Applied Statistics for the Behavioral Sciences

Chapter 2

Frequency Distributions and Graphs

Chapter 2 Outline

- Organization of Data
  - Simple Frequency Distributions
  - Grouped Frequency Distributions
- Graphs
  - Frequency polygon
  - Histogram
  - Bar graphs
- Describing Distributions

Moving beyond raw data

- Unorganized/interpretable
- Imposing organization
- Ordering
- N = total number of observations (scores)
- *f* = frequency of each score

Simple frequency distribution

- find highest and lowest values
- write all the numbers between these values
  - descending order
  - headed by the variable name
- work through all unorganized scores placing tally mark in row with that score

Simple frequency dist. (cont.)

- count tally marks at the end
- put this number in a column labeled (*f*) for frequency
- verify that you got all scores
  - add up all the values in the (*f*) column
  - they should sum to N.

Simple frequency distributions

- useful because:
  - highest and lowest values clear
  - can figure out vicinity of central tendency
- final version no tally marks or scores not in the data set

Grouped frequency distributions

- difficult to extract information about the form of the distribution from simple frequency distribution
- too many pieces of information to try to consider simultaneously.
- *i* = class interval

Class intervals

- intervals - ranges of scores that we use to tally the observations in the dataset
- Class intervals always equal sized
- have an upper and lower limit just like scores
- add half unit to upper-bound
- subtract half a unit from the lower bound

Class intervals (cont.)

- interval: 21-23, has an actual range from 20.5-23.5
- midpoint of each interval is the value used to represent all scores falling in interval
- selecting intervals 4 steps (selecting and organizing intervals)

Class interval steps

- 1. The number of class intervals should be between 10 and 20
- 2. Choose a convenient size for (*i*)
  - Odd numbers nice because midpoint of interval will be a whole number
  - 3, 5, 10, and multiples of 10 are all popular interval sizes.
- 3. Begin class intervals with a multiple of *i*
- 4. Largest scores go at the top

Grouped freq. dist. steps

- 1. Find the highest and lowest scores
- 2. Determine range:lower limit of low score from upper limit of high score.
- 3. Determine *i*
  - divide range by possible values of *i*
  - gives you the number of intervals
  - want 10-20 class intervals and convenient interval size
- 4. Begin bottom interval with a multiple of *i*, end interval with number that makes scores in the interval = *i,* next interval at previous interval's start value + *i*
- 5. Tally marks are placed according to ranges

Candy Calories/Serving Raw Data

310	230	220	170	200	260	190	190	230	230	200	390
250	200	200	220	280	160	240	340	200	230	250	270
240	220	260	280	230	420	450	250	240	250	250	160
280	160	230	170	260	250	190	240	410	230	280	230
210	250	125	262	220	220	200	230	230	230	280	190
230	210	250	210	250	230	410	240	230	200	190	410
240	210	230									

Candy Calories/Serving Ordered Data

450	420	410	410	410	390	340	310	280	280	280	280
280	270	262	260	260	260	250	250	250	250	250	250
250	250	250	240	240	240	240	240	240	230	230	230
230	230	230	230	230	230	230	230	230	230	230	230
220	220	220	220	220	210	210	210	210	200	200	200
200	200	200	200	190	190	190	190	190	170	170	160
160	160	125									

Simple Frequency Table

Calories	*f*	Calories	*f*
450	1	250	9
420	1	240	6
410	3	230	15
390	1	220	5
340	1	210	4
310	1	200	7
280	5	190	5
270	1	170	2
262	1	160	3
260	3	125	1
		N =	75

Grouped Frequency Table

Calories (class interval)	Midpoint (X)	*f*
475-499	487	0
450-474	462	1
425-449	437	1
400-424	412	3
375-399	387	1
350-374	362	0
325-349	337	1
300-324	312	1
275-299	287	5
250-274	262	14
225-249	237	21
200-224	212	16
175-199	187	5
150-174	162	5
125-149	137	1
100-124	112	0

Frequency table from SPSS

calories

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	125.00	1	1.3	1.3	1.3
	160.00	3	4.0	4.0	5.3
	170.00	2	2.7	2.7	8.0
	190.00	5	6.7	6.7	14.7
	200.00	7	9.3	9.3	24.0
	210.00	4	5.3	5.3	29.3
	220.00	5	6.7	6.7	36.0
	230.00	15	20.0	20.0	56.0
	240.00	6	8.0	8.0	64.0
	250.00	9	12.0	12.0	76.0
	260.00	3	4.0	4.0	80.0
	262.00	1	1.3	1.3	81.3
	270.00	1	1.3	1.3	82.7
	280.00	5	6.7	6.7	89.3
	310.00	1	1.3	1.3	90.7
	340.00	1	1.3	1.3	92.0
	390.00	1	1.3	1.3	93.3
	410.00	3	4.0	4.0	97.3
	420.00	1	1.3	1.3	98.7
	450.00	1	1.3	1.3	100.0
	Total	75	100.0	100.0	

Graphical presentations of data

- large datasets make graphical presentation of data more necessary
- picture worth 1000 words in summarizing patterns in data.
- Parts of a basic graph
  - X axis (horizontal, abscissa, baseline)
  - Y axis (vertical, ordinate)
  - positive directions are upward and to the right.

Frequency distributions

- frequency polygon
- histogram
- bar graph
- used to present the kind of information we've been looking at about distributions.

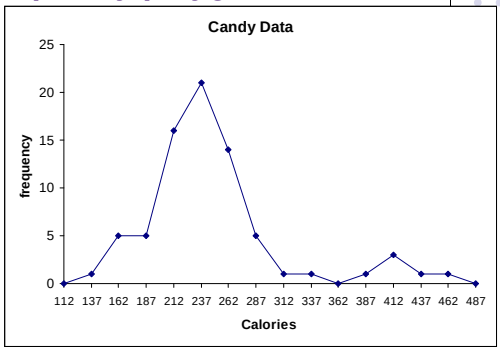
Selecting graph types

- dictated by the data we have collected
- frequency polygon or histogram is used for quantitative data
- bar graphs are generally used for qualitative data

Frequency Polygon

- appropriate for quantitative data
- each point represents frequency of class interval on X axis
- midpoints usually shown on X axis
- if scale starts above zero denoted with slash marks
- closed by connecting the curve to the X-axis at both ends.

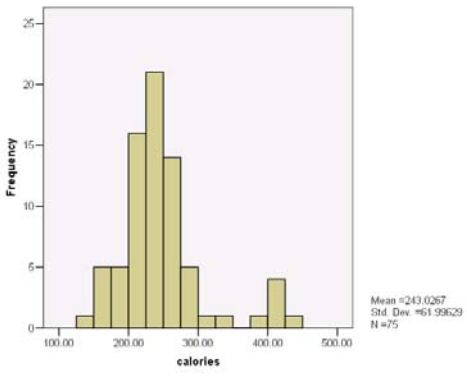
Frequency polygon



Histograms

- appropriate for quantitative data
- bars touching one another symbolize continuity of underlying scale
- better choice for discrete data

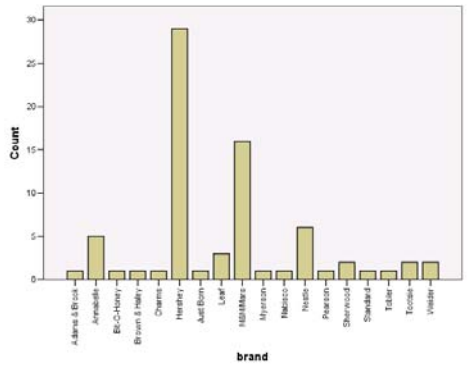
Histogram from SPSS



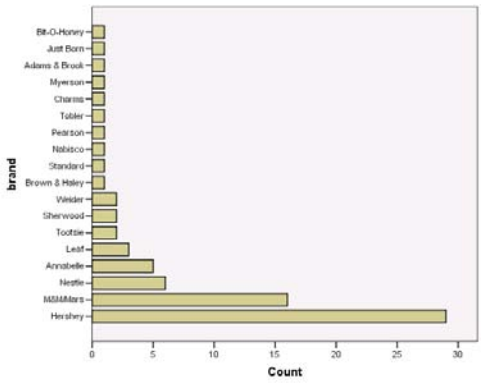
Bar graphs

- used for qualitative data
- large number of categories, horizontal orientation can be easier to read
- convention is to orient bars vertically
- if variable is ordinal level, ordering of attributes dictates order on graph
- if nominal, order doesn't matter, alphabetical often used

Bar graph from SPSS



Bar Graphs



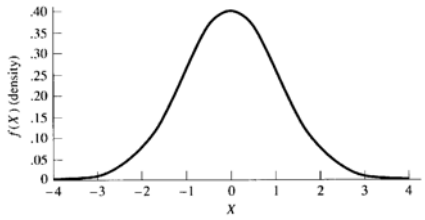
Line graphs

- Used to express one variable as a function of another
- Example in text, pg. 41, serial position effect

Describing distributions

- every one of the standard distributions we will look at has a mathematical equation associated with it
- we will generally not be interested in these.
- can use pictures and words to describe
- Bell shaped curve
  - most observations concentrated in middle
  - special case of bell-shaped distribution is called the normal distribution or curve
  - very well-known properties.

Normal distribution



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}(e)^{-(x-\mu)^2/2\sigma^2}$$

Skewed distributions

- distribution is not symmetrical
- tail trails off in one direction or the other
- greatest frequency of observations not in middle
- skewness is in the direction the tail trails off
- Other shapes of distributions
  - Rectangular (also called uniform distribution)
  - Bimodal (mode = most frequently occurring value)
  - J-curves (highly skewed)