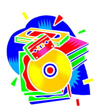
Central tendency

"I say what I means and I means what I say!."

Popeye

- Normal distribution video clip
- To view an unedited version visit:

http://www.learner.org/resources/series65.html#



mean for metric data

- 2 important properties
  - 1) sum of deviations from the mean = 0
  - 2) sum of + deviations = sum of - deviations
- 2 advantages
  - 1) more stable than other measures
  - 2) other important statistics can be derived using it
- Technically it is called the arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The mean

- variance and standard deviation
- problems
  - a) fractional values
  - b) cannot be computed if data is open ended
  - c) strongly affected by extreme cases

Worktable for Calculating Arithmetic Mean of Washington, D.C., Precipitation Data

Observation i	Precipitation X_i
1	41.11
2	54.29
3	35.09
...	...
38	34.98
39	35.96
40	50.50
Total	1598.00

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{41.11 + 54.29 + \ldots + 50.50}{40} =$$

$$\frac{1598.00}{40} = 39.95$$

Grouped data

- mean for grouped data

$$x = \frac{\sum x f_t}{N}$$

Worktable for Calculating Grouped **Mean of Washington, D.C., Precipitation Data**

Class interval j	Class midpoint X_j	Class frequency f_j	$X_j f_j$
25–29.99	27.5	4	110.0
30–34.99	32.5	5	162.5
35–39.99	37.5	12	450.0
40–44.99	42.5	9	382.5
45–49.99	47.5	5	237.5
50–54.99	52.5	4	210.0
55–59.99	57.5	1	57.5
Total		40	1610.0

$$\overline{X}_w = \frac{\Sigma X_j f_j}{n} = \frac{1610.0}{40} = 40.25$$

The weighted mean

If the weights are all equal then it's the same as the arithmetic mean

$$\overline{x} = \frac{\sum x_i w_i}{\sum w_i}$$

w_i= weight associated with ith case weights compensate for the higher chances of selecting some cases than others

Why use it?

- Each individual data value might actually represent a value that is used by multiple people in your sample. The weight, then, is the number of people associated with that particular value.
- Your sample might deliberately over represent or under represent certain segments of the population. To restore balance, you would place less weight on the over represented segments of the population and greater weight on the under represented segments of the population.

- Some values in your data sample might be known to be more variable (less precise) than other values. You would place greater weight on those data values known to have greater precision.

dichotomous data

- mean for dichotomous data

$$\overline{x} = p$$

- where p is the proportion of successes or cases coded 1

Sensitivity of the Mean to a Single Outlier

Values	Statistics
$21,000	Total = $500,000
21,000	
22,000	Mode = $21,000
26,000	
27,500	Median = $26,000
32,500	
349,000	Mean = $500,000/7 = $71,428.57

Sum of squares

- these measures of central tendency tell us nothing about the variability in the data or the dispersion
- one way to do this is compare the values with the mean value
- the simplest way is to subtract the mean from each value to see if it is higher or lower
- if you do this you get both + and - values
- if we summed them to get a sort of index we would get 0 as a total, to get around this we square the differences $|x_i - \bar{x}|$ this known as the **sum of squares**

Sum of squares

- or the total squared variation about the mean
- from this we can derive the variance and the standard deviation
- variance is the sum of the squared deviations from the mean divided by N for the population and n-1 for a sample
- remember that sample statistics are estimates of the population statistics
- the sample uses n-1 because it has been shown that the use of N for a sample results in an underestimation of the population variance

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1}$$

Standard deviation

- a short cut formula for the sample variance is

$$s^2 = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)}$$

standard deviation $s = \sqrt{s^2}$

a large standard deviation means a large variability in the data

Alternatively it can written as

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

The first term in the numerator is called the "*raw sum of squares*" or 'uncorrected sum of squares'
and the second term is called the "*correction term for the mean*"
A name for the numerator is the "*corrected sum of squares*",
and this is usually abbreviated by *Total SS*.

- It is called corrected because theoretically there is no error in the sum of squares
- This comes up again in analysis of variance later in the course
- In geog 201 it was denoted SS_x

Grouped data variance

- variance can also be calculated for grouped data
- $s^2 = \sum \frac{(x-M)^2 f}{N} = \sum \frac{x^2 f}{N} - M^2$
- where f_i=frequency of classes
- M=grouped mean

- A B FOR a $\Sigma X_i=3+7+9+2+4+6=31$
- 3 30 $\Sigma X^2 = 3^2+7^2+9^2+2^2+4^2+6^2=195$
- 7 70 $(\Sigma X_i)^2= 31^2=961$
- 9 90 $\bar{x} = 31/6 = 5.16$
- 2 20
- 4 40
- 6 60 $s^2=6(195)-961/6(6-1)=6.96$
- s=2.639
- for b
- x = 51.6
- s^2=696.6
- s=26.39

- problem with variance and standard deviation is that for the purpose of comparison, they are sensitive to the magnitude of the data

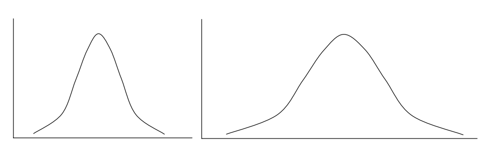
  for example in the previous data the variance and standard deviation of b was 10 times that of a
- to compare a and b we need to standardize
  - coefficient of variation $cv = \frac{s}{\bar{x}}$
- for a and b the coefficient of variation is 2.639/5.155 = .51 or 26.39/51.6=.51

Measure of Spread

- Characteristics of s and s^2
  - Always positive (why)
  - Related to mean; so can only use with mean
  - Like mean, large outliers exaggerate standard deviation

Normal curve

- Special curvature of normal curve
  - Can be fully described by mean and standard deviation
    - Mean tells where curve centered on number line
    - Standard deviation tells how steep



Normal curve

- Special curvature of normal curve
  - Can be fully described by mean and standard deviation
  - Always follows 68-95-99.7 rule
    - 68% of all observations within 1 sd of mean
    - 95% of all observations within 2 sd's of mean
    - 99.7% of observations within 3 sd's of mean

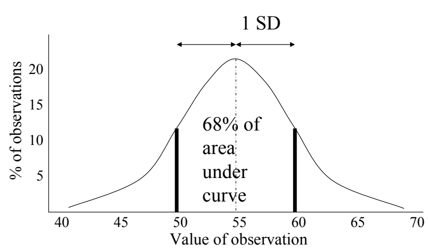
Normal curve formula

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

- Note you only need to know the mean and the variance to create the curve

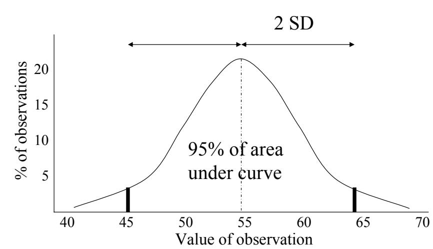
Normal curve

- 68-95-99.7 rule



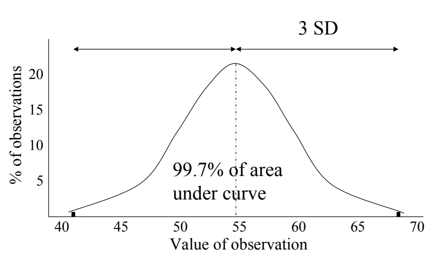
Normal curve

- 68-95-99.7 rule



Normal curve

- 68-95-99.7 rule



Normal distribution

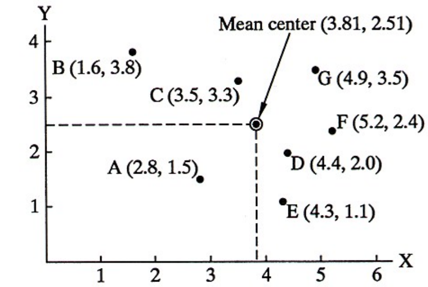
- For normal distribution the mean is the most efficient and therefore the least subject to sample fluctuations of all measures of central tendency.
- The sum of squared deviations of scores from their mean is lower than their squared deviations from any other number.

distribution statistics for spatial distributions

- the bivariate mean
- in geography the centre of an area may be of interest, can calculate the weighted bi-variate mean centre or the weighted centroid

$$w_i \bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

$$w_i \bar{y} = \frac{\sum w_i y_i}{\sum w_i}$$



Graph of Locational Coordinates and Mean Center

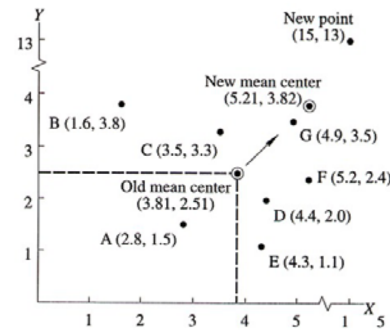
Worktable for Calculating Mean Center

Point	Locational coordinates* X_i	Y_i
A	2.8	1.5
B	1.6	3.8
C	3.5	3.3
D	4.4	2.0
E	4.3	1.1
F	5.2	2.4
G	4.9	3.5

$n = 7$ $\quad \Sigma X_i = 26.7$ $\quad \Sigma Y_i = 17.6$

$$\bar{X}_c = \frac{\Sigma X_i}{n} = \frac{26.7}{7} = 3.81 \qquad \bar{Y}_c = \frac{\Sigma Y_i}{n} = \frac{17.6}{7} = 2.51$$

Mean center coordinates: (3.81, 2.51)



Affect of an Outlier on Mean Center Location

Geographic Center of U.S. Population, 1790–1990 (Source: Bureau of the Census, U.S. Dept. of Commerce.)

The spatial mean?

Euclidean median

- Central location that minimizes the *unsquared* distances rather the squared ones
- It is methodically complex and has to be solved iteratively

$$(X_e, Y_e) = \min \sum \sqrt{(X_i - X_e)^2 + (Y_i - Y_e)^2}$$

Weighted Euclidean median

$$(X_{we}, Y_{we}) = \min \sum f_i \sqrt{(X_i - X_{we})^2 + (Y_i - Y_{we})^2}$$

Weighted Euclidean median

- Has important applications in geography
  - Weber location problem
  - Used in public and private facility algorithms
    - Urban fire station
    - Store site for clothing store
  - Can be extended to multiple locations to solved at one time
    - Neighbourhood health centers

standard distance

dispersion has it counterpart in bivariate descriptive statistics

- because distances are deviations in the geographic sense, it is defined as the equivalent of a standard deviation

$$SD = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1} + \sum \frac{(y_i - \bar{y})^2}{n-1}}$$

Worktable for Calculating Standard Distance

Point	Locational coordinates X_i	Y_i	X_i^2	Y_i^2
A	2.8	1.5	7.84	2.25
B	1.6	3.8	2.56	14.44
C	3.5	3.3	12.25	10.89
D	4.4	2.0	19.36	4.00
E	4.3	1.1	18.49	1.21
F	5.2	2.4	27.04	5.76
G	4.9	3.5	24.01	12.25

From earlier calculation of mean center:

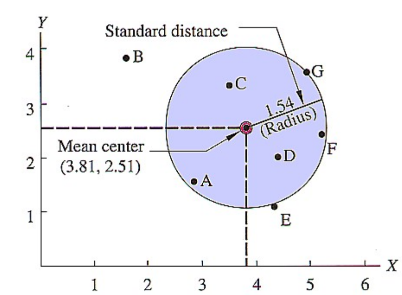
$\bar{X}_c = 3.81$ $\bar{Y}_c = 2.51$ $\bar{X}_c^2 = 14.52$ $\bar{Y}_c^2 = 6.30$

$n = 7$ $\Sigma X_i^2 = 111.50$ $\Sigma Y_i^2 = 50.80$

$$S_D = \sqrt{\left(\frac{\Sigma X_i^2}{n} - \bar{X}_c^2\right) + \left(\frac{\Sigma Y_i^2}{n} - \bar{Y}_c^2\right)}$$

$$= \sqrt{\left(\frac{111.50}{7} - 14.52\right) + \left(\frac{50.80}{7} - 6.30\right)}$$

$$= 1.54$$



Graph of Point Locations, Mean Center, and Standard Distance

Standard ellipse

- if you want to take possibility of an ellipse rather than a circle then we can calculate standard distance separately for X and Y

$$SD_x = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}$$

$$SD_y = \sqrt{\sum \frac{(y_i - \bar{y})^2}{n-1}}$$

for weighted observations

$$SD_w = \sqrt{\frac{\sum w_i (x_i - \bar{x})^2}{\sum w_i} + \frac{\sum w_i (y_i - \bar{y})^2}{\sum w_i}}$$

this is far too tedious to do by hand
so we would have to use a computer program

Worktable for Calculating Weighted Standard Distance

Point	f_i	X_i	X_i^2	$f_i(X_i)^2$	Y_i	Y_i^2	$f_i(Y_i)^2$
A	5	2.8	7.84	39.20	1.5	2.25	11.25
B	20	1.6	2.56	51.20	3.8	14.44	288.80
C	8	3.5	12.25	98.00	3.3	10.89	87.12
D	4	4.4	19.36	77.44	2.0	4.00	16.00
E	6	4.3	18.49	110.94	1.1	1.21	7.26
F	5	5.2	27.04	135.20	2.4	5.76	28.80
G	3	4.9	24.01	72.03	3.5	12.25	36.75

From earlier calculation of weighted mean center :

$\bar{X}_{wc} = 3.10 \quad \bar{Y}_{wc} = 2.88 \quad \bar{X}_{wc}^2 = 9.61 \quad \bar{Y}_{wc}^2 = 8.29$

$\Sigma f_i = 51 \quad \Sigma f_i(X_i)^2 = 584.01 \quad \Sigma f_i(Y_i)^2 = 475.98$

$$S_{WD} = \sqrt{\left(\frac{\Sigma f_i(X_i)^2}{\Sigma f_i} - \bar{X}_{wc}^2\right) + \left(\frac{\Sigma f_i(Y_i)^2}{\Sigma f_i} - \bar{Y}_{wc}^2\right)}$$

$$= \sqrt{\left(\frac{584.01}{51} - 9.61\right) + \left(\frac{475.98}{51} - 8.29\right)}$$

$$= 1.70$$

- A good free program for doing these simple spatial statistics is Crimestat
- http://www.icpsr.umich.edu/NACJD/crimestat.html