Geographic Data and Mapping

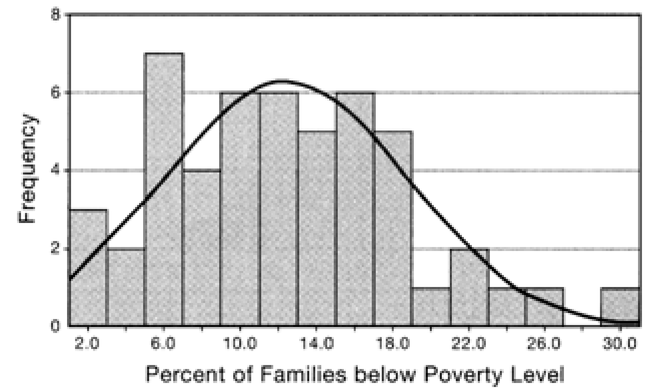
GEOG380 FA 2018

Statistical foundation

- Numerical approaches for map analysis
  - Methods for analyzing spatial data
  - Numeric summaries for analyzing data
    - Measures of central tendency, outliers, ranges
    - Stem-and-leaf plot & histogram
    - Variance & standard deviation
    - Rates, proportions, and percentages
  - Standardizing data

Numerical approaches for map analysis : Methods for analyzing spatial data

- Graphs
  - Histogram
    - Class: a group between multiple values
    - Height: amount of the values in each class
    - Normal distribution
      - Most of the observations locate near the mean (middle of the distribution)
      - Fewer observations locate in both tails
    - Positively / negatively skewed distributions (where are the tails?)
      - Examples: figure 3.2 (p.39, next slide)



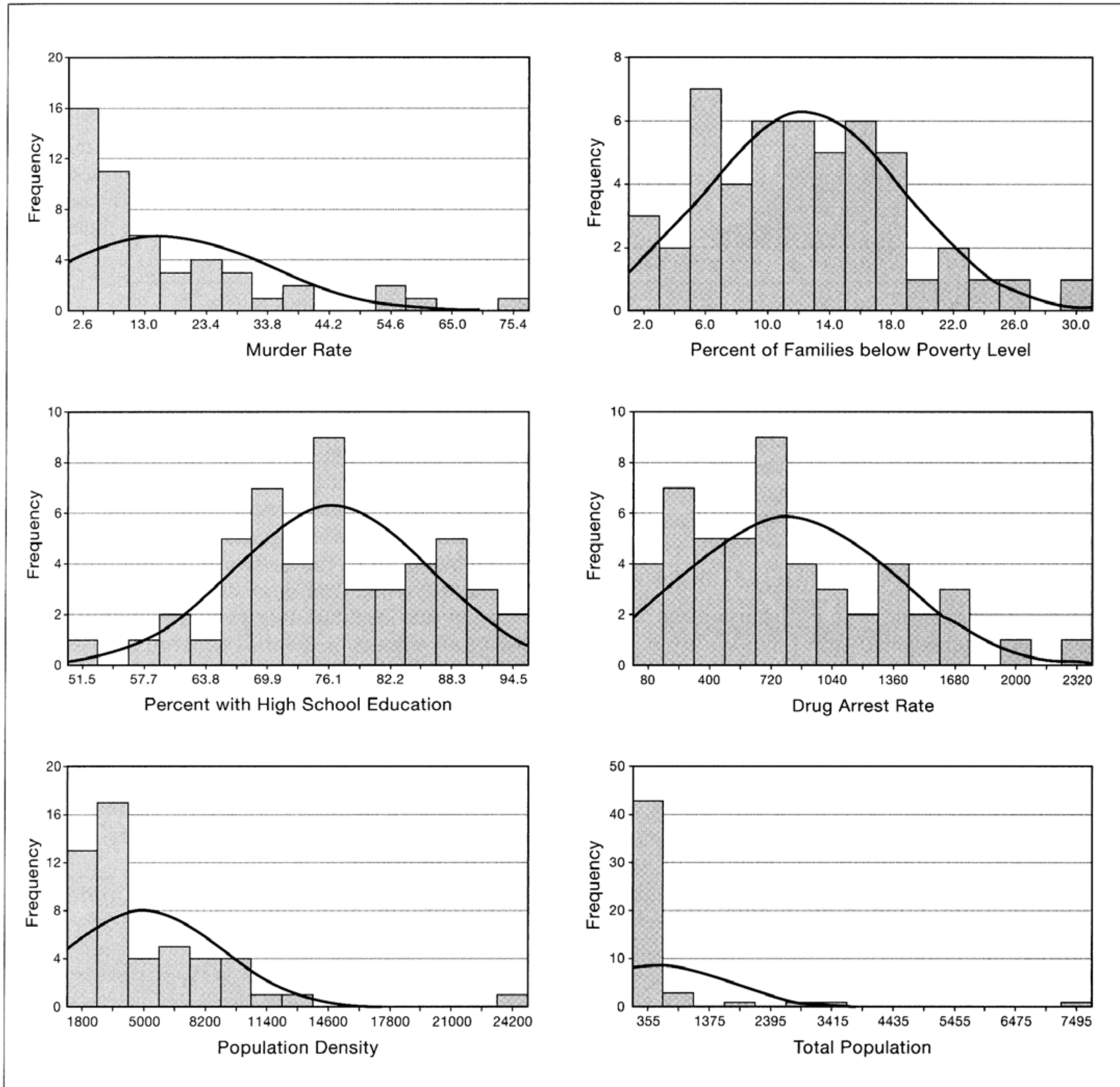


FIGURE 3.2 Histograms for the data presented in Table 3.1.

Methods for analyzing spatial data

- Tables
  - Raw table
    - Table 3.1 (p.36)

TABLE 3.1 Sample data for 50 U.S. cities (sorted on murder rate)

City	Murder Rate*	Families below Poverty Level (%)	High School Graduates (%)	Drug Arrest Rate†	Population Density‡	Total Population (in Thousands)
Irvine, CA	0.0	2.6	95.1	780	2607	110
Cedar Rapids, IA	0.9	6.6	84.5	110	2034	109
Overland Park, KS	0.9	1.9	94.1	255	2007	112
Livonia, MI	1.0	1.7	84.7	665	2823	101
Lincoln, NE	1.6	6.5	88.3	294	3033	192
Madison, WI	1.6	6.6	90.6	57	3311	191
Glendale, CA	1.7	12.3	77.2	452	5882	180
Allentown, PA	1.9	9.3	69.4	1078	5934	105
Tempe, AZ	2.1	7.0	89.9	295	3590	142
Boise City, ID	2.4	6.3	88.6	512	2726	126
Lakewood, CO	2.4	5.2	88.2	216	3100	126
Mesa, AZ	3.1	6.9	84.8	223	2653	288
Pasadena, TX	3.4	11.1	69.8	370	2727	119
San Jose, CA	4.5	6.5	77.2	1289	4568	782
Waterbury, CT	4.6	9.9	66.8	1326	3815	109
Springfield, MO	5.0	11.6	77.0	446	2068	140
Chula Vista, CA	5.2	8.6	75.7	808	4661	135

Numeric summaries

2.6, 6.6, 1.9, 1.7, 6.5, 6.6, 12.3, 9.3

- If you use some kind of numeric information as data, you should also provide an explanation for that numeric such as…
  - Central tendency measurement
    - Mode: the most frequently occurring value (=6.6)
    - Median: the middle value in an ordered set of data (1.7, 1.9, 2.6, 6.5, 6.6, 6.6, 9.3, 12.3) (=6.55)
    - Mean: the average of the data (=5.938)
    - Standard deviation (SD): average of distance between each value and the mean of the data (=3.749)
      - ±1 SD from mean, ±2 SD from mean (later in details)

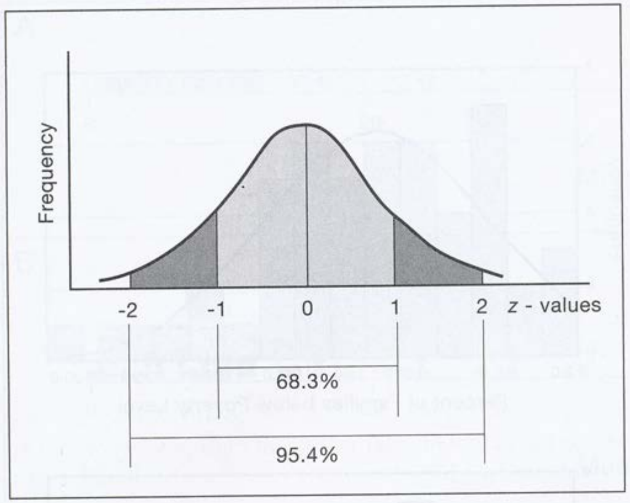
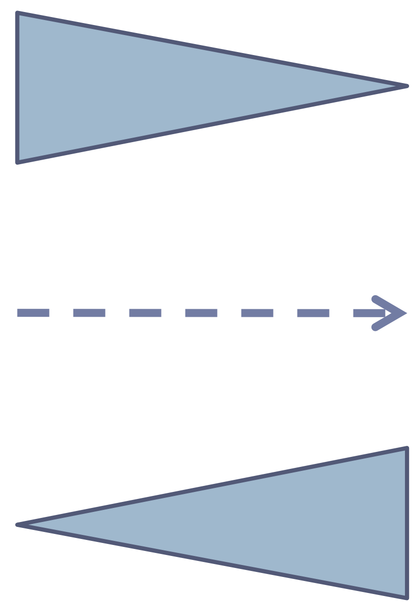


FIGURE 3.3 An example of a normal curve. Histograms will approximate this shape if the data are normal. For a perfectly normal data set, approximately 68 percent and 95 percent of the observations will fall within 1 and 2 standard deviations, respectively, of the mean.

Two types of statistics:

- Descriptive
  - Reducing lots of data to manageable and digestible pieces of information — this is what most mapping is all about

  - e.g., generalization, shorelines



- Inferential
  - Understanding what conclusions can be drawn from limited information. Often in the form of samples of a population
    – e.g. sampling, geostatistics

The problem: LOTS of data:

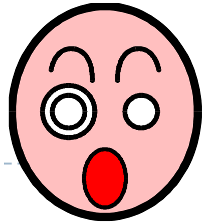
Number of deer-vehicle accidents between counties in Ohio

COUNTY_NAME	DEERVEH02	DEERVEH03	DEERVEH04
MONROE	32	36	26
CHAMPAIGN	79	56	98
LAWRENCE	202	135	101
HOCKING	146	127	112
MADISON	149	147	121
BELMONT	160	185	149
OTTAWA	167	151	154
VINTON	229	233	155
PREBLE	214	214	158
VAN WERT	118	194	158
PIKE	146	119	162
PAULDING	159	181	177
MERCER	179	177	182
PUTNAM	121	130	183
NOBLE	228	213	186
CLARK	265	218	191
HARRISON	220	219	191
FAYETTE	195	223	195
HENRY	170	186	201
MORGAN	186	191	204
GALLIA	361	261	206
CARROLL	245	279	214
PERRY	279	243	218
MEIGS	204	207	220
FULTON	181	190	234
JEFFERSON	270	253	238
SCIOTO	252	206	238
AUGLAIZE	238	266	259
HARDIN	199	275	262
DARKE	224	272	267
CRAWFORD	238	286	268
ERIE	219	318	268
WASHINGTON	312	336	275
WYANDOT	244	322	277
ATHENS	413	388	280
PICKAWAY	302	336	283
ADAMS	323	332	287
CLINTON	295	290	290
MARION	296	308	297
SANDUSKY	256	320	299
HURON	320	359	302
JACKSON	376	318	305
MONTGOMERY	415	395	325
UNION	332	364	333
MIAMI	327	390	342
BROWN	382	359	351
WOOD	294	337	357
LICKING	338	324	364
COLUMBIANA	381	447	372
GEAUGA	425	443	374
ALLEN	404	405	376
HIGHLAND	399	394	376
DEFIANCE	335	354	378
GREENE	493	445	391
SHELBY	339	373	393
SENECA	308	381	399
MEDINA	389	426	400
GUERNSEY	475	435	410
HOLMES	392	383	416
WAYNE	434	506	440
COSHOCTON	612	577	455
STARK	519	591	455
HANCOCK	373	500	460
PORTAGE	494	500	461
ASHTABULA	554	583	464
WILLIAMS	378	453	472
ASHLAND	432	488	473
MAHONING	443	516	476
WARREN	425	482	478
BUTLER	484	498	483
CUYAHOGA	476	525	485
FAIRFIELD	493	505	485
TRUMBULL	477	482	506
FRANKLIN	489	511	515
ROSS	518	555	518
TUSCARAWAS	502	591	529
CLERMONT	561	541	537
LORAIN	419	517	540
DELAWARE	547	577	560
KNOX	632	612	576
LOGAN	561	449	612
SUMMIT	674	642	618

Suggestions?

Ordering the numbers

26,98,101,112,121,149,154,155,158,158,162,177,
182,183,186,191,191,195,201,204,206,214,218,220,
234,238,238,259,262,267,268,268,275,277,280,
283,287,290,297,299,302,305,308,308,310,325,333,
342,351,357,364,372,374,376,376,378,391,393,
399,400,410,416,440,455,455,460,461,464,472,473,
476,478,483,485,485,506,515,518,529,537,540,
560,576,612,618,670,714,718



Measures of central tendency

- The *mean*
  - *Add all the values*
  - *Divide by the number of values*

$$\frac{\sum x}{n}$$

339

- The *median*
  - *Rank all the values*
    26,98,101,112,121,149,154,155,158,158,162,177,182,183,186,191,191,195, 201,204,206,214,218,220,234,238,238,259,262,267,268,268,275,277,280, 283,287,290,297,299,302,305,308,308,310,325,333,342,351,357,364,372, 374,376,376,378,391,393,399,400,410,416,440,455,455,460,461,464,472, 473,476,478,483,485,485,506,515,518,529,537,540,560,576,612,618,670, 714,718
  - *Find the middle value(s) (between the 44^{th} ~ the 45^{th} values from 88 observations)*

309

- The *mode (more useful for class/category data)*
  - The most common value/class

158, 191, 238, 268, 376, 455, 485

Outliers

- An *outlier* is *an extreme value* a long way from the mean or median

- To identify outliers and other types of characteristics that relate to the distribution of data, we need measures of *data spread, variation, or dispersion* such as…
  - Range, Inter-quartile range (IQR)
  - Variance, Standard Deviation
    (next slides)

Ranges

- **Range**
  - Maximum value – minimum value
    718 - 26 = 692
- ***Quantiles***
  - Divides ranked observations into equally-large sets
    - Percentiles…into 100 sets (88 observations in total)
    - Deciles…into 10 sets
    - Quartiles…into 4 sets
- ***Inter Quartile Range (IQR)***
  - A range that includes the middle-half of the ranked data or…

$$IQR = P_{75} - P_{25} = 460 - 214 = 246$$

26
98
101
112
121
149
154
155
158
158
162
177
182
183
186
191
191
195
201
204
206
214
218
220
234
238
238
259
262
267
268
268
275
277
280
283
287
290
297
299
302
305
308
308
310
325
333
342
351
357
364
372
374
376
376
378
391
393
399
400
410
416
440
455
455
460
461
464
472
473
476
478
483
485
485
506
515
518
529
537
540
560
576
612
618
670
714
718

Stem-and-leaf plot: a graphical summary

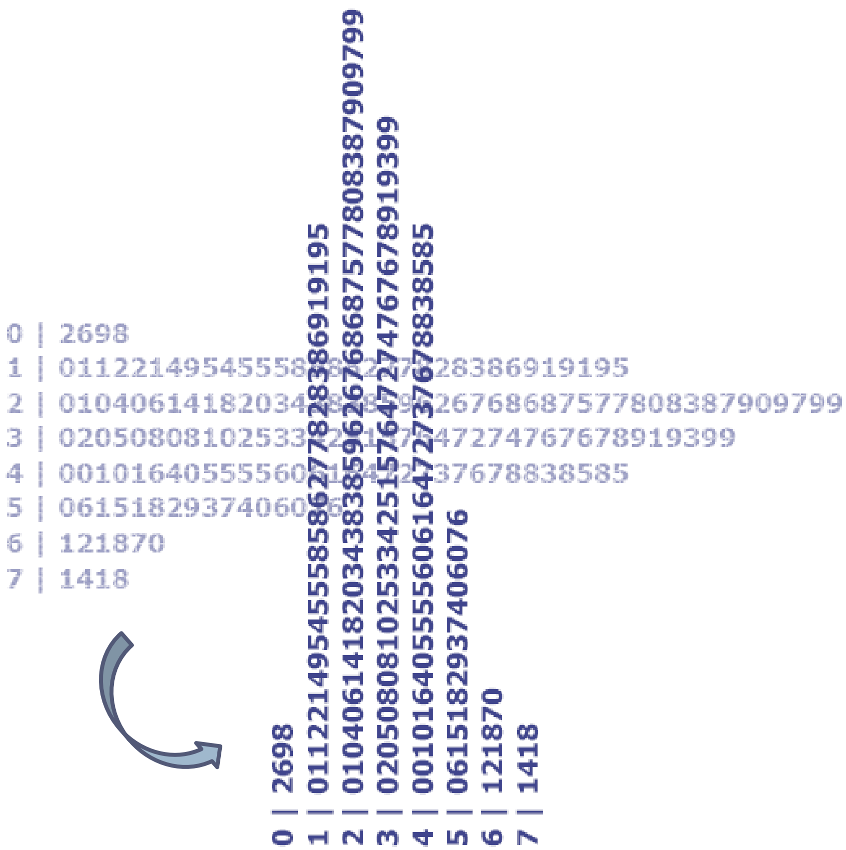
- This gives a good visual summary of the data, without too much graphical efforts
- Useful for just getting a rough idea of the whole data

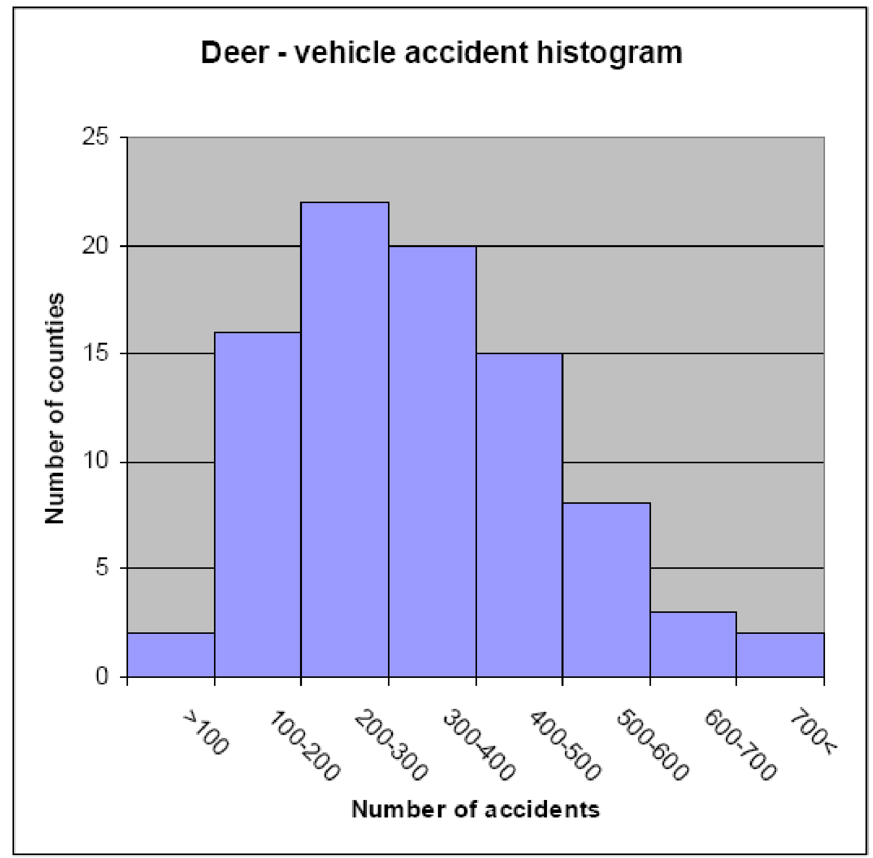
0 | 2698
1 | 011221495455585862778283869191 95
2 | 010406141820343838596267686875778083879097 99
3 | 020508081025334251576472747676789193 99
4 | 00101640555560616472737678838585
5 | 06151829374060 76
6 | 121870
7 | 1418

71, 74, 71, 78

From Stem-and-leaf plot to Histogram

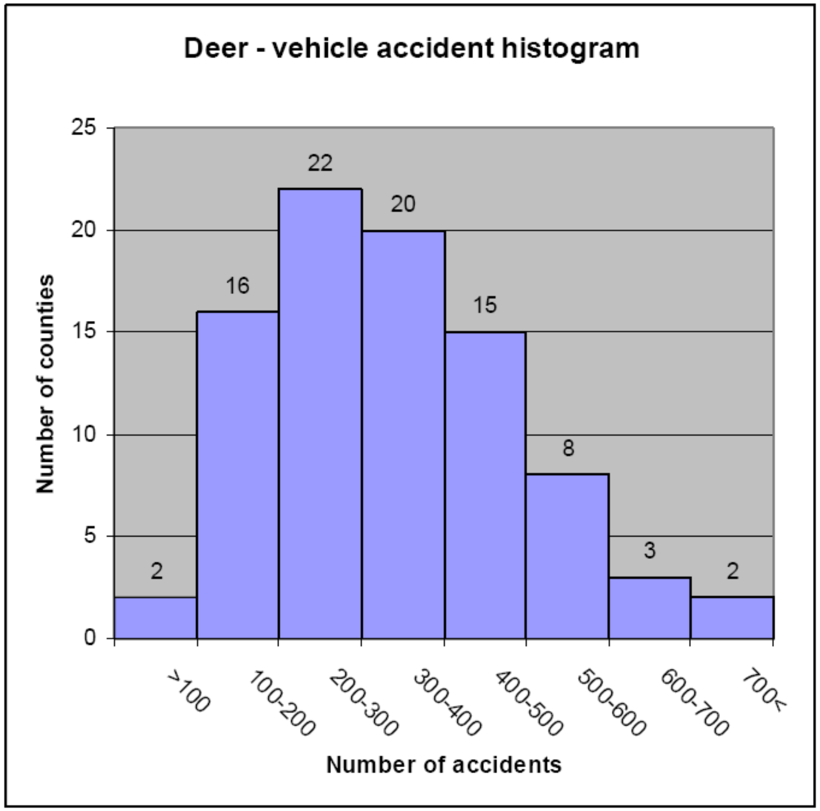
- Turn the stem-and-leaf plot 90 degrees counterclockwise





recap: Histograms

- A more refined form of stem-and-leaf plots
- Divide the range of the data into a series of equal intervals (ex. 0 to 100, 100 to 200, and so on)
- Count how many cases lie in each interval
- Plot the counts (or frequencies) as vertical bars



More about Histograms

- Some important points about the intervals:
  - Use simple bounds (i.e., 0.5-1.0, NOT 0.46-0.98)
  - Respect natural breakpoints (i.e., 0 °C, pH 7, 50%)
  - No overlaps (mutually-exclusive categories) between classes (i.e., NOT 0~10, 9~20)
  - Cover all values (i.e., NOT 0~10, 12~20)
  - Same interval-widths between classes (i.e., NOT 0~10, 11~15)
  - Appropriate number of classes (3 classes, 100 classes)

ALWAYS LABEL EVERYTHING in a histogram!

Variance and Standard Deviation

- Variance
  - Calculates an average of how much each value differs from the mean in squared

$$\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}$$

1) Sum all differences
2) Use a square to avoid negative values
3) Divide by the number of values (n-1 for a sample, n≥2) ← Q. why n≥2?

- Standard Deviation (SD or Std. Dev. , σ, sigma)
  - A measure of how dispersed numbers are
  - Square root of the variance

$$\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}}$$

- Ex) the deer-vehicle accident data
  - Variance = 22405, Standard deviation = about 150
  - Variance shows much bigger and positive numbers than std. dev.

Standard Deviation (SD or Std.Dev.)

- So what does the SD mean to us?
- Measure the SD and add/subtract from the mean, and you get a range of deviation from the mean
- Then often you can apply the statistical empirical rule
  - Given a set of *n measurements of a normally distributed* variable,
    - The mean ± 1SD includes roughly 68% of the observations
    - The mean ± 2SD includes roughly 95% of the observations
    - The mean ± 3SD includes roughly 99.7% of the observations

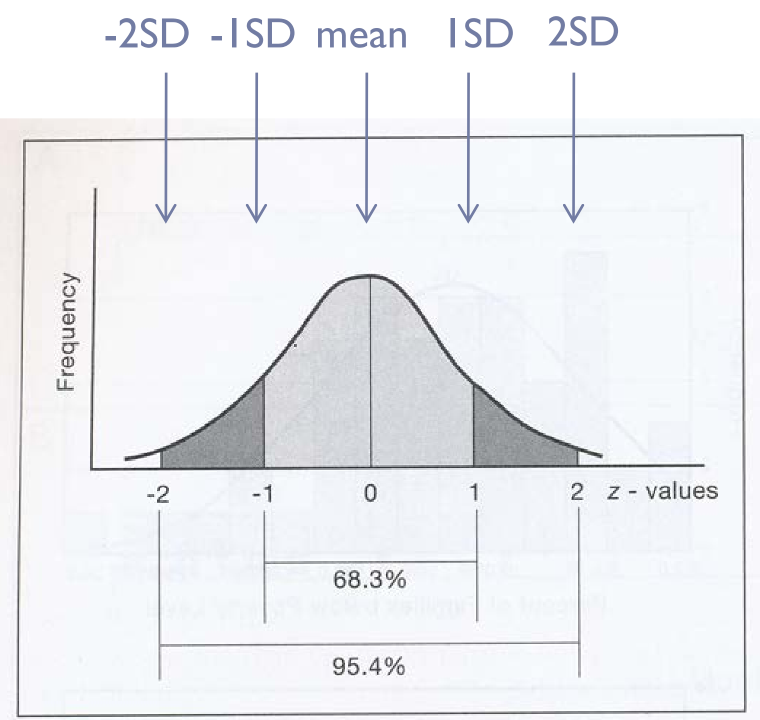
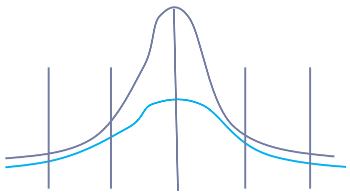


FIGURE 3.3 An example of a normal curve. Histograms will approximate this shape if the data are normal. For a perfectly normal data set, approximately 68 percent and 95 percent of the observations will fall within 1 and 2 standard deviations, respectively, of the mean.

Std.Dev. continued…

- For the deer-vehicle accidents data we have *mean*=158, *SD*=150, then…
  - 158 (mean) ± 1 x 150 (1SD) = 8 ~ 308 (should be ≈ 68% of all counties if the data shows normal distribution)
  - 158 (mean) ± 2 x 150 (2SD) = *0* ~ 458 (should be ≈ 95%)
  - 158 (mean) ± 3 x 150 (3SD) = *0* ~ 608 (should be ≈ 99.7%)
- Was this close in the real data?
  - The *range* was 692 (|26-718|), so…
    - 8 ~ 308 is ≈ 50% of the observations
    - 0 ~ 458 ≈ 74% of the observations
    - 0 ~ 608 ≈ 94% of the observations

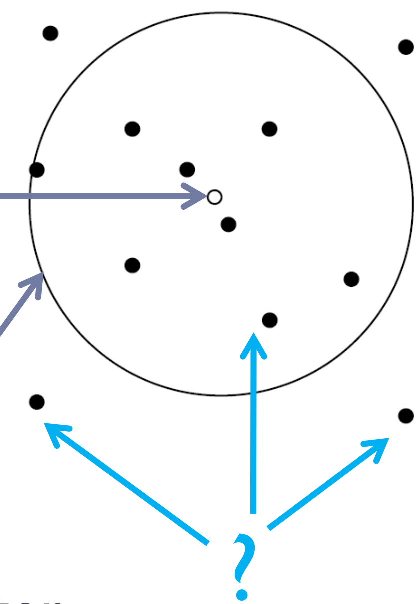


→ Data does not show normal distribution but dispersed

Mean Center and Dispersion Measures (p.51)

- Central tendency
  - Mean center

$$\bar{s} = \left(\mu_x, \mu_y\right) = \left(\frac{\sum_{i=1}^{n} x_i}{n}, \frac{\sum_{i=1}^{n} y_i}{n}\right)$$



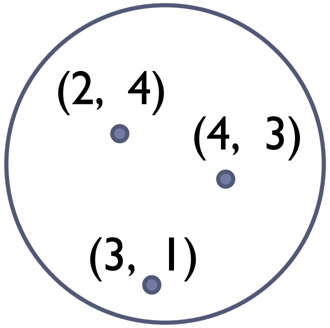
- Dispersion (measure of spread)
  - How much is the data dispersed?
    → use standard distance *d* from the mean center
    larger *d*: the data are more dispersed

$$d = \sqrt{\frac{\sum_{i=1}^{n}\left(\left(x_i - \mu_x\right)^2 + \left(y_i - \mu_y\right)^2\right)}{n}}$$

Similar to Std.Dev. = $\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{n-1}}$

Exercise

- What is the mean center of the given points in the figure below?
  -
- What is the dispersion measure?
  -



(map not to scale)

Derived indices

- Rates make data more comparable than raw values
  - E.g. Rate of vehicle accidents *per* population
- You often hear rates reported as an index
  - Often an index expresses each value as a percentage of some base value, or as standardized z-scores (in details later)
  - *Aspatial* examples
    - Dow Jones, Consumer Price, Poverty, Sustainability, GNP, …
  - Spatial examples
    - Location quotient (local economy VS. reference economy), Heat, Wetness, UV, Normalized Difference Vegetation Index (NDVI), The Average Watershed Nitrogen Leaching Index (AWNLI), …

Rates, Proportions, and Percentages

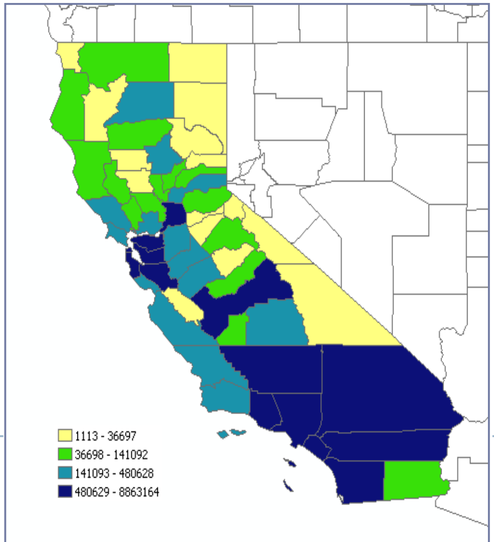
- Rates are a way of standardization of data to a common measure for comparison purposes

State	Population	Robbery	Total offenses	Robbery / 1000 p.	Robbery / all offenses
Colorado	4,417,714	3,555	186,379	0.80	1.9 %
Delaware	796,165	1,156	32,267	1.45	3.6 %

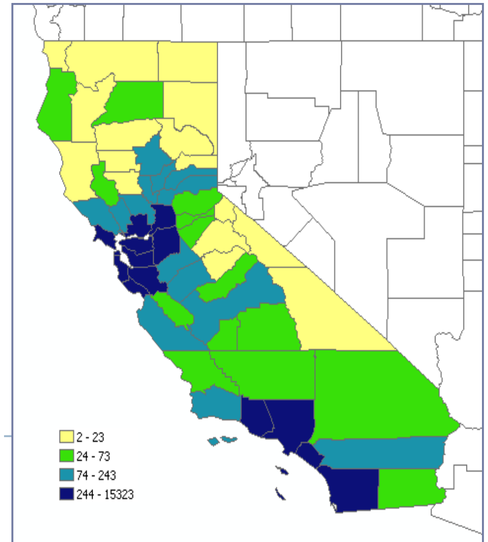
Source: U.S. Department of Justice

- Geographically… (LA county Vs. San Bernadino county)

1990 Population, CA (raw data)



1990 Population Density, CA (rates)



Rate calculations – general notes

- Choose some basis of the unit value
  - Usually population, area, total income, or number of households… of an areal unit
  - However, some variables are relevant for total count
    - For example, …?

- Decide how to express the unit of the results
  - Per person, per 100,000 people, per unit area, etc.
  - The calculation is:

    Rate = Count / Basis & Unit ex. 10people/mi^2

ALWAYS LABEL THE UNIT!

September 23rd, 2009

Map of the day, McDonald's edition

Posted by: Felix Salmon
Tags: consumption, charts

Post a comment
(3)

Are you happy with the map…?

This beautiful map comes from Stephen Von Worley, who has mapped the

September 24th, 2009
5:05 am GMT
[permalink]

If you normalised this by dividing by population density I'd imagine it's pretty smooth. Cool that you can see the highways in the West (route 80, 84 and 15).

– Posted by Nic Fulton

http://blogs.reuters.com/felix-salmon/2009/09/23/map-of-the-day-mcdonalds-edition/

For next time

- Readings
  - Ch. 3
- Worksheet 1