

Dispersion

Why should we care?

- Central tendency, is that enough?
- E.g. Disadvantages are related to the problem of dispersion. The better the mean, more symmetric the distribution, which “mean” is sensitive to extreme values!
- E.g. We can make a lot of distributions with the same centrality measure!
- E.g. Sociological concerns: 20: 80 society! Who rules our society? Monopoly, centralization, inequality, diversity, homogeneity...
- The level of measurement again, as always!

Categorical Variables

- Let's think about heterogeneity of population. One group composed of 20 people using IBM and 10 people using Mac.
 - 1) When dispersion(variation) has maximum, the frequencies are equally distributed across categories: heterogeneity. In this case, 15 (IBM) and 15 (Mac)
 - 2) When dispersion has min, they are unequally distributed: homogeneity. In this case, all of students use IBM.

- Index of qualitative variation

$$\frac{k(N^2 - \sum f^2)}{N^2(k-1)}$$

- The basic idea is (total differences observed)/(maximum possible differences). In our case, total differences observed is $20 \times 10 = 200$. Maximum possible difference is $15 \times 15 = (30/2) \times (30/2) = 225$. Hence, IQV is $200/225$.
- If you extend this logic to categories more than 2, you can lead to the above equation.
- SPSS doesn't provide this measure.
- Skip the index of segregation.

Interval/Ratio Variables

- 1) Range=Max-Min. What's the problem? In terms of sensitive to what again?
- 2) Inter-quartile range= 75^{th} - 25^{th} . What's still the problem?
 - This is why another measure of dispersion based on the average deviation is needed.
- Q) What is the range of "age" in our data? What is the IQR of age in our data?



1 : year 1978

	year	id	marital	sibs	childs	age	educ	
1	1978	1967	2	5	4	26	10	
2							12	
3							12	
4							15	
5							12	
6							12	
7							4	
8							12	
9							12	
10							12	
11							16	
12							12	
13							12	
14							8	
15							16	
16							11	
17							16	
18							12	
19							18	
20	1978	963	1	8	2	45	9	

Frequencies

Variable(s): GSS YEAR FOR T AGE OF RESPONDEN

OK

Frequencies: Statistics

Percentile Values

- Quartiles
- Cut points for: 10 equal groups
- Percentile(s): 75

Add 25 50

Change

Remove

Central Tendency

- Mean
- Median
- Mode
- Sum

Values are group midpoints

Dispersion

- Std. deviation
- Variance
- Range
- Minimum
- Maximum
- S.E. mean

Distribution

- Skewness
- Kurtosis

Continue

Cancel

Help

- 1) Mean Absolute Deviation: How far are cases from the mean on the average?
 - SPSS does not provide this measure.
- 2) Variance(s^2): the square of the distance of each case from the mean instead of its absolute value.
 - Population (N) vs Sample (N-1)
 - Why is this measure more important than MAS?
- 3) Standard deviation(s): the average distance of each case from the mean. More easy to understand.
 - Why so important?: Besides dispersion, calculation of z-score and hypothesis testing in inferential statistics.
- Q) What is the variance and standard deviation of "tvhours" in our data?



1 : year 1978

	year	educ
43	19	12
44	19	11
45	78	16
46	59	14
47	27	12
48	30	12
49	39	14
50	33	5
51	26	16
52	24	16
53	2	12
54	1	12
55	0	12
56	2	12
57	3	12
58	2	14
59	2	13
60	3	14
61	0	12
62	1	8

Descriptives

Variable(s): # HOURS PER DAY WA

OK
Paste
Reset
Cancel
Help
Options...

Descriptives: Options

Mean Sum

Dispersion

Std. deviation Minimum

Variance Maximum

Range S.E. mean

Distribution

Kurtosis Skewness

Display Order

Variable list

Alphabetic

Ascending means

Descending means

Continue
Cancel
Help

- 4) Coefficient Variation: $SD/mean$ or $SD*100/mean$, which means the “relative” deviation from the mean.
- Why important? (e.g.) The standard deviation of the income in two countries (Mexico and US) is the same, say 1000, but the average income is 10,000 and 50,000. What do you think which country is more homogeneous? (Hint) The meaning of 1000 depends on the magnitude of observations. 1000 might be big enough for poor country.
- Q) What is the CV of “tvhours” in our data? (SPSS does not provide this measure separately because we can calculate this easily when you know mean and SD)