3190 Week 2

Describing data

Ordered arrays

- The simplest organizational tool for working with data is to order it
- An ordered array is a list of numerical values associated with a variable in rank order from the smallest value to the largest value
- So, are the unemployment data an ordered array?

Frequency Distributions

- These are tables of your grouped data that show the frequency of cases in each group
- The groups are in the left column
- The frequencies are in the adjacent column to the right
- Percentages are in a third column to the right

Central Tendency

- These are calculations that represent the central or typical value in a distribution of values for a variable
  - Three types
    - Mean
    - Mode
    - Median

Measures of Dispersion

- Calculations that depict the amount of spread or variability in a set of data values of a variable
  - There are several
    - The range
    - The interquartile range
    - The variance
    - The standard deviation
    - The coefficient of variation

The Mean

- The arithmetic average of a sample calculated as summation of the scores divided by the number of cases for a variable
  - Notation
    - $\sum$ = "sum of"
    - x_i = "a particular score for a case"
    - n = " total number of cases" or "sample size"
    - μ = "population mean"
    - $\overline{X}$ = "sample mean"

$$\overline{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Why different symbols…?

- It is important to note that a sample is a sub-set of values from a population

- Thus, any statistic calculated from a sample is an estimate of the population parameter.

- So, we use two symbols to keep them straight; we rarely have population level data

Unweighted vs. weighted mean

data from Table 2.6, Table 3.2 McGrew & Monroe

- When we use the raw data for a distribution, as the formula indicates, the mean is unweighted (the normal way of calculating)
    - = 1598/40 = 39.95 inches
- When we use grouped data, we weight the mean.

$$\overline{X}_w = \frac{\sum X_j f_j}{n}$$

- Simply multiply the class midpoint by the frequency of cases in each class
    - = 1610/40 = 40.25 inches

TABLE 2.6

Annual Precipitation for Washington, D.C.: A Ranked 40-Year Record (in Inches)

26.87	35.20	39.86	45.62
26.94	35.38	40.21	46.02
28.28	35.96	40.54	47.73
29.48	36.02	41.11	47.90
31.56	36.65	41.34	48.02
32.78	36.83	41.44	50.50
33.07	36.99	41.46	51.17
33.62	38.15	41.94	51.97
34.98	39.34	43.30	54.29
35.09	39.62	43.53	57.54

Source: National Climatic Data Center, U.S. Dept. of Commerce.

TABLE 3.2

Worktable for Calculating Weighted 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

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

The Weighted Mean

- Why is it weighted? Because it is not based on the original data distribution.

- Comes in handy when all we have to work with is a frequency distribution, a histogram, or a frequency polygon

That is, when we do not have access to the original data

The mode

- The mode is the most frequently occurring score in a data distribution
- Often there is no mode, or there is more than one mode in a distribution
- There is no mode in Table 2.6 because each score is different

The crude mode

- This is the “weighted” mode, based on grouped data; it is the midpoint of the most frequently represented group

- Look at Table 3.2, which group has the most members?
  - Although there was no mode; there is a crude mode for the grouped data
  - It is 37.5 inches

The median

- An ordered array's middle value; the value with an equal number of cases above and below it
- When there is an odd number of cases, it is the middle value, aka the 50th percentile
- When there is an even number of cases it is half way between the two middle scores
  - Easily, efficiently calculated as $$Median = \frac{(n+1)}{2} \text{ ordered observation}$$

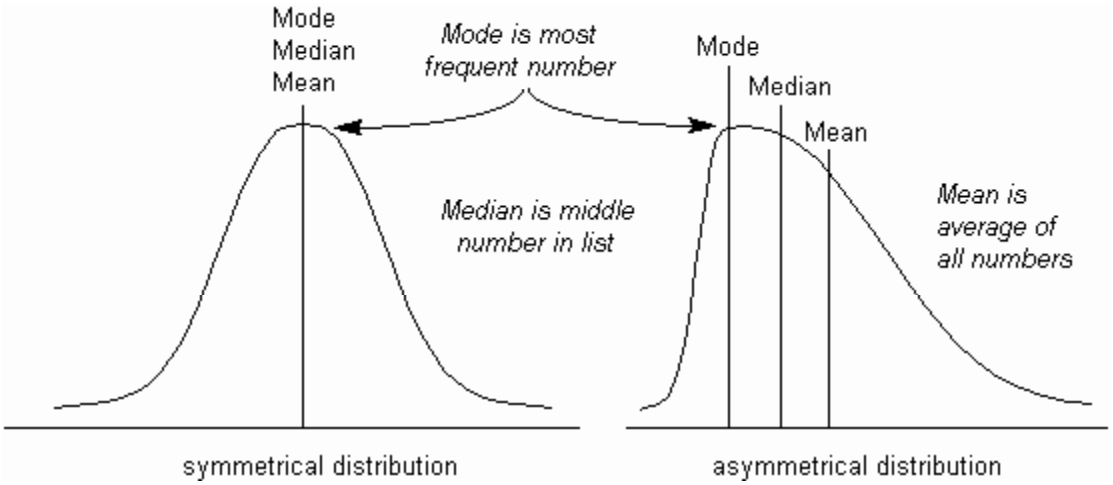
An example

- For Table 2.6
  - (40 + 1) / 2 = position 20.5
  - So the median score is halfway between the scores at 20 & 21
  - **Score at P20 = 39.62; at P21 = 39.86**
  - Sum the two scores, and divide by 2 to get the score of the median

= 39.74 inches

Which one should I use…?

- The mean, median, & mode are the same in a perfectly symmetrical distribution
- Mode is not useful if the distribution is multimodal



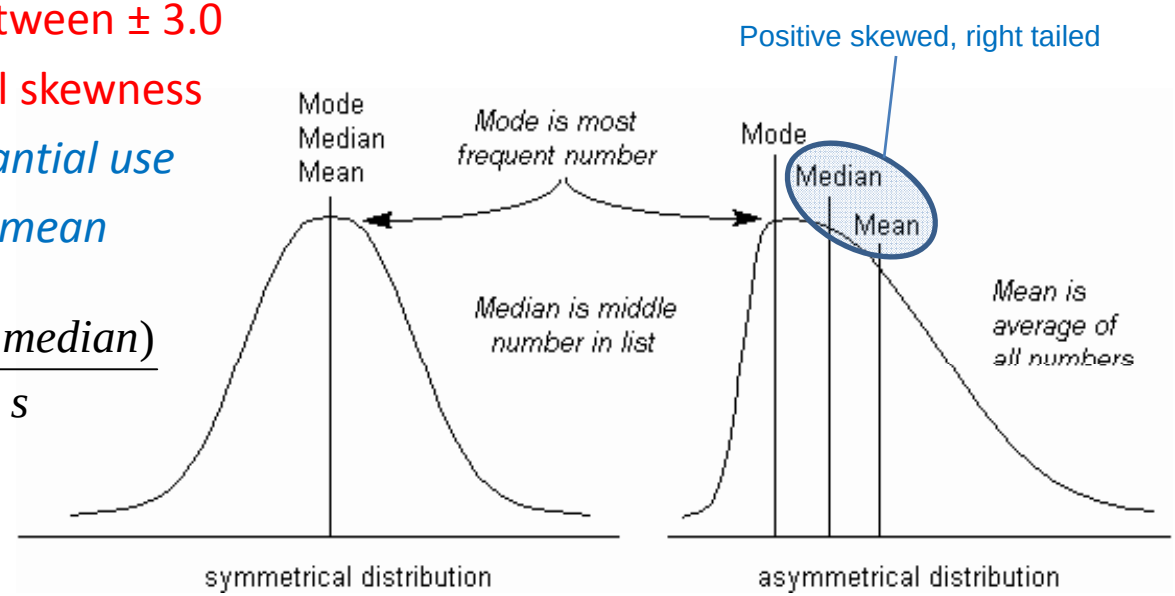
svgue.com/improvement/Median.htm

- The mean is the preferred measure <u>for symmetrical</u> (or nearly symmetrical) <u>distributions</u> because there are many statistical inference tools designed specifically for analyzing means later in the course (see syllabus)

Pearson's skewness

- Based on the logic that the mean is more affected by skewness than the median
  - So we can use the difference between them to assess severity and direction of skewness
  - Pear. Skw. varies between ± 3.0
  - > ± 0.6 is substantial skewness
  - *If skewness is substantial use the median not the mean*

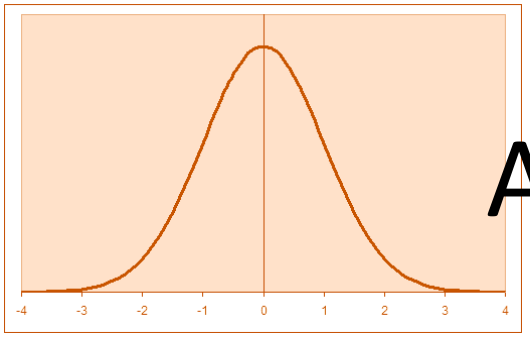
$$Pearson's \quad skewness = \frac{3(\overline{X} - median)}{s}$$



Important terminology

- Mean, standard deviation (based on mean), all referred to as "**parametric**," "interval-scale" or "absolute" statistics in this course
  - Based on original data scores in data distributions
  - Related inferential tests use more assumptions (e.g., symmetry, normality)

- Median, interquartile range, range, ordered array, Pearson's skewness all referred to as "**non-parametric**," "ordinal-scale" or "relative" statistics in this course
  - Relies on positions or scores at positions in ordered arrays
  - Related inferential tests use few assumptions (does not assume normality)



Assuming normality

- To assume normality is to assume that the shape of the distribution of a variable for a population is unimodal & symmetrical
  - remember “parametric” = “about population”
- We would like to be able to assume normality
  - Then we can use parametric statistics, which are more powerful
  - More powerful because we can use the normal probability distribution to make predictions
- If our sample is random, we can assume normality at samples $n \geq 30$, why?
  - We will discuss normality in more detail later on, but learn these basics

Dispersion

Non-parametric & parametric alternatives

The five points of data summary

- With 5 **non-parametric stats** we can learn a lot about a dataset
  - The minimum (score at the lowest position)
  - The 25th percentile (the score at the position with 25% of the cases below it; 75% above)

    25th percentile = **score** at position (n+1)/4
  - The median (50th percentile; the score at the position with 50% of the cases above & below it)

    Median = **score** at position (n+1)/2
  - The 75th percentile (the score at the position with 75% of the cases below it; 25% above)

    75th percentile = **score** at position 3(n+1)/4
  - The maximum (**score** at the highest position)

TABLE 2.6

Annual Precipitation for Washington, D.C.: A Ranked 40-Year Record (in Inches)

26.87	35.20	39.86	45.62
26.94	35.38	40.21	46.02
28.28	35.96	40.54	47.73
29.48	36.02	41.11	47.90
31.56	36.65	41.34	48.02
32.78	36.83	41.44	50.50
33.07	36.99	41.46	51.17
33.62	38.15	41.94	51.97
34.98	39.34	43.30	54.29
35.09	39.62	43.53	57.54

Source: National Climatic Data Center, U.S. Dept. of Commerce.

TABLE 3.2

Worktable for Calculating Weighted 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

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

Measures of Dispersion

- Calculations that depict the amount of spread or variability in a set of data values of a variable
  - There are several
    - The range
    - The interquartile range
    - The variance
    - The standard deviation
    - The coefficient of variation

The Range

- The simplest measure of variability or dispersion
  - Simply subtract the minimum score from the maximum score
  - Precipitation data = 57.54 – 26.87 = 30.67

- Weakness is that extreme scores can be misleading.
  - Let's say we add year 41 & its precipitation is 72.13 inches
  - The range is now 72.13 – 26.87 = 45.26
  - But most of the scores are not different than w/o year 41

Interquartile Range (IQR)

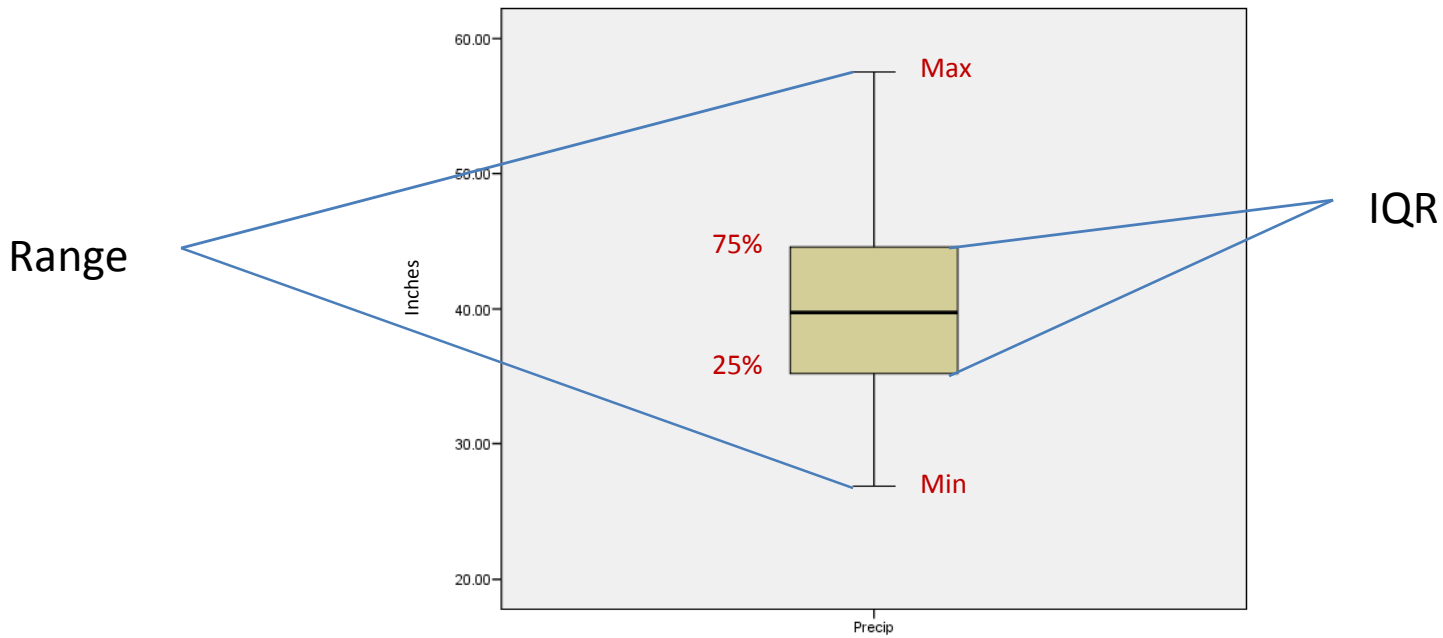
- The difference between the 75th and the 25th percentile… the "middle half of the data"
  - Precip. data 75th = P30.75 = 45.10 inches
  - 25th = P10.25 = 35.12 inches
  - IQR = 9.98 inches

- Add a 41st year to the dataset at 72.13 and
  - 75th = 45.82; 25th = 35.145; IQR = 10.675 inches
  - So the IQR did not change as much as the range

When to use…

- Use the range and IQR as non-parametric descriptions *when you do not want to assume the data are symmetrical*

- Very handy to use with the 5 points of data summary & boxplots



Parametric Measures of Dispersion

- These rely on <u>deviation</u> from the mean of scores for cases in a sample
  - Notation $x_i - \overline{X}$ where X_i is a score for a case
  - To calculate the "deviation" of a score from the mean you simply subtract the mean from it
    - If the score is > the mean it will be +
    - If the score is < the mean it will be –

  - **We are interested in a <u>summary</u> of how all scores for cases together vary about the mean**
    - **How can it be determined?**

The Variance

- The average squared deviation of scores around $\bar{X}$
- What is an average (mean)? $\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$
  - The sum of scores ÷ the n of cases
  - But we are not interested in *average scores* (here), but average deviation of scores from the mean

S^2 = variance

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

These equations are <u>similar</u>

Sum of scores for all cases

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

Number of cases
(sample size)

Sum of deviations for all cases

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Number of cases

Two questions:

1) Why n-1?
2) Why do we square S?

The Variance

- The sum of raw deviations is always 0
  - Of no use to us because – outweigh +
  - Must square to get rid of –

Pencil	Length (inches)	Deviation	Deviation^2
1	10	10 - 3.7 = 6.3	39.69
2	4	4 - 3.7 = 0.3	0.09
3	2	2 - 3.7 = -1.7	2.89
4	1.5	1.5 - 3.7 = -2.2	4.84
5	1	1 - 3.7 = -2.7	7.29
	Mean 3.7	Sum 0	Sum 54.8

$$S^2 = \frac{54.8}{4}$$

$$= 13.7$$

  - Then we divide by sample size (n-1)

The Variance

- So for two samples with the <u>same</u> mean…
- The higher the variance, the greater the dispersion around the mean
  - It is always a positive number

- But, the variance is in "units^2"
  - It is <u>average squared deviation</u>
  - Inches^2 in our pencil example
- It is of greater use to us if we change that

The Standard Deviation

- We can take the square root of the variance to get *back to the original units*
  - We still have a good measure of dispersion because we summed <u>before</u> we took the square root
  - Called the Standard Deviation
    - $S^2 = 13.7$, S = 3.701 inches
      - A pencil of 7.401 inches is 1 S above the mean (= 3.7)
      - This is a better measure of average deviation (not squared)

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

An example

- Let's say you sample two counties (A & B) for chlorine-contaminated water wells
- n_A = 30 wells; n_B = 30 wells
  - Same mean of 1250 mg/L
  - So, it should require the same effort to clean up both counties, correct?
    - Not necessarily… S_A = 125 mg/L but for S_B = 410 mg/L
    - Because S_B is quite a bit higher, there are higher (and lower) deviants in the sample (the sample is more dispersed)
      - Those high concentration sites may be very difficult to clean up

- But this direct comparison with S only works **<u>if the means are similar</u>…**

Test 1

Student	Test Scores A	Deviation	Deviation^2
Athena	80	30	900
Achilles	65	15	225
Zues	50	0	0
Aphrodite	35	-15	225
Hercules	20	-30	900

Mean	Sum	Sum
50	0	2250

S^2	S
562.5	23.71708245

Student	Test Scores B	Deviation	Deviation^2
Mercury	51	1	1
Apollo	50.5	0.5	0.25
Hera	50	0	0
Socrates	49.5	-0.5	0.25
Neptune	49	-1	1

Mean	Sum	Sum
50	0	2.5

S^2	S
0.625	0.79056942

Test 2

Student	Test Scores A	Deviation	Deviation^2
Athena	90	37.4	1398.76
Achilles	65	12.4	153.76
Zues	50	-2.6	6.76
Aphrodite	35	-17.6	309.76
Hercules	23	-29.6	876.16

Mean	Sum	Sum
52.6	0	2745.2

S^2	S
686.3	26.19732811

Student	Test Scores B	Deviation	Deviation^2
Mercury	100	47.4	2246.76
Apollo	92.5	39.9	1592.01
Hera	88	35.4	1253.16
Socrates	85	32.4	1049.76
Neptune	49	-3.6	12.96

Mean	Sum	Sum
82.9	0	6154.65

S^2	S
391.05	19.7749842

The Coefficient of Variation

- We can make S comparative by using it to calculate the CV

$$CV = \frac{S}{\overline{X}} \cdot 100\%$$

- The coefficient of variation expresses sample standard deviation as a <u>percentage of the sample mean</u>
  - Answers, relative to the mean, how large is S?
  - Because it is a relative measure, samples with unequal means can be compared in terms of dispersion

Test 2

Student	Test Scores A	Deviation	Deviation^2
Athena	90	37.4	1398.76
Achilles	65	12.4	153.76
Zues	50	-2.6	6.76
Aphrodite	35	-17.6	309.76
Hercules	23	-29.6	876.16

Mean	Sum	Sum
52.6	0	2745.2

S^2	S	CV
686.3	26.19732811	49.8048063

Student	Test Scores B	Deviation	Deviation^2
Mercury	100	47.4	2246.76
Apollo	92.5	39.9	1592.01
Hera	88	35.4	1253.16
Socrates	85	32.4	1049.76
Neptune	49	-3.6	12.96

Mean	Sum	Sum
82.9	0	6154.65

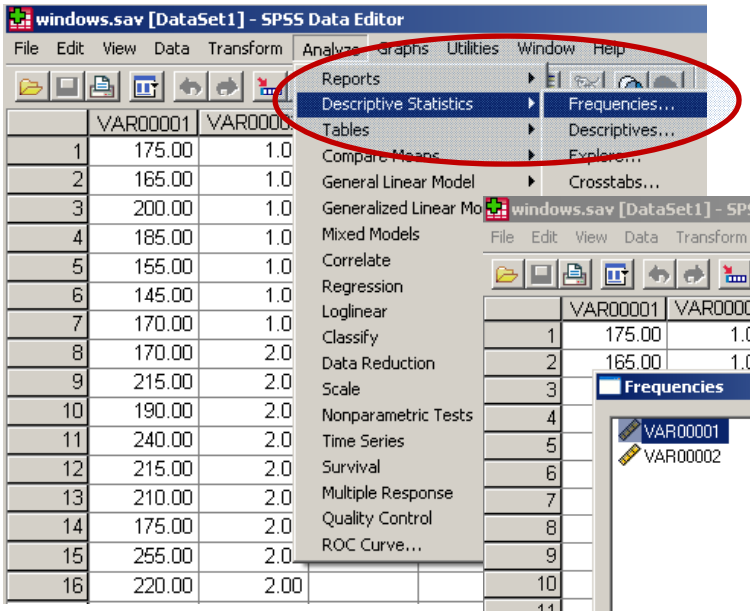
S^2	S	CV
391.05	19.7749842	23.85402195

Why describe dispersion (variability)

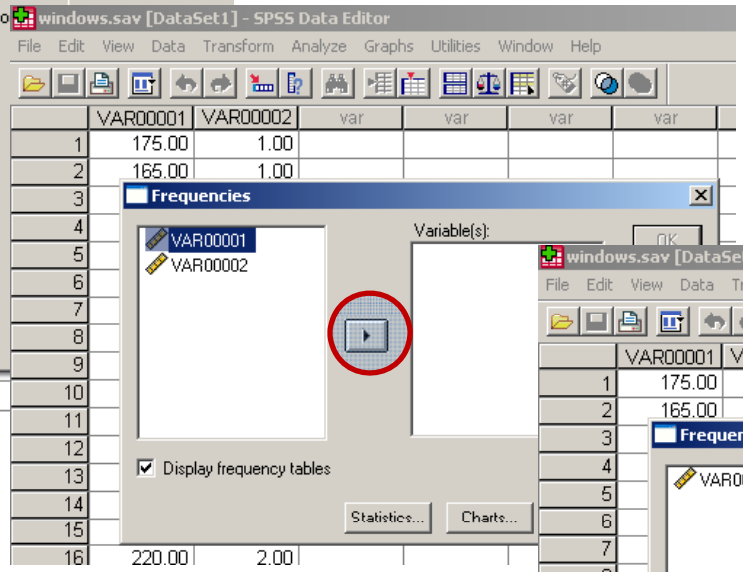
- Allows us to determine between two or more samples, which one has a broader dispersion of scores
  - In grading, if one lab is highly variable and one is not there might be many reasons
    - Attendance is poor in the early lab, but those who come to class get more attention (fewer students) = broad variability in scores
    - More people come to the second lab, but there is less attention = narrow variability
    - Or, the professor is tired early on, and only mathematically gifted students understand him/her, but others do poorly = broad variability
    - The second lab gets the professor when he/she is more awake, and all students do better with out the broad dispersion of scores = narrow variability
- Allows us to target reasons for differences between samples (deer size)

Basic Descriptives in SPSS

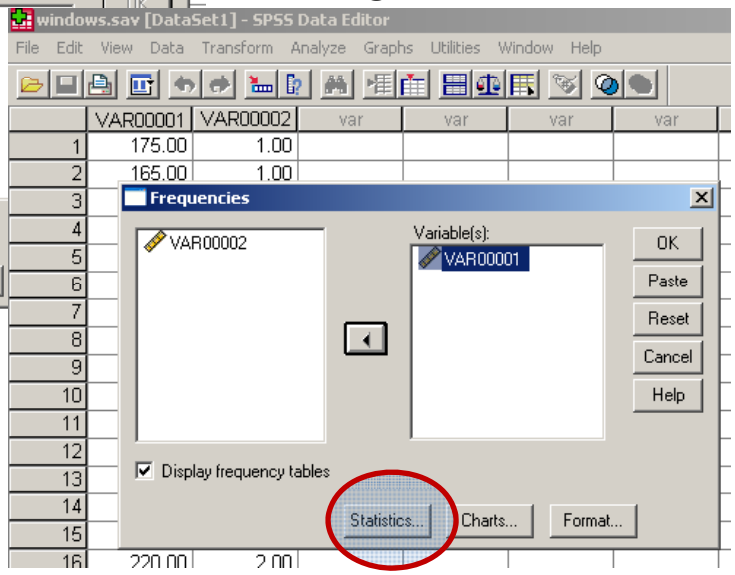
1



2

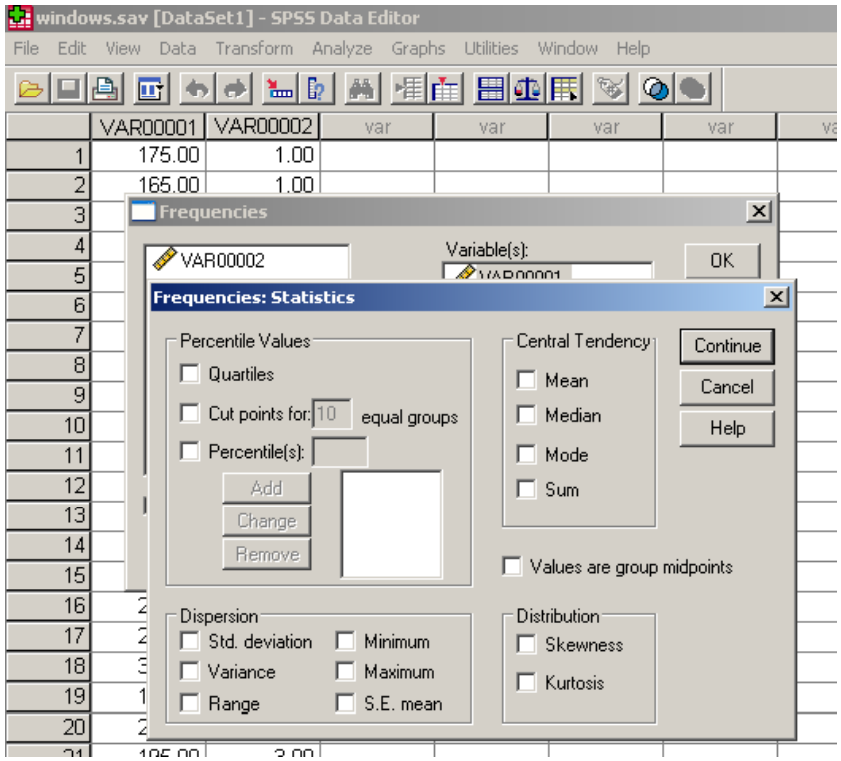


3

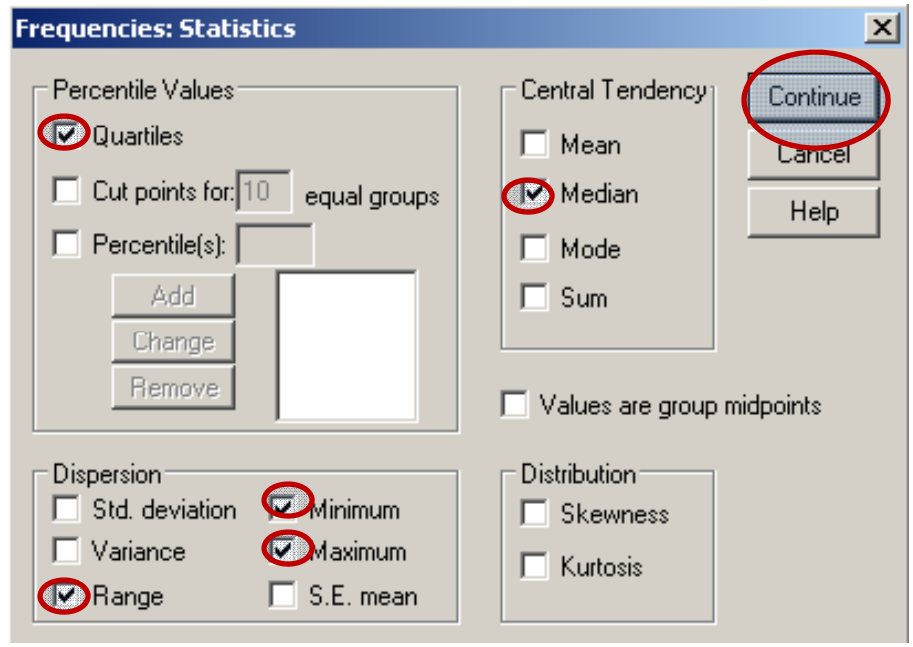


5 pts of data summary (& range)

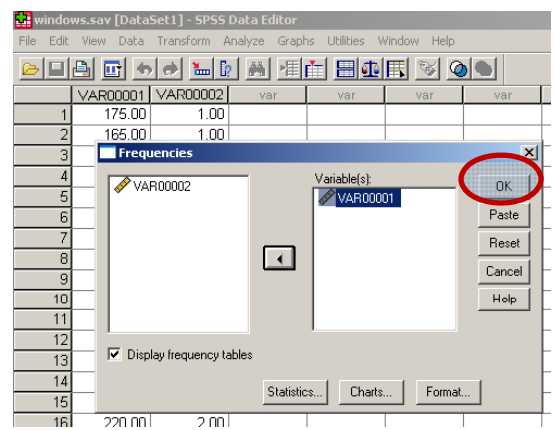
1



2



3



Output

Statistics

VAR00001

N	Valid	23
	Missing	0
Median		200.0000
Range		280.00
Minimum		145.00
Maximum		425.00
Percentiles	25	175.0000
	50	200.0000
	75	240.0000

Now, calculate the IQR

Then, redo the descriptives using parametric statistics