Applied Statistics

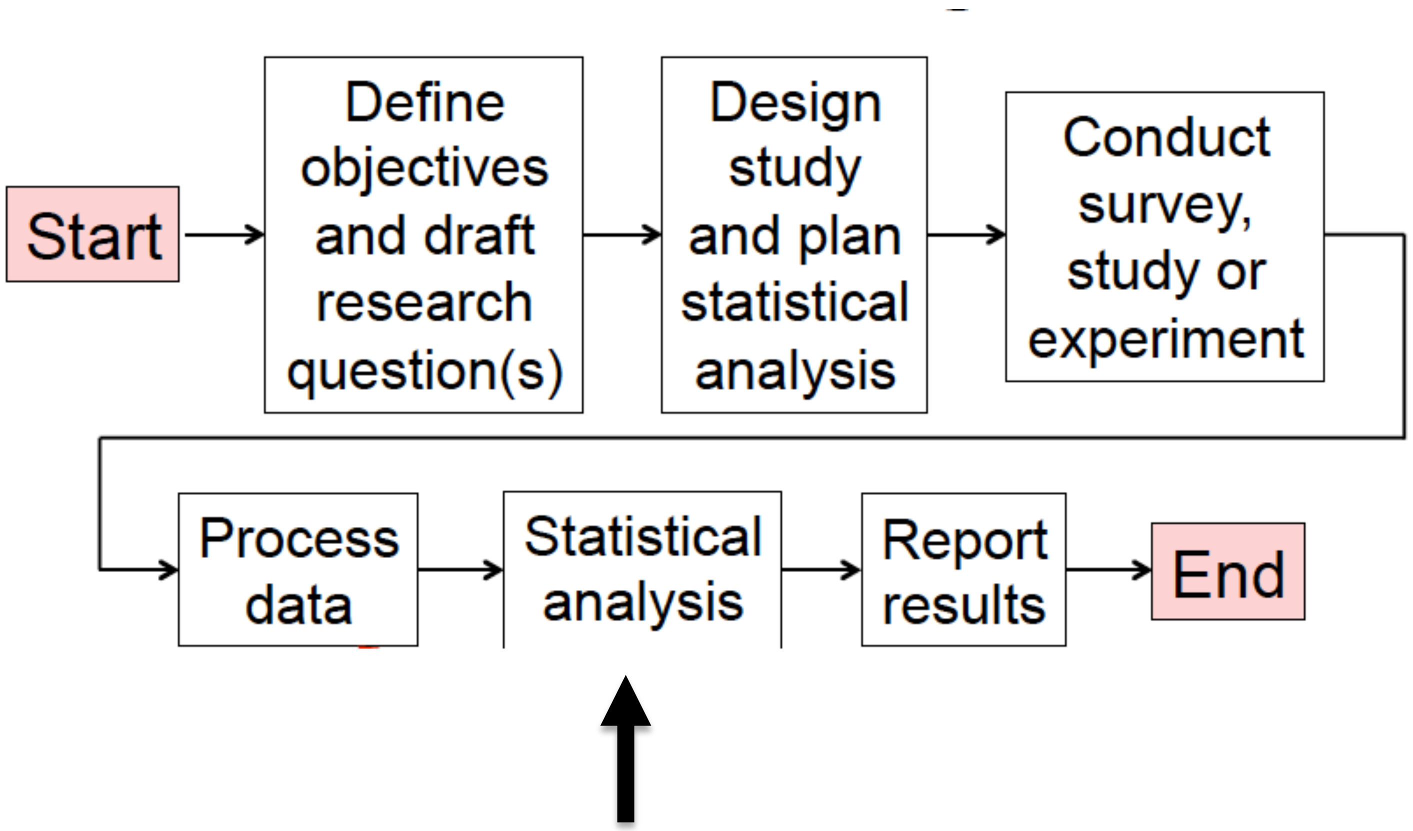
Feb 4th 2021

Prof.ssa Chiara Seghieri,

Laboratorio di Management e Sanità, Istituto di Management,
Scuola Superiore Sant'Anna, Pisa
c.seghieri@santannapisa.it

EXPLORING AND SUMMARIZING DATA

The research study process


Start
Define objectives and draft research question(s)
Design study and plan statistical analysis
Conduct survey, study or experiment
Process data
Statistical analysis
Report results
End

Summarize Data to Reveal Meaningful Information, Patterns, and Relationships

- How you do this depends on the nature of the data, e.g.,
  - nominal, ordinal, etc.

- One variable at a time (*Univariate Analysis*)
- Two variables at a time (*Bivariate Analysis*)
- Multiple variables at a time (*Multivariate Analysis)*

- Two stages:
  - reduce the data to a single relatively compact *table* (**frequency table**, crosstabulation, control table, etc.) or corresponding ***chart*** (bar graph, histogram, dot chart, box chart, scattergram, etc.)
  - reduce it further, if possible and depending on the nature of the variable, to one or several ***summary statistical measures*** (measures of central tendency, dispersion, correlation, etc.).
- We first look at the process of summarizing *data* down to *frequency tables*, *bar graphs*, and *histograms*.
  - Then (univariate) measures of central tendency and dispersion.

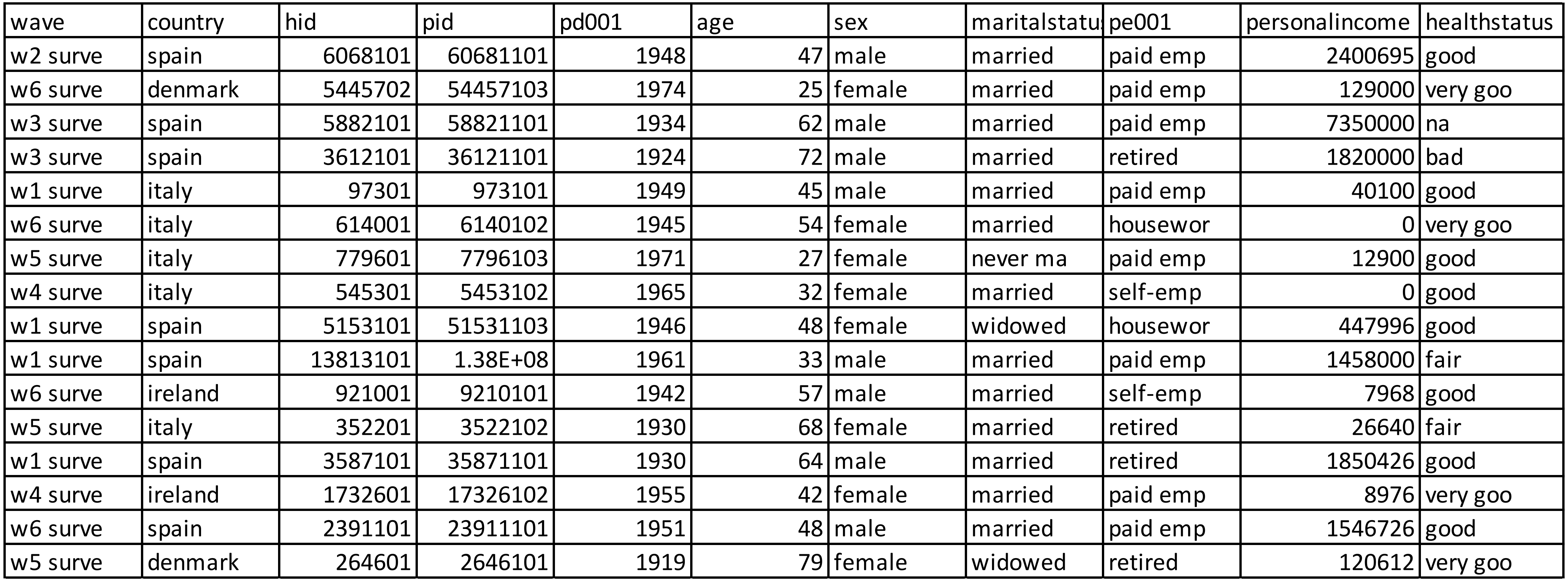
We first look at the process of summarizing *data* down to *frequency tables*, *bar graphs*, and *histograms*.

- Then (univariate) measures of central tendency and dispersion.

Frequency tables

- After collecting data, the first task for a researcher is to organize and simplify the data so that it is possible to get a general overview of the results.
- This is the goal of descriptive statistical techniques.
- One method for simplifying and organizing data is to construct a **frequency table**.

Data Matrix (nxp): individual (person) level



wave	country	hid	pid	pd001	age	sex	maritalstatu	pe001	personalincome	healthstatus
w2 surve	spain	6068101	60681101	1948	47	male	married	paid emp	2400695	good
w6 surve	denmark	5445702	54457103	1974	25	female	married	paid emp	129000	very goo
w3 surve	spain	5882101	58821101	1934	62	male	married	paid emp	7350000	na
w3 surve	spain	3612101	36121101	1924	72	male	married	retired	1820000	bad
w1 surve	italy	97301	973101	1949	45	male	married	paid emp	40100	good
w6 surve	italy	614001	6140102	1945	54	female	married	housewor	0	very goo
w5 surve	italy	779601	7796103	1971	27	female	never ma	paid emp	12900	good
w4 surve	italy	545301	5453102	1965	32	female	married	self-emp	0	good
w1 surve	spain	5153101	51531103	1946	48	female	widowed	housewor	447996	good
w1 surve	spain	13813101	1.38E+08	1961	33	male	married	paid emp	1458000	fair
w6 surve	ireland	921001	9210101	1942	57	male	married	self-emp	7968	good
w5 surve	italy	352201	3522102	1930	68	female	married	retired	26640	fair
w1 surve	spain	3587101	35871101	1930	64	male	married	retired	1850426	good
w4 surve	ireland	1732601	17326102	1955	42	female	married	paid emp	8976	very goo
w6 surve	spain	2391101	23911101	1951	48	male	married	paid emp	1546726	good
w5 surve	denmark	264601	2646101	1919	79	female	widowed	retired	120612	very goo

n =1000 individuals (sample size) on the rows
p=number of variables on the colums

Notation

$X =$ variable

$n =$ sample size

$k =$ num of values of X

$x_i =$ value i of X

$n_i =$ absolute frequency of xi

$f_i =$ relative frequency of xi

X	n_i	$f_i = n_i / n$
x_1	n_1	f_1
x_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k
Totale	n	1

tab sex

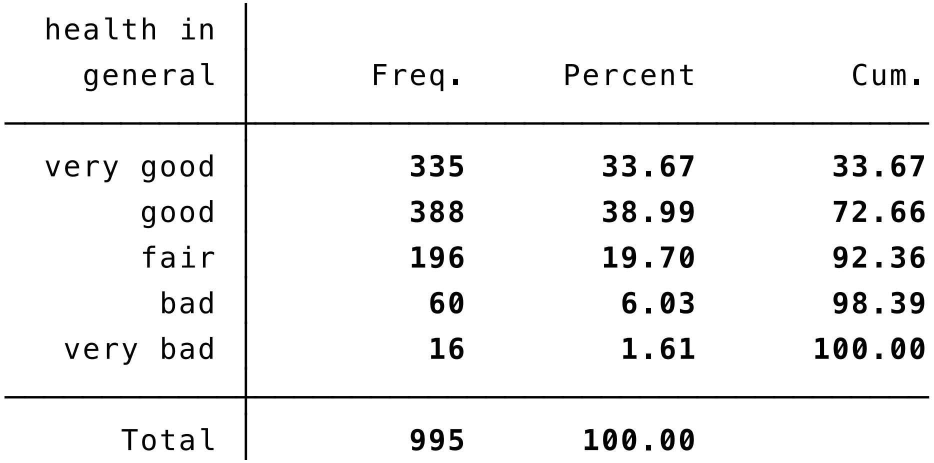
```
                 n: absolute   f: relative    cumulative
                     freq       freq (%)       freq (%)

     sex of │
 individual │      Freq.     Percent        Cum.
────────────┼───────────────────────────────────
       male │        457       45.70       45.70
     female │        543       54.30      100.00
────────────┼───────────────────────────────────
      Total │      1,000      100.00
```

tab pd005

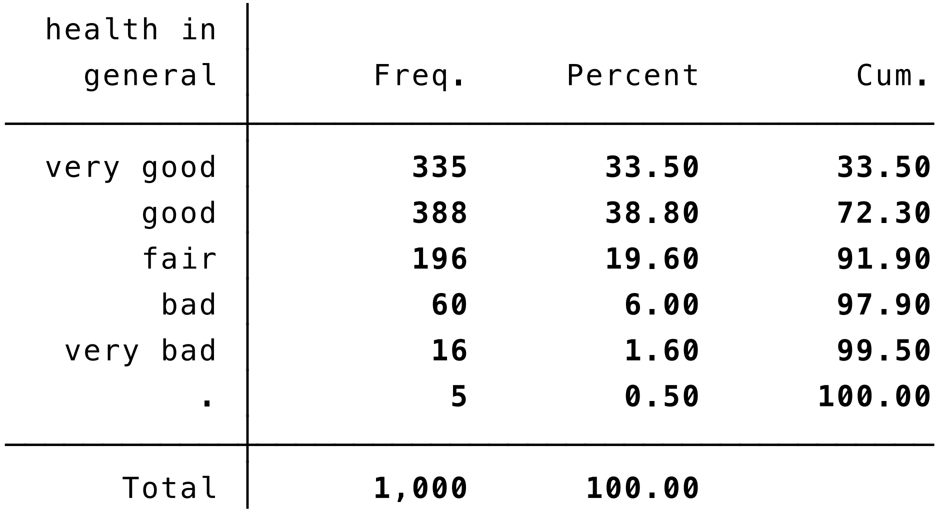
```
     marital │
      status │      Freq.     Percent        Cum.
─────────────┼───────────────────────────────────
     married │        584       58.40       58.40  =f1
   separated │         20        2.00       60.40  =f1+f2
    divorced │         17        1.70       62.10  =f1+f2+f3
     widowed │         80        8.00       70.10   =........
never married │        299       29.90      100.00
─────────────┼───────────────────────────────────
       Total │      1,000      100.00
```

tab ph001



```
  health in |
    general |      Freq.     Percent        Cum.
------------+-----------------------------------
  very good |        335       33.67       33.67
       good |        388       38.99       72.66
       fair |        196       19.70       92.36
        bad |         60        6.03       98.39
   very bad |         16        1.61      100.00
------------+-----------------------------------
      Total |        995      100.00
```

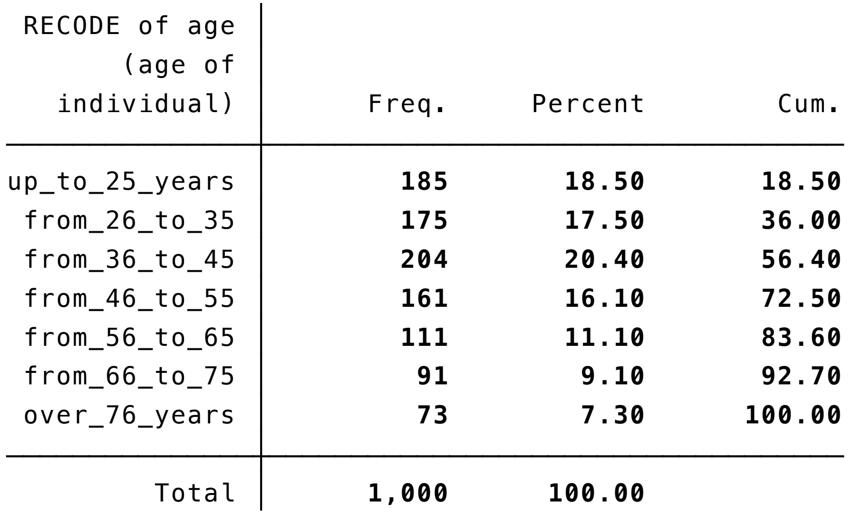
tab ph001, m



```
  health in |
    general |      Freq.     Percent        Cum.
------------+-----------------------------------
  very good |        335       33.50       33.50
       good |        388       38.80       72.30
       fair |        196       19.60       91.90
        bad |         60        6.00       97.90
   very bad |         16        1.60       99.50
          . |          5        0.50      100.00
------------+-----------------------------------
      Total |      1,000      100.00
```

Age is a continuous variable, we need to create a new variable in which age is divided in classes!

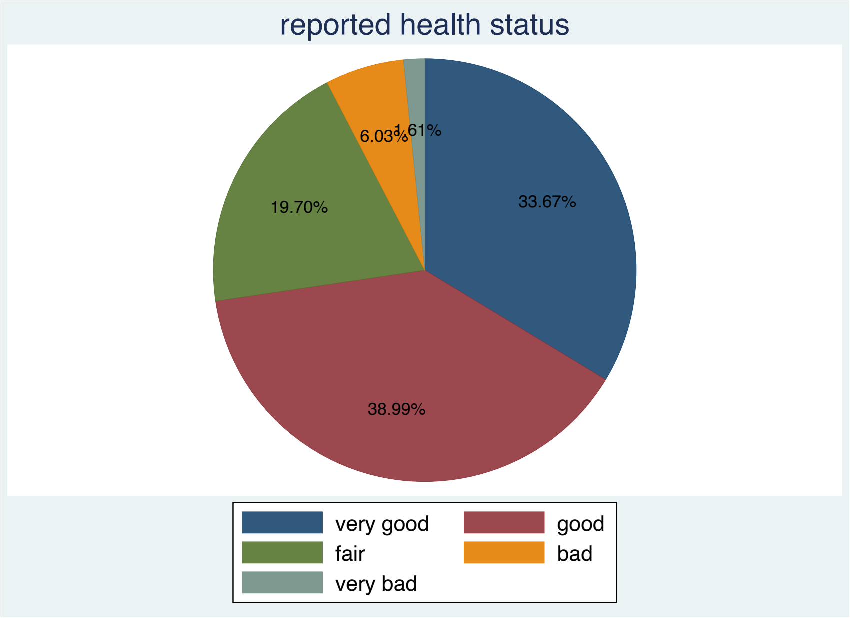
tab agegr

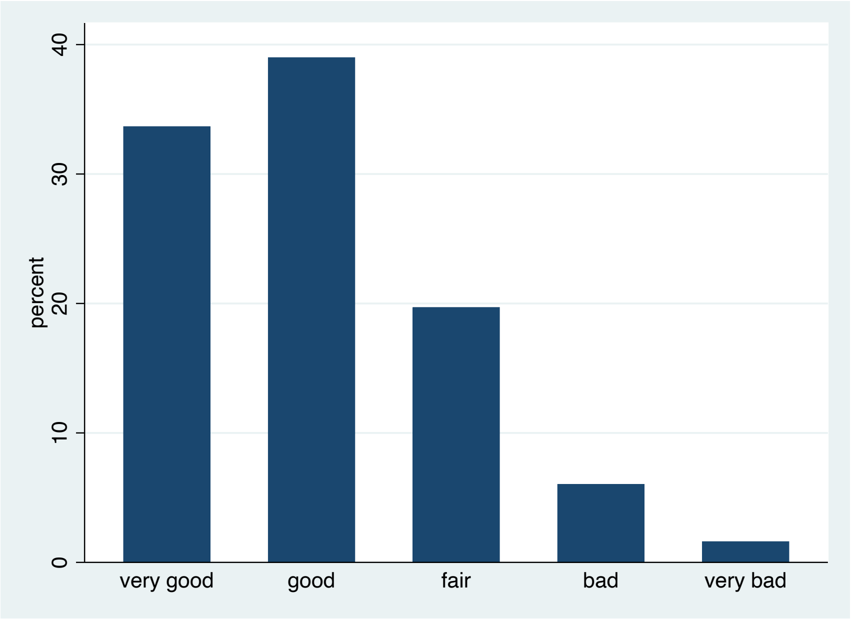


```
RECODE of age |
      (age of |
  individual) |      Freq.     Percent        Cum.
--------------+-----------------------------------
up_to_25_years|        185       18.50       18.50
 from_26_to_35|        175       17.50       36.00
 from_36_to_45|        204       20.40       56.40
 from_46_to_55|        161       16.10       72.50
 from_56_to_65|        111       11.10       83.60
 from_66_to_75|         91        9.10       92.70
 over_76_years|         73        7.30      100.00
--------------+-----------------------------------
        Total |      1,000      100.00
```

Pie Charts and Bar Charts

```
   health in │
     general │      Freq.     Percent        Cum.
─────────────┼───────────────────────────────────
   very good │        335       33.67       33.67
        good │        388       38.99       72.66
        fair │        196       19.70       92.36
         bad │         60        6.03       98.39
    very bad │         16        1.61      100.00
─────────────┼───────────────────────────────────
       Total │        995      100.00
```



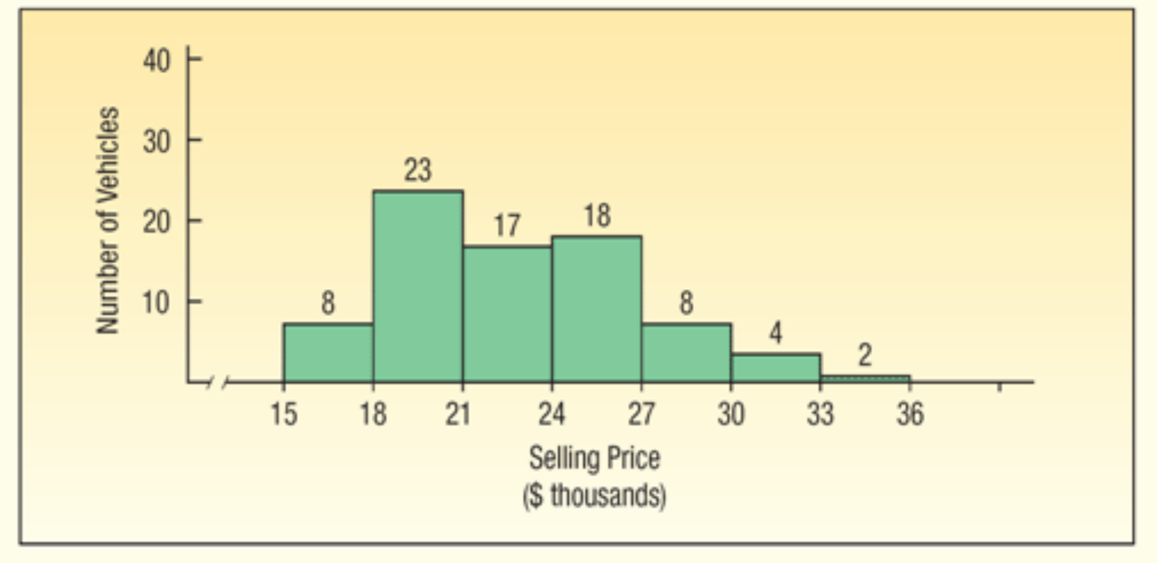


for qualitative variables!

the frequencies are on the vertical axis and are proportional to the heights of the bars

Histogram

A graph in which the classes are marked on the horizontal axis and the class frequencies on the vertical axis. The class frequencies are represented by the heights of the bars and the bars are drawn adjacent to each other.

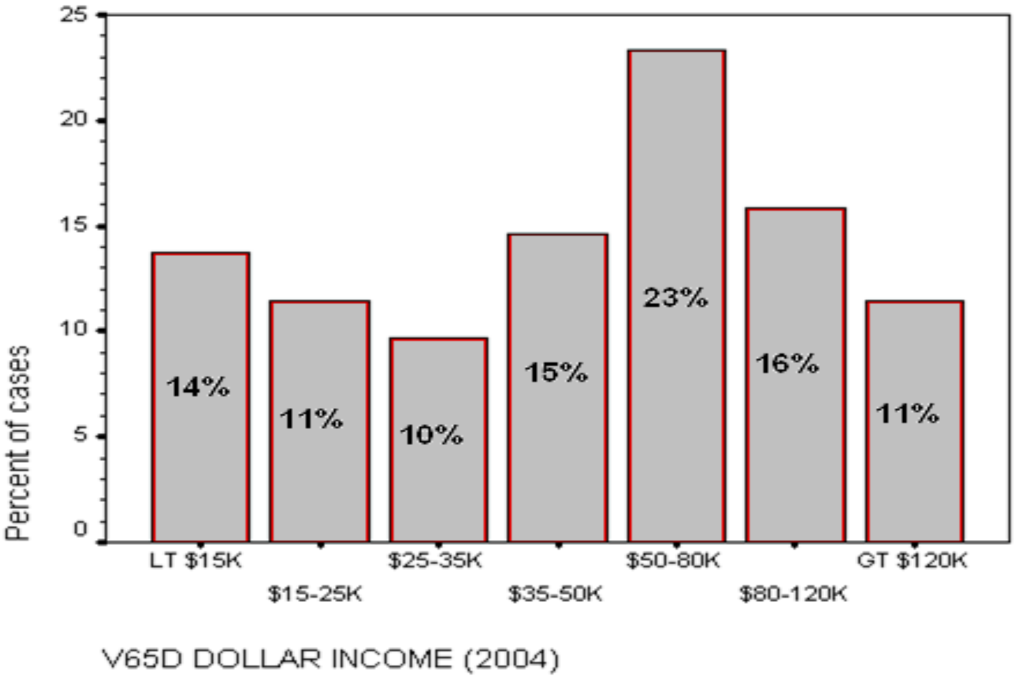


Frequency table for income

		Freq.	Percent	Valid %	Cum. %
	Less than $15,000	145	12.0	13.7	13.7
	$15,000 to $25,000	121	10.0	11.4	25.2
	$25,000 to $35,000	102	8.4	9.7	34.9
	$35,000 to $50,000	154	12.7	14.6	49.5
	$50,000 to $80,000	246	20.3	23.3	72.8
	$80,000 to $120,000	167	13.8	15.8	88.6
	More than $120,000	120	9.9	11.4	100.0
	Total	1055	87.0	100.0	
Missing	NA	157	13.0		
Total		1212	100.0		

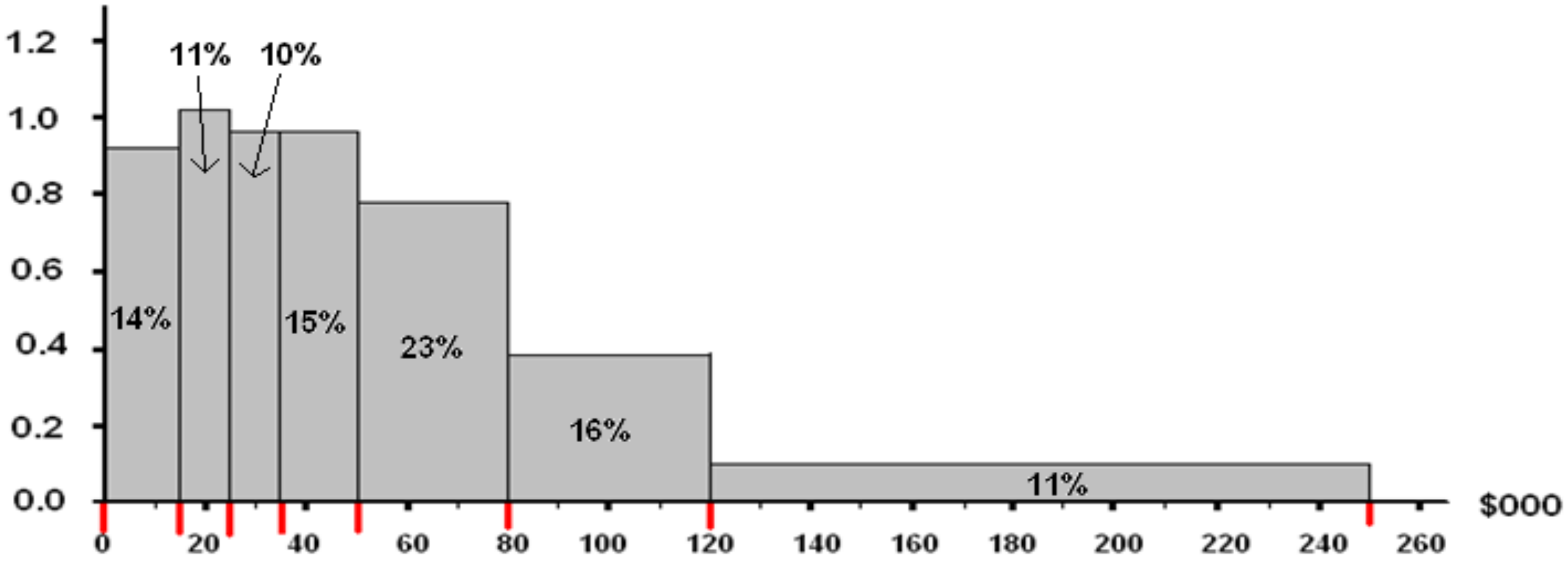
example from SPSS

Bar Chart



- The bar chart appears to display a distribution of income that is approximately "uniform" – that is, all bars are approximately the same height, except for a distinctive peak (or "mode") in the third highest income category.
  - Indeed, the impression the bar graph conveys to the eye is that there are more well-off than not-so-well-off people.
- However, this impression is quite misleading, as you can begin to understand when you look more closely at the income class intervals and notice that they are not of equal width.
- Here is the histogram of the same INCOME data =>

Histogram



- The fundamental difference between a bar graph and a histogram:
  - in a bar graph, *frequency is represented by the* <u>*height*</u> *of the bars* (all of which have the same width);
  - in a histogram, *frequency is represented by the* <u>*area*</u> *of the "bars"* (which may have different widths, reflecting the different "widths" of the class intervals).
- With equal class intervals, the area of a bar depends only on its height, so Histogram ≈ Frequency Bar Chart
- But with unequal class intervals, the area of a bar depends on both its height and its width, so Histogram ≠ Frequency Bar Chart

Histogram (cont.)

- the *area* [*not* height] of each rectangle is proportional to the frequency associated with that class interval.
- How tall should each rectangle be?
- The width of each rectangle is the width of the class interval, and you should remember that:

  Area = *Height* × *Width* so *Height* = *Area* / *Width*
- Since *Area* here represents *Frequency*, we have the formula:

  Height = *Frequency* / *Width*,

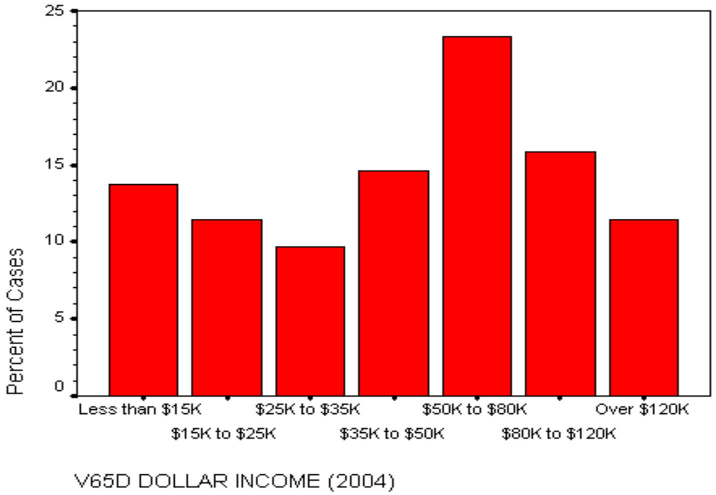
  where *Width* is the width of the class interval.

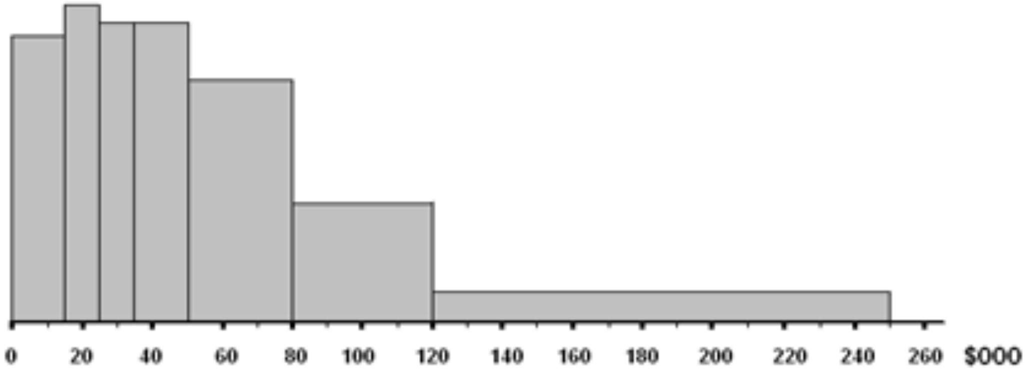
Histogram (cont.)

- Now we can calculate the following (relative) heights of all the bars/rectangles. (Since only relative magnitudes matter, we can ignore the $000 = $K in INCOME values.)

Class Interval	*Width*	*Freq.*	Freq/Width		*Height*
0-15	15	13.7	13.7 / 15	=	0.913
15-25	10	11.4	11.4 / 10	=	1.140
25-35	10	9.7	9.7 / 10	=	0.970
35-50	15	14.6	14.6 / 15	=	0.973
50-80	30	23.3	23.3 / 30	=	0.777
80-120	40	15.8	15.8 / 40	=	0.395
120-250	130	11.4	11.4/130	=	0.088

- Now we can draw the appropriate scale on the vertical axis.
- The tallest rectangle has a height of about 1.14





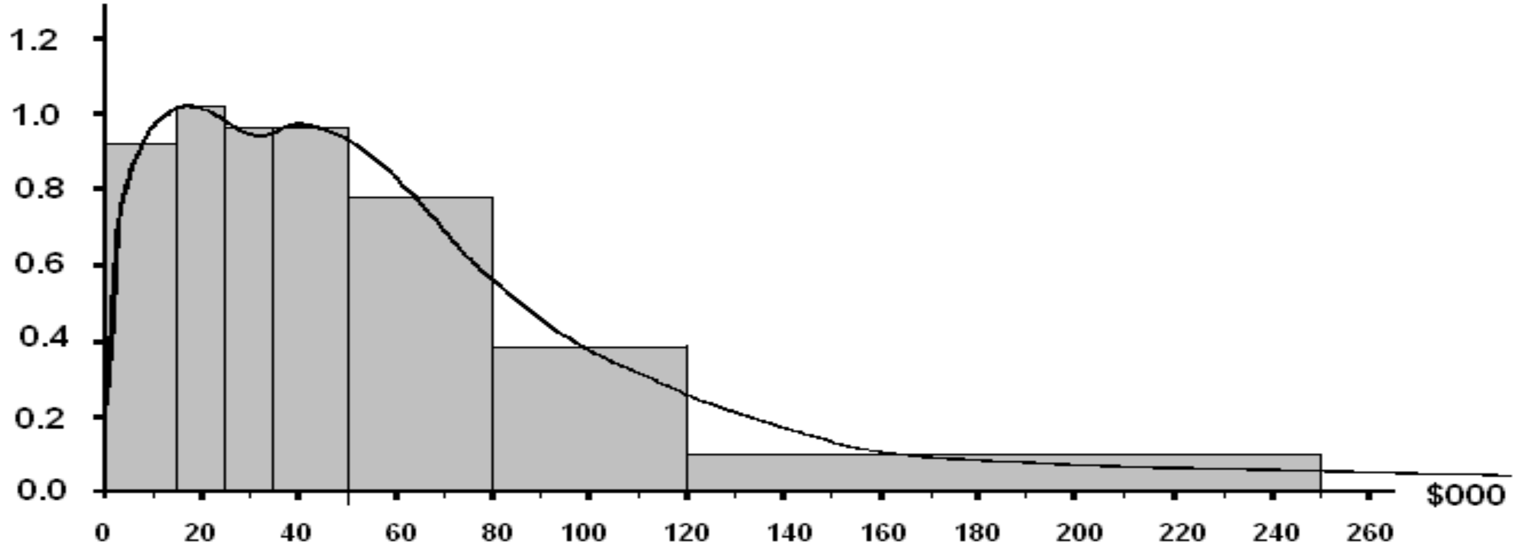
- Given that height in a histogram does *not* represent frequency, what does it represent?
- The answer is that height represents *density* — that is, *how densly observed values of cases are "packed into" each class interval*.
  - Note that the class interval $50-80K includes about twice as many cases (23.3%) as the interval $15-25K (11.4%).
  - This fact is reflected in the *bar graph* in Figure 1 by the fact that the bar on the $50-80K interval is about *twice as high* as the bar over the $15-25K interval.
  - It is reflected in the histogram in Figure 2 by the fact that the "bar" (rectangle) on the $50-80K interval has about *twice the area* of the bar on the $15-25K interval.
  - But the 23.3% of the cases in the $50-80K interval are spread over an income interval that is three times as wide as the interval into which the 11.4% of the cases in the $15-25K interval are packed, so the height of the former (wide) bar is actually less than the height of the latter (thin) bar.

Histogram vs. Frequency Bar Graph

- If *all class intervals all have the same width*, then the histogram is essentially no different from a bar chart

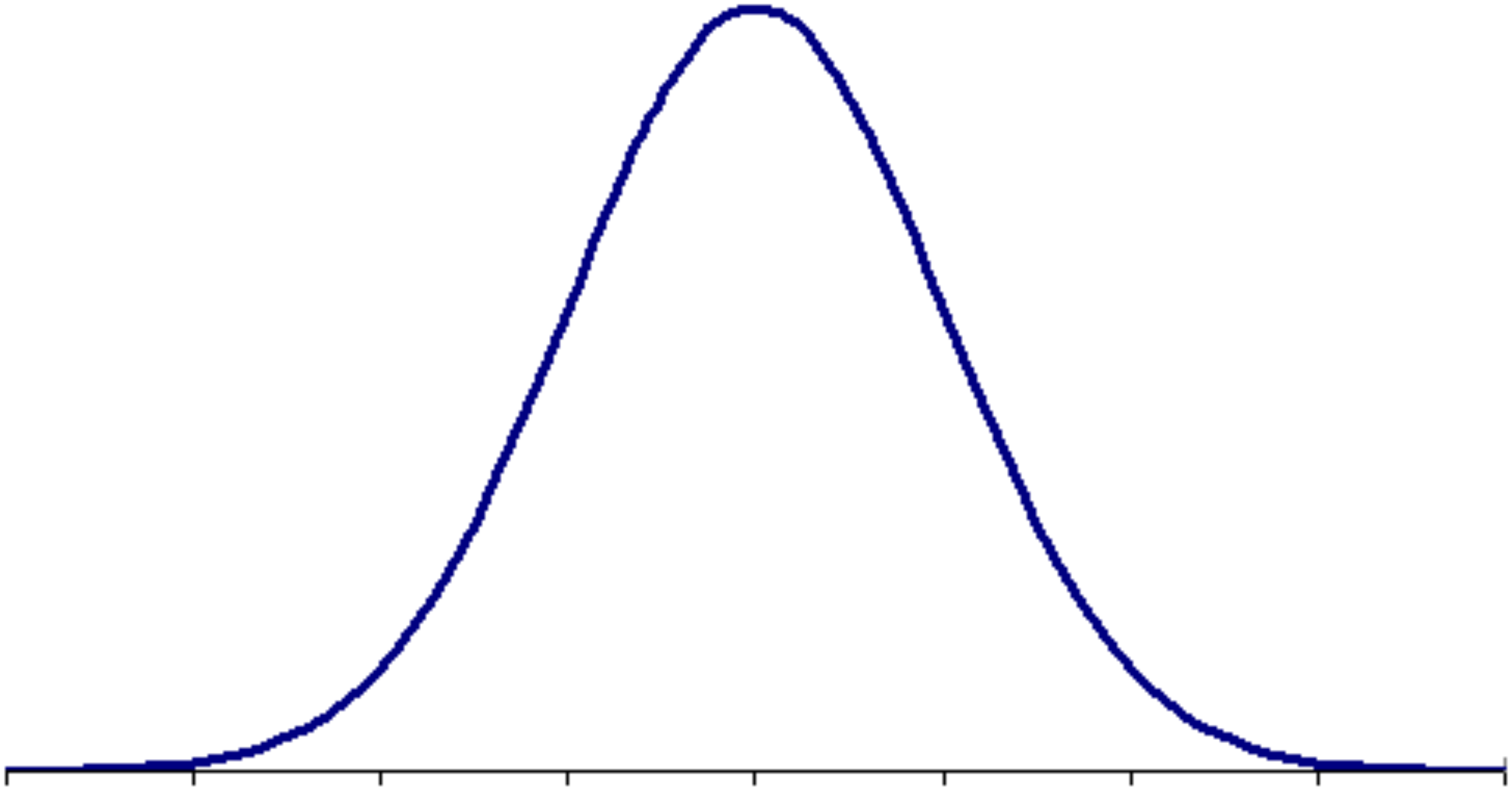
- Otherwise (i.e., if the class intervals are not all of equal width), a bar chart and a histogram of the same data may look quite different,
  - in which event the bar chart presents a misleading picture of the data,
  - while the histogram presents a more accurate picture.
- The histogram, unlike the bar chart takes account of the *interval* property of the variable.

A Continuous Income Density Curve ["Eyeball estimate"]

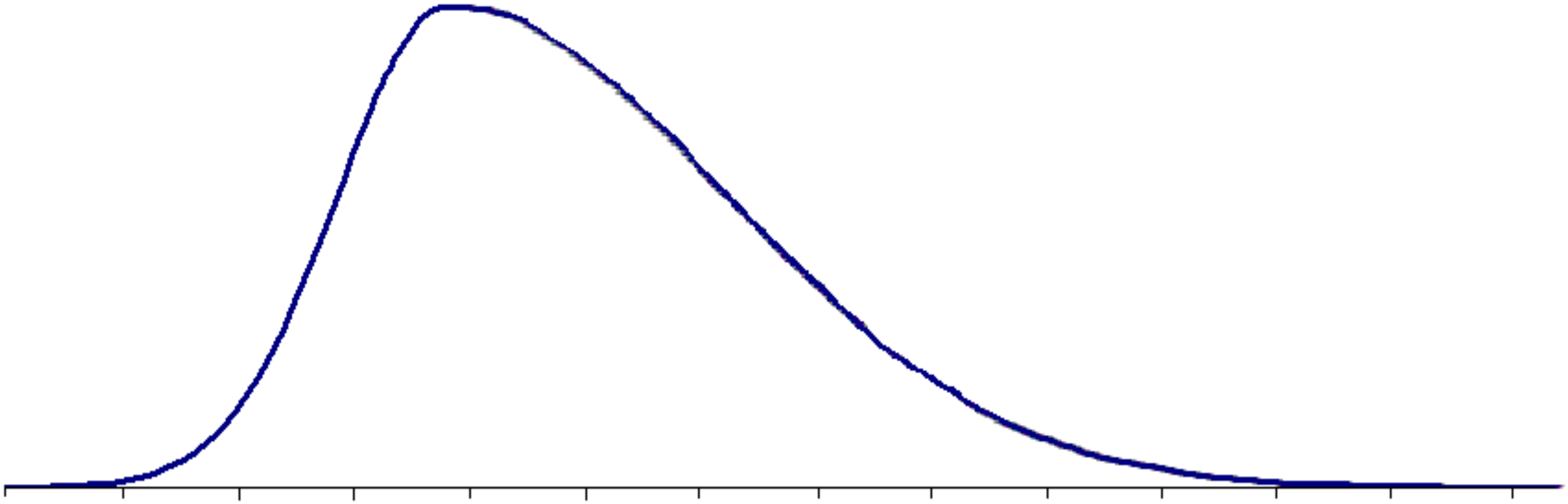


- Contrary to the (hypothetical) continuous density curve, INCOME data suggests that the distribution of household income has two "peaks" (or *modes*), one at about \$18K and another at about \$43K, with a slight "valley" between them.
- This probably results from the fact that there are two types of households: family or multi-person households (typically two or more adults and often children as well) and single-person households (typically widows/widowers or young adults). On average, the former type of household has (and needs) higher income than the latter. This tends to produce two peaks in the overall distribution of household income.

A Symmetric [Normal] Density Curve



An Asymmetric Density Curve



Key Concept

A histogram is an important type of graph that portrays the nature of the distribution.

Two-way frequency tables – bivariate analysis

Also known as *contingency tables*, crosstabs help you to analyze the relationship between two or more categorical variables

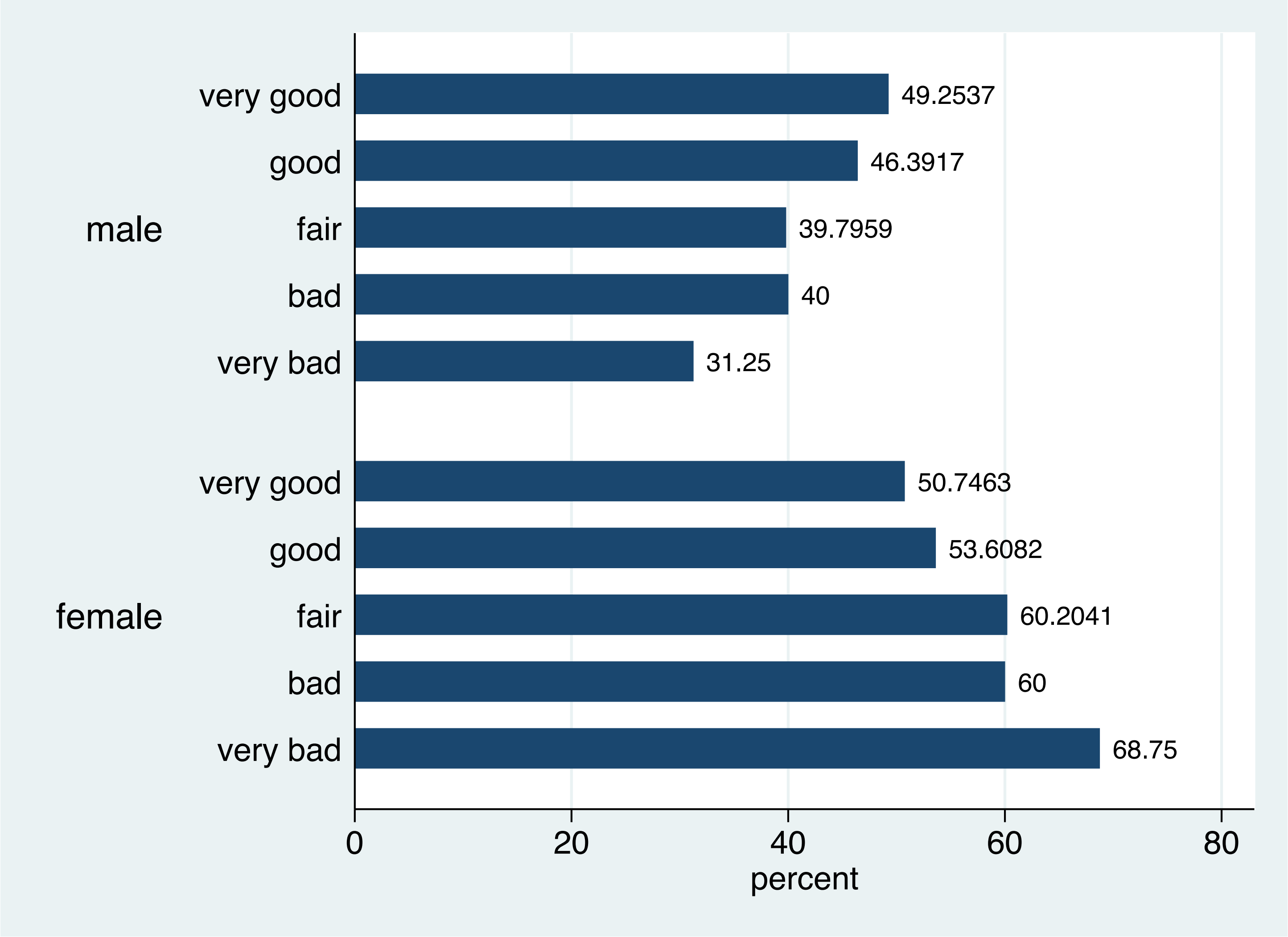
tab ph001 sex, row col

```
 health in |   sex of individual
   general |      male     female |     Total
-----------+----------------------+----------
 very good |       165        170 |       335
           |     49.25      50.75 |    100.00
           |     36.50      31.31 |     33.67
-----------+----------------------+----------
      good |       180        208 |       388
           |     46.39      53.61 |    100.00
           |     39.82      38.31 |     38.99
-----------+----------------------+----------
      fair |        78        118 |       196
           |     39.80      60.20 |    100.00
           |     17.26      21.73 |     19.70
-----------+----------------------+----------
       bad |        24         36 |        60
           |     40.00      60.00 |    100.00
           |      5.31       6.63 |      6.03
-----------+----------------------+----------
  very bad |         5         11 |        16
           |     31.25      68.75 |    100.00
           |      1.11       2.03 |      1.61
-----------+----------------------+----------
     Total |       452        543 |       995
           |     45.43      54.57 |    100.00
           |    100.00     100.00 |    100.00
```

The first value in a cell: the number of observations for each xtab. In this case, 165 respondents are ‘male’ and reported to be in a ‘very good’ health status, 170 are ‘female’ and and reported to be in a ‘very good’ status.

The second value in a cell: row percentages for the first variable in the xtab. Out of those who report to be in ‘very good’ health status, 49.25% are males and 50.75% are females.

The third value in a cell: column percentages for the second variable in the xtab. Among males, 36.50% report a ‘very good’ health while 31.31% of females report a ‘very good’ health *catplot ph001 sex, percent(ph001) blabel(bar)*


male
very good
good
fair
bad
very bad
female
very good
good
fair
bad
very bad
percent