22S:30/105 Statistical Methods and Computing

Measures of Center, continued Measures of Dispersion

Lecture 3
January 23, 2017

Kate Cowles
374 SH, 335-0727
kate-cowles@uiowa.edu

The mean is meaningful only for quantitative data (either discrete or continuous).

- Example regarding a discrete variable: We hear reports such as that the average number of children per family is 1.9.
- The mean is not meaningful for nominal or ordinal data.

Exception: if a binary variable is coded as 0 and 1.

Then the arithmetic mean is the proportion of observations in the dataset that have value 1.

The median

The median is the 50th percentile of a set of observations.

- Values must be sorted from smallest to largest.
- If the number of observations is odd, then the median is the middle value.

  75 80 82 88 95

  The median is 82.
- If the number of observations is even, then the usual way to define the median is as the **mean** of the **two** middle values.

  75 80 82 88 95 97

  The median is $\frac{82+88}{2} = 85$.

The median is **not** strongly affected by a few extreme values in the dataset.

Example 1:

75 80 82 88 95

- mean = 84
- median = 82

Example 2:

25 80 82 88 95

- mean = 74
- median = 82

The median is *robust* to extreme values.

The median can be used as a measure of center for **ordinal** data as well as for discrete and continuous data.

Example: The NYC poll

city1yr	Frequency	Percent	Cumulative Frequency
Worse	593	61.64	593
Same	252	26.20	845
Better	111	11.54	956

- 956 people answered this question regarding whether they thought the condition of the city in June, 2003, was better, worse, or the same as one year earlier.
- If the values are sorted from smallest to largest (Worse, Same, Better), then the median will be the average of the 478th and 479th values.
- We can use the cumulative frequencies in the table to figure out what these have to be. They are both in category "Worse."
- Thus the median is Worse.

The mode

- The mode of a set of values is the value that occurs most frequently.
- Example: in the NYC poll data, the mode of the "city1yr" variable is Worse.
- Example: There is no mode in the birthweights data, because no value occurs more than once.
- There may be more than one mode in a set of values.
- The mode may be reported for **all** types of data.

When is each measure of central tendency appropriate?

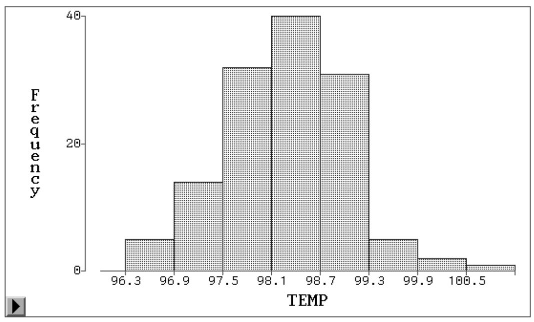
Depending on data type

- Nominal data
  - mode only
  - possible exception: binary data coded 0 and 1
- Ordinal data
  - mode or median
- Quantitative data
  - mean, median, or mode

Depending on the shape of the distribution of values (quantitative variables)

- if the shape is approximately symmetric and has only one mode
  - mean and median will be close in value
  - mean is typically reported

  Example: the body temperature data



From a statistical computer package:

- mean = 98.24
- median = 98.3

- if the distribution is highly skewed
  - if skewed to the right, mean will be larger than median
  - if skewed to the left, mean will be smaller than median
  - mean may not be a “typical” value

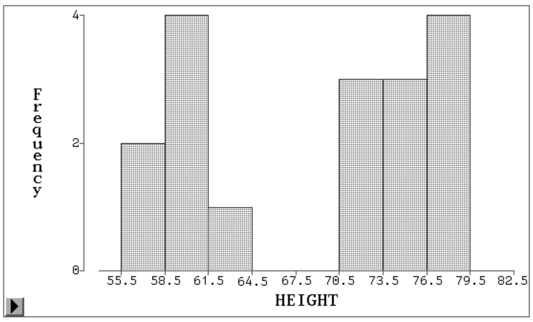
Example: the billionaire data

From a statistical computer package:

- mean = 2.7 billion
- median = 1.8 billion

- if the distribution has more than one mode
  - neither the mean nor the median may be representative values
  - may be best to report all modes and/or to display a graph
  - may occur if two or more different subgroups are represented in the sample

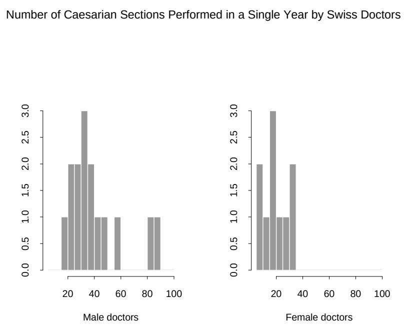
Example:



From a statistical computer package:

- mean = 69.0
- median = 72.0

In getting the “overall picture” of quantitative data, the spread is just as important as the center of the data.



Numerical measures of dispersion

- the range
- the interquartile range
- the standard deviation

The range

- The range is the difference between the largest and the smallest observations.
- For the male Swiss doctors,
  - largest value = 86
  - smallest value = 20
  - range = $86 - 20 = 66$
- For the female Swiss doctors,
  - largest value = 33
  - smallest value = 5
  - range = $33 - 5 = 28$

The range shows the full spread of the data, but may be exaggerated if the largest and/or smallest values are atypical (outliers)

- Example: the 1992 billionaire data
  - With Bill Gates:
    range = 37 - 1 = 36 billion
  - If Bill were deleted:
    range = 24 - 1 = 23 billion
- Example: the male Swiss doctors data
  - With the largest two values
    range = 86 - 20 = 66 billion
  - If the two largest values were deleted:
    range = 59 - 20 = 39 billion
- So additional measures are needed to give a more complete picture of the spread of values.

The quartiles and the interquartile range

- The *first quartile* is the same as the 25th percentile
  - one quarter of the observations in a dataset have values less than or equal to the 1st quartile, and the other three quarters have values greater than or equal to the first quartile
- The *third quartile* is the same as the 75th percentile
  - three quarters of the observations in a dataset have values less than or equal to the 3rd quartile, and the other one quarter have values greater than or equal to the 3rd quartile

- The interquartile range (IQR) is the difference between the 3rd and 1st quartiles
- For the male Swiss doctors,
  - third quartile = 50
  - first quartile = 27
  - IQR = 50 - 27 = 23
- For the female Swiss doctors,
  - third quartile = 29
  - first quartile = 14
  - IQR = $29 - 14 = 15$
- For the 1992 billionaires,
  - third quartile = 3 billion
  - first quartile = 1.3 billion
  - IQR = $3 - 1.3 = 1.7$ billion

The IQR is considered less sensitive to outliers than the range.

- Example: the 1992 billionaire data
  - With Bill Gates:
    IQR = $3 - 1.3 = 1.7$ billion
  - If Bill were deleted:
    IQR = $2.9 - 1.3 = 1.6$ billion
- However, in a small dataset, deletion of a few outliers may affect the IQR substantially.
- Example: the male Swiss doctors
  - IQR with the two largest values included:
  - IQR = $50 - 27 = 23$
  - IQR with the two largest values deleted:
  - IQR = $37 - 27 = 10$

The five-number summary

- The five-number summary provides a reasonably-complete numeric summary of the center and dispersion of a set of values.
- The five-number summary consists of
  - the minimum value
  - the first quartile
  - the median
  - the third quartile
  - the maximum value

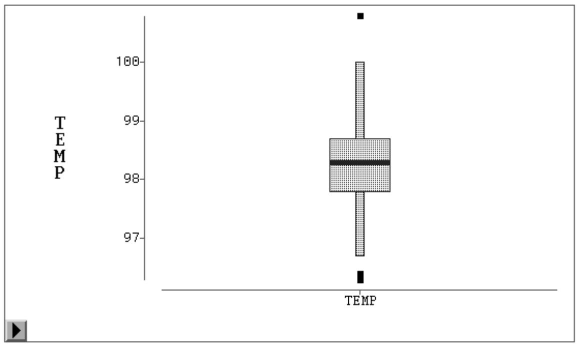
The five-number summary for the billionaire data may be extracted from the following computer output:

```
                 Quantiles(Def=5)

     100% Max        37        99%         14
      75% Q3          3        95%        6.2
      50% Med       1.8        90%        4.5
      25% Q1        1.3        10%        1.1
       0% Min         1         5%          1
                                1%          1
     Range           36
     Q3-Q1          1.7
     Mode             1
```

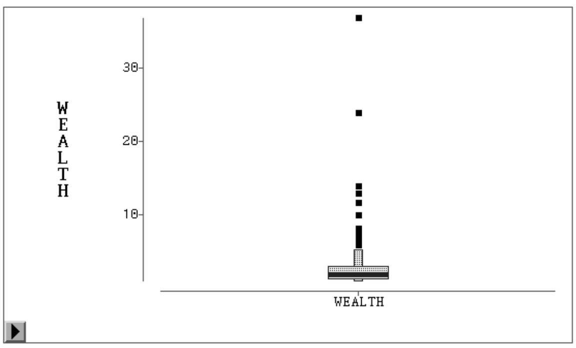

Boxplots

- are used to summarize the distribution of a continuous variable



- box extends from 1st quartile to 3rd quartile of data
- line in middle of box marks 50th percentile

- "whiskers" sticking out of box extend to *adjacent values*
  - adjacent values are most extreme observations that are not farther away from the edge of the box than 1.5 times the height of the box
- points farther out than the adjacent values are considered *outliers*
  - represented by circles or squares
  - probably are not typical of the rest of the data



The standard deviation

- The standard deviation measures spread by looking at how far the observations are from their mean.
- Example: quiz scores

  75 80 82 88 95

  The mean is

$$\bar{x} = \frac{75 + 80 + 82 + 88 + 95}{5} = 84$$

  points.

- We want a measure of typical distance between an individual value and this mean.

An idea that won't work for measuring the spread: take the average of the "deviations" of the individual observations from the mean.

Observed Value	Deviation from mean	Squared deviation
75	75 - 84 = -9	$(-9)^2 = 81$
80	80 - 84 = -4	$(-4)^2 = 16$
82	82 - 84 = -2	$(-2)^2 = 4$
88	88 - 84 = 4	$4^2 = 16$
95	95 - 84 = 11	$11^2 = 121$
sum	0	238

Because the sum of the deviations is always 0, the average deviation is always 0!

Solution: Square the individual deviations before adding them up!

The variance and the standard deviation

- The variance s^2 is the sum of the squared deviations divided by one less than the number of observations.

$$\begin{aligned} s^2 &= \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\ &= \frac{238}{4} = 59.5 \end{aligned}$$

  – We can think of the variance roughly as the average of the squared deviations.

- The standard deviation is the square root of the variance.
  $s = \sqrt{59.5} = 7.71$ points.

Facts about the standard deviation s

- s measures the spread of values around the *mean*
  – thus s should be used as a measure of dispersion only when mean has been chosen as the measure of center
- s is always greater than 0 unless all the observations have the same value
- s has same units of measurement as original observations
- s is sensitive to extreme observations
  – like the mean
- s is the most commonly-used measure of dispersion (is often used when it is not the best choice!)

The mean and standard deviation together provide a reasonable numeric summary of a set of values if the distribution is approximately **symmetric**.

- Example: the body temperature data

Variable	N	Mean	Std Dev
TEMP	130	98.2492308	0.7331832

- Example of inappropriate use of $\bar{x}$ and s to summarize a distribution: the billionaire data

Analysis Variable : WLTH

N	Mean	Std Dev
233	2.6815451	3.3188403