PSY 201: Statistics in Psychology

Lecture 05

Central tendency

Does a company deserve a tax break?

Greg Francis

Purdue University

Fall 2019

DISTRIBUTION USES

- summarize data
- indicate most frequent data values
- indicate amount of variation across data values
- allows us to interpret a single score in the context of other scores
- we are exploring quantitative methods to describe distributions

LIMITATIONS

- Last time we discussed percentiles and percentile ranks
- very useful for comparing a score to a distribution of scores
- not so good for talking about a distribution overall
- want to quantify ideas of central tendency (most of scores, average score,...)
    - mode
    - median
    - mean

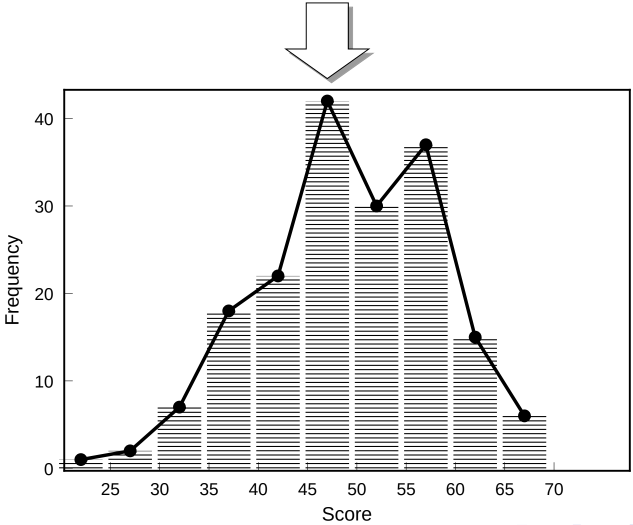
MODE

- the most frequent data value (score)
- easy to find from a table of frequency scores

Exact Limits	Midpoint	f	cf	%	c%
64.5–69.5	67	6	180	3.33	100
59.5–64.5	62	15	174	8.33	96.67
54.5–59.5	57	37	159	20.56	88.34
49.5–54.5	52	30	122	16.67	67.78
44.5–49.5	**47**	**42**	**92**	**23.33**	**51.11**
39.5–44.5	42	22	50	12.22	27.78
34.5–39.5	37	18	28	10.00	15.56
29.5–34.5	32	7	10	3.89	5.56
24.5–29.5	27	2	3	1.11	1.67
19.5–24.5	22	1	1	0.56	0.56

MODE

- top of a hill on a frequency distribution graph
- easy to find

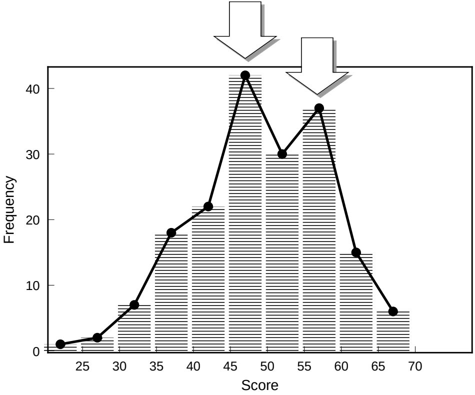


MODE

- With a frequency distribution we actually look for a **modal interval** and consider the midpoint of the interval to be the mode
- **unimodal distribution:** when there is a single mode. (single hill)
- **multimodal distribution:** when there are several modes. (many hills)
- **bimodal distribution:** when there are two modes. (two hills)
- NOTE: the use of the terms are not quite consistent!

BIMODAL

- this distribution might be called bimodal, even though there is really only one mode!

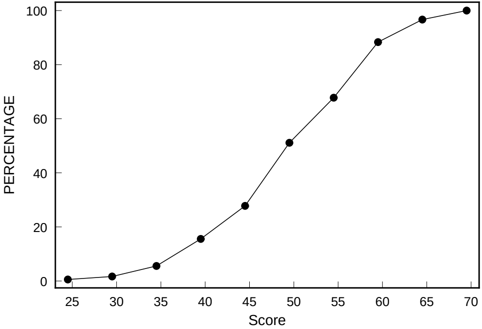


- not very useful for mathematics!

MEDIAN

- the point below which 50% of scores fall
- the 50th percentile

$$\text{Mdn} = P_{50} = ll + \left(\frac{n(0.5) - cf}{f_i} \right) (w)$$



CALCULATIONS

- for our data set
  - $ll = 44.5$; exact lower limit of the interval containing the $n(0.5)$ score
  - $n = 180$; total number of scores
  - $0.5 = 50/100$, proportion corresponding to 50th percentile (decimal form)
  - $cf = 50$; cumulative frequency of scores **below** the interval containing the $n(0.5)$ score
  - $f_i = 42$; frequency of scores **in** the interval containing the percentile point
  - w =5; width of class interval

$$\text{Mdn} = P_{50} = 44.5 + \left(\frac{180(0.5) - 50}{42} \right) (5) = 49.26$$

CALCULATIONS

- when the raw scores are used (instead of class intervals)
  1. Arrange the scores in ascending order (from lowest to highest).
  2. If there is an odd number of scores, the median is the middle score.
  3. If there is an even number of scores the median is halfway between the two middle scores.
- Will this always give the same value as for the frequency distribution approach?

CALCULATIONS

Name	Sex	Score
Aimeé	Female	94
Greg	Male	95
Ian	Male	89
Jim	Male	92

- scores: 89, 92, 94, 95 (even number of scores)
- the median is: halfway between 92 and 94 $=$ 93

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92
Bob	Male	83

- scores: 83, 89, 92, 94, 95 (odd number of scores)
- the median is: the middle score $=$ 92

MEAN

- arithmetic average of scores in a distribution
- mean of a **population** is designated as μ
- mean of a **sample** is designated as $\overline{X}$
- Calculated as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - X_i = the ith score
  - n = total number of scores

- sometimes just written as

$$\overline{X} = \frac{1}{n}\sum X_i$$

CALCULATIONS

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92

$$\overline{X} = \frac{1}{n} \sum X_i$$

$$\overline{X} = \frac{1}{4} [X_1 + X_2 + X_3 + X_4]$$

$$\overline{X} = \frac{1}{4} [95 + 89 + 94 + 92] = \frac{370}{4} = 92.5$$

COMPARISON

- mean can only be used on interval or ratio data.
- mode can be used on nominal data
- mode and median can be used on ordinal data
- mean can be manipulated mathematically
- mean can be sensitive to extreme scores

TAX BREAKS

Position	Number	Salary
President	1	$540,000
Ex. vice pres.	1	160,000
Vice pres.	2	140,000
Controller	1	52,800
Senior sales	3	50,000
Junior sales	4	42,800
Foreman	1	37,000
Machinists	12	25,000

- The company wants a tax break from the city. Is it a good corporate citizen?
- Mean = $67,640
- Median = $37,000
- Mode = $25,000
- The numbers answer different questions; which answer is best depends on what you care about.

MEAN OF MEANS

Name	Sex	Score
Greg	Male	95
Ian	Male	89
Aimeé	Female	94
Jim	Male	92

$$\overline{X} = \frac{\sum X_i}{n}$$

$$\overline{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

$$\overline{X} = \frac{95 + 89 + 94 + 92}{4} = \frac{370}{4} = 92.5$$

COMBINED GROUPS

- the mean of means is **not** the same thing as the mean of all the scores in all the groups!!!
- Consider our small data set

$$\overline{X}_F = \frac{94}{1} = 94.0$$

$$\overline{X}_M = \frac{95 + 89 + 92}{3} = 92.0$$

- the mean of the means is:

$$\frac{\overline{X}_F + \overline{X}_M}{2} = 93.0$$

- but we already found that the mean of all the scores was

$$\overline{X} = 92.5$$

- too much weight on the "female" group

COMBINED GROUPS

- correct calculation goes like

$$\overline{X} = \frac{n_F \overline{X}_F + n_M \overline{X}_M}{n_F + n_M}$$

- where
  - $\overline{X}_F$ is the mean for the females
  - $\overline{X}_M$ is the mean for the males
  - n_F is the number of females
  - n_M is the number of males

$$\overline{X} = \frac{(1)(94.0) + (3)(92.0)}{1 + 3}$$

$$= \frac{94 + 276}{4} = 92.5$$

- same as direct calculation of $\overline{X}$!

COMBINED GROUPS

- in general, given
  - $\overline{X}_i$ = individual group means
  - n_i = number of observations in individual groups
  - $N = \sum n_i$ = total number of observations in all groups

$$\overline{X} = \frac{\sum n_i \overline{X}_i}{N}$$

PROPERTIES OF MEAN

- The sum of deviations of all scores from the mean is zero.
- The sum of squares of the deviations from the mean is smaller than the sum of squares of deviations from any other value.

- deviations: data value minus mean

$$x_i = X_i - \overline{X}$$

- pluses and minuses cancel each other out!

$$\sum x_i = \sum (X_i - \overline{X})$$

$$= (95 - 92.5) + (89 - 92.5) + (94 - 92.5) + (92 - 92.5)$$
$$= (2.5) + (-3.5) + (1.5) + (-0.5) = 0$$

SUM OF SQUARES

- Let's ignore the **direction** of deviation, and consider the squared magnitude of deviation

$$\sum x_i^2 = \sum (X_i - \overline{X})^2$$

$$= (95 - 92.5)^2 + (89 - 92.5)^2 + (94 - 92.5)^2 + (92 - 92.5)^2$$
$$= (2.5)^2 + (-3.5)^2 + (1.5)^2 + (-0.5)^2$$
$$= 6.25 + 12.25 + 2.25 + 0.25 = 21.0$$

- sum of squared deviations from any other value is larger

$$\sum (X_i - 90)^2 = 25 + 1 + 16 + 4 = 46.0$$

$$\sum (X_i - 100)^2 = 25 + 121 + 36 + 64 = 246.0$$

- these properties will be important later!

CONCLUSIONS

- central tendency
  - mode
  - mean
  - median

NEXT TIME

variation
variance
standard deviation
z scores
How to make IQ scores look good.