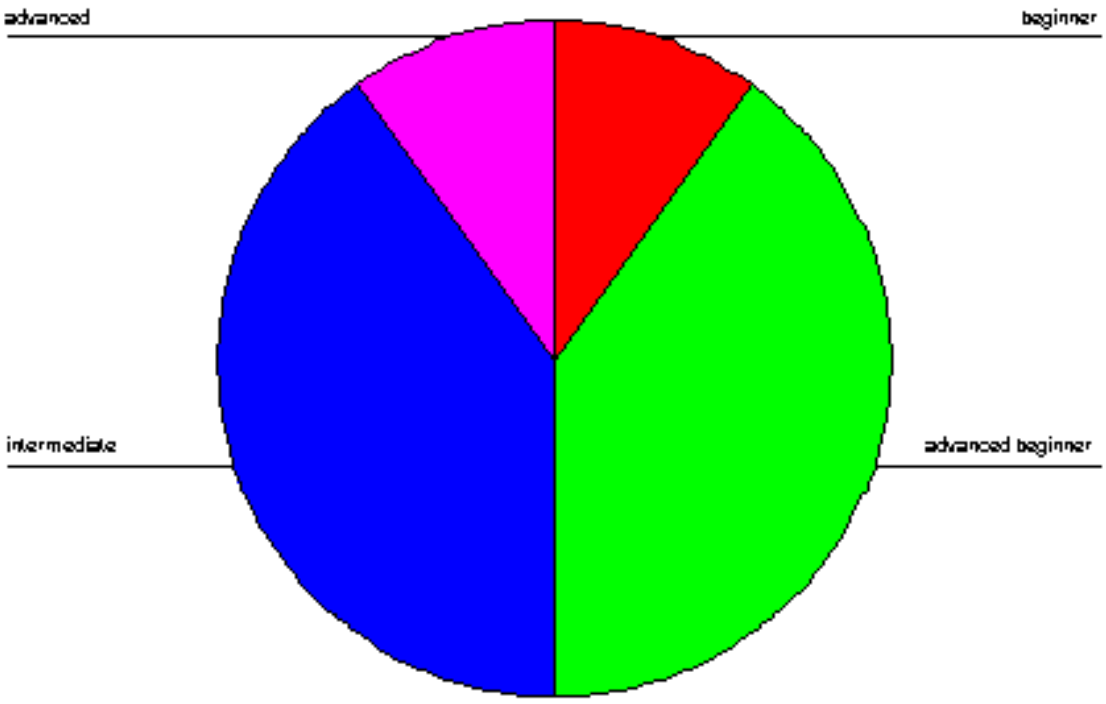
pie charts show relative distribution as portions of a "pie"

Dutch for Foreigners – Levels (All)



Pie Charts

Dutch for Foreigners – Levels (European Students)

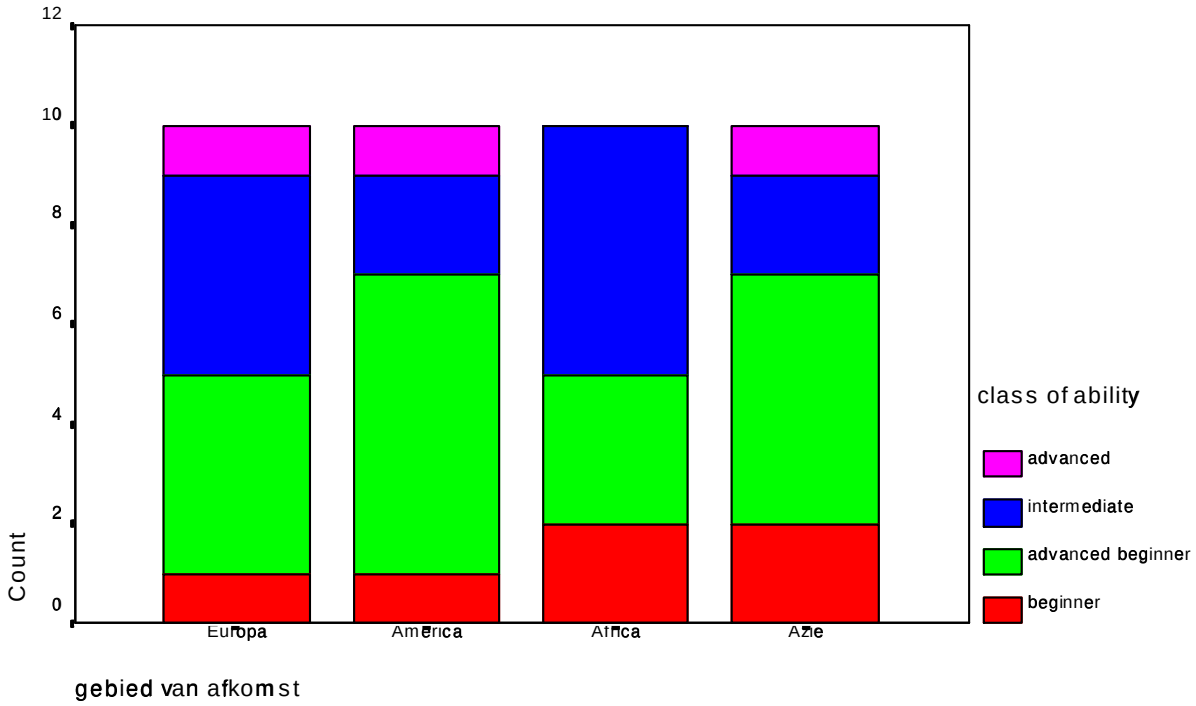


How to compare two subgroups? in absolute & relative size?

Tufte: "most overused" form of graph

Area and Ability

To examine now the relation between area of origin and language proficiency level in the test, we may examine a segmented bar chart (and a cross-table, of course).



But recall the language proficiency was originally a numeric variable, which we classified for convenience.

There are other ways of analyzing the relation between one numeric and one nonnumeric variable.

Numeric Values

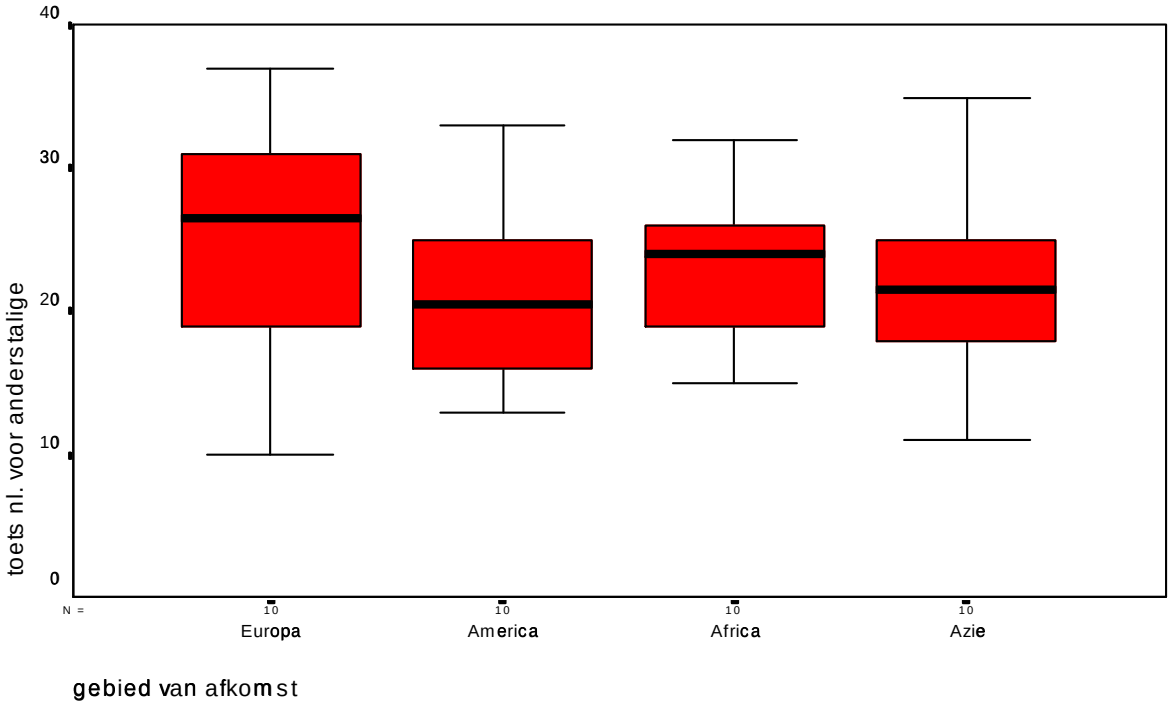
Numeric values such as age, income, weight, height, . . . have

- high and low values
- central values (average or median)
- central range of values

Box 'n Whisker diagrams show these.

Box 'n Whiskers Diagrams

GC



compares various "areas of origin"

Each of the four box 'n whiskers diagrams

- shows center (median) — middle line
- shows middle 50% of distribution — in box
- shows extreme 50% of distribution — in whiskers

favorite means of display for many statisticians

RuG

Box 'n Whiskers Diagrams

The box 'n whiskers display reveals more of the numeric structure than the tables, stem 'n leaf diagrams, histograms, pie charts, and segmented bar charts.

But it also hides the exact numeric scores — the designer must be reponsible for this. Often, there is too much data for tabular display, which justifies the summary immediately.

It uses (multiple) numeric displays to compare the cases of the nominal variable.