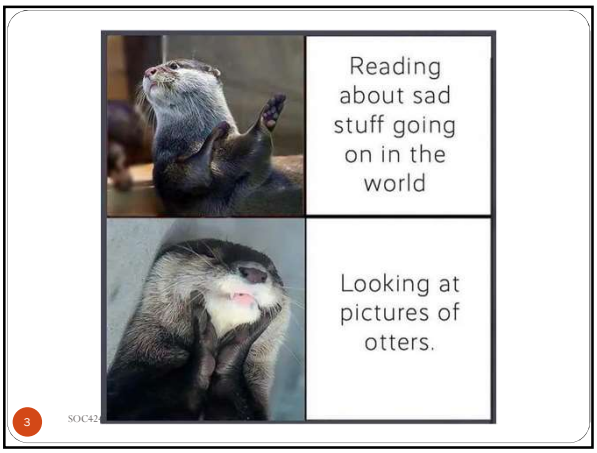


Schedule Pointer

- **Submit HW1 asap if not already done!**
- **Should be able to do HW2 *and* HW3 after Lab 5**
- **Don't delay homeworks –**
  - deadlines are later by design – I *expect* you to beat some!
  - 5% penalty per class meeting late – 2% bonus per class early!

SOC424 – Statistics w/ Dr. Ellis Godard

Dispersion:
Measuring Variation Among & Across Values

What's New & What Isn't

- **You learned those stats in 5th grade** (maybe 4th? 6th)
- But I recognize that **there *are* new things** about them now
  - **Opening a dataset** (FILE>OPEN>DATA)
  - **Getting those stats from SPSS**
    - Always check 6: mean, median, mode, quartiles, std dev, & range
  - **Choosing variables** (nominal, ordinal, & interval)
    - Don't use the Measure column – look @ the Values column
  - **Choosing which statistic to use for which variable**
    - Guidance throughout (and again @ the end of) this lecture
  - **Putting them in a meaningful sentence**
    - Not for a mathematician – for a relative, boss, neighbor, etc.
    - Reporting & Interpreting them – not just the number; what's it *mean?*
    - Describe the sample / Summarize the data

Outline for Today…

- **Admin matters**
  - Schedule Pointer
  - Reminders about what's new (& what isn't)
  - Assignment Pointer
  - Roadmap
  - CT Review
- **Dispersion – long!** *(breathe deeply & roll your neck)*
  - Concept of Dispersion
  - Measures of Dispersion
  - Calculated Example
  - Illustrated Example
  - SPSS Demonstration – Shape, CT, Dispersion
- **Lab #5 – Dispersion in SPSS – IT HAS TWO PARTS!!!**

Assignments Pointer…

- **Example Instruction:**
  - Report & interpret the central tendency of an interval variable.
- ***Thru lab 5, you might have said…***
  - For MAEDUC, mean is 12.4 and median 12. Skewed.
  - That has the pieces and values, but is incomplete.
  - You must *use* the data to *describe* the sample!!
- **Say something like this** (which is complete & descriptive):
  - The mean education of respondents' mothers (that is, the arithmetic average of case scores) was 12.4 years, while the median (that is, the value of middle case when the cases are ordered) was 12. A mean higher than a median indicates that the distribution is skewed to the right, possibly due to some outliers (that is, particular mothers in the sample who received much more education than the typical respondents' mother).
- **Report & Interpret are *not* the same thing!**
  - Interpret = make sense of it (middle = HS; but some PhDs?)

Roadmap up to this point...

- **Building towards inference**
  - **Saying things about population based on sample distribution**
- **2 main ways to describe distribution**
  - **Central Tendency:** What is typical?
  - **Dispersion:** How far spread out are they?

CENTER vs. SPREAD

- Purpose of both is to describe the shape of a distribution with one or two numbers, rather than some vague idea about the shape and rather than having to list all of the cases
- If they have the same shape but different middles, the mean and standard deviation will tell us that

CENTRAL TENDENCY

Three ways to measure typicality:

- **Mode**
  - most frequently observed value
- **Median**
  - middle value, when cases are ordered by values of that variable
- **Mean**
  - arithmetic average, "center of gravity"

What each asks about cases...

Central Tendency	Dispersion
What's most common?	How much do they differ?
What's the majority/plurality?	How well are they balanced?
Can we identify an exemplar?	Focus on the periphery.

DISPERSION: Concept

Consider 2 samples, spread 0 to 20, and 6 to 14…

Sample 1 (more varied)		Sample 2 (more clustered)	
0	12	6	10
2	14	6	12
4	16	8	12
6	18	8	14
8	20	10	14
10		10	

For either, the median is 10, and the mean = 110 / 11 = 10.
The samples differ – but the central tendencies don't!

DISPERSION: Measures

- **Need another way to measure distributions**
  - Central tendency only looks @ middle/typical
  - Ignores spread, extremes, endpoints…
- **4 ways to measure the spread of a distribution**
  - **Range** or **IQR** (and IQR usually better)
  - **Standard Deviation** (built from Variance, but that's *not* a choice)
  - **Variation Ratio** (not in text, and not labelled in SPSS)
- **For all four:**
  - The bigger the value of the statistic, the more dispersed the data are

Criteria for Selection

Same as central tendency (see those notes):

- Level of measurement (!)
- Shape of distribution (!)
- Robustness
- Efficiency
- When in doubt, use several (but *not all 3!)*
  - Using all 3 tells me that you don't know how to pick
  - You need to be showing me that you *can* pick

RANGE: Problems

- **Doesn't use all data, just the extremes**
  - only uses 2 cases to ID what's typical
  - very affected if those extremes are outliers
  - May simply measure # of values used (~ IQV…)
- **Only shows maximum spread**
  - not how tightly the data is clustered
  - doesn't fully represent the *shape* of the distribution
  - could be U or V
- **Must be able to rank order the data**
  - so can use for ordinal, but not nominal
  - use for interval if sample range is less than pop range

RANGE: Meaning

- **Distance between Maximum and Minimum**
  - Largest *value observed in the sample* minus Smallest *value observed in the sample*
  - *Not* the smallest and highest frequency or percentage
  - *Not* the smallest or largest value in the *measurement*
    - e.g. Age 10 to 90, but no cases are over 30
- **Advantage:**
  - Easy to compute and understand
  - Best option for ordinal data

Note on Reporting

- What it means:
  - The range is NOT a measure of what the typical case is or does
  - That's the mean (and other measures of CT).
  - Measures of dispersion assess how much cases DIFFER from each other
- What the value is:
  - So, if the range is 3, do NOT report the label for the third value (e.g. "$30-39K).

(More) Warnings about the Range

- The range is *not* a particular value or category
  - For example, satisfaction from 1-5
    - Very Un, Unsatisfied, Neutral, Satisfied, Very Satisfied
  - The range is 5-1=4
    - That's it – *just 4* – **not** "Satisfied"
- SPSS provides range (& any stat) for any variable
  - 1=male, 2=gender -> range is 1?
  - 1=Baptist,2=Catholic,3=Athiest -> range2?
    - No Baptists -> range 1?
  - But nominal values aren't ordered; can't measure spread the same way (along a line)

INTERQUARTILE RANGE (IQR)

- **Width of the middle 50% of cases**
  - Width between the upper and lower quartiles
  - Distance from 25th to 75th percentile
    - Range of *values* between those *cases*
- Example, if 11 cases:
  - 25th percentile = (n*.25)th case
    - eg 0.25 * 11 = 2.75 -> 3 (round up)
  - 75th percentile = (n*.75)th case
    - eg 0.75 * 11 = 8.25 -> 9 (round up)

IQR: More Examples

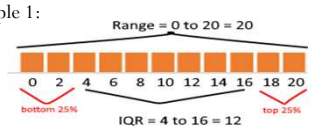
- **Examples, from earlier slide:**
  - Sample 1 = 16 – 4 = 12
  - Sample 2 = 12 – 8 = 4
- **In SPSS:**
  - Could use EXPLORE
    - (but *don't* – we'll use that for something else *later*)
  - Easiest to get quartiles (four percentiles) from ANALYZE > DESCRIPTIVES > FREQUENCIES / "Statistics" and then subtract the 25th percentile (such as 4 or 8, above) from the 75th (such as 16 and 12, above)

RANGES: IQR

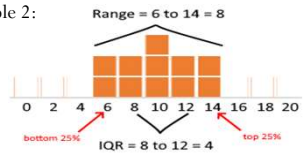
- **Advantages**
  - more stable than range b/c ignores outliers/extremes
  - More useful if have more cases
- **Disadvantages**
  - not using all the information
    - outliers matter sometimes, tails almost always do
    - ignores half the cases, not just outliers
  - Harder to calculate/understand (vs range)
- **When to use**
  - Better than range for interval, esp. if heavily skewed
    - Possibly also better than std. dev.?
  - Good for ordinal, esp. if observed range = value range
  - ordinal too if sample not small & range not short

Range vs. IQR

- Sample 1:



- Sample 2:



DEVIATIONS

- A "deviation" is a difference or distance
  - e.g. how far a given score is from a mean
- It's an error, in a sense
  - if we use the mean to summarize the data, it differs from each case by some amount
- There will always be deviations
  - Otherwise, everyone would have the same score
- Want to talk about the *typical* deviation
  - Could sum and divide by n if wanted the average deviation and didn't need to make inferences from the data
  - But there's almost no use for variance w/o inferences, so we use n-1 for denominator – more on that later, w/ dfs

IQR: Evaluation

- Probably better than range, especially for ordinal
  - For ordinals, the range of the *sample* is often the same as the range of the *variable*, if there's a least one case for each value (e.g. on a scale of 1-5)
  - We want to know how spread out the *data* are, how much the *cases* differ, not how much variation was permitted by the measurement(s) taken
  - The IQR focuses on the middle, and better emphasizes the spread of *cases*

DEVIATIONS

- **Variance** $s^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$
  - Avg. squared deviation from mean
  - Sample "s" estimates the population σ^2
  - always 0 or more; never negative
    - if s=0, then there's no variation in sample
- **Standard Deviation** $s = \sqrt{s^2} = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$
  - Square root of variance
  - "typical" deviation from the mean
  - Sample s^2 estimates population σ

Interpretations of Deviations

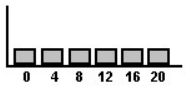
- **Variance has NO clear interpretation**
  - Later, we'll use it to build more powerful stats
  - For now, it's just a step to std. deviation..
- **Interpret stdev in terms of the mean**
  - add or subtract it from mean-
    - e.g. what values are 1 SD above or below the mean?
  - interpret IQR in reference to *median* (50th percentile)
    - Std dev is much more useful than IQR
    - tells us almost everything need to know about deviation

Standard Deviation

- **Now: Use to compare one sample to another**
  - sample 1 = s = 6.6; sample 2 s = 2.8
    - data more clustered together in second example
  - If sample standard deviation is larger, then we probably suspect that the pop's is too
- **Coming: individual scores in *terms* of stdevs**
  - Subtract each from mean, divide by Std Dev
  - If do for a full set of data (all cases for a variable), then you've *standardized the* scores for that variable
- **Empirical rule:**
  - Approximately 68% of all cases fall within 1 std dev of the man
  - Approximately 95% fall w/i 2 stdevs

Illustrated Example

- **Three families w/ different TV habits**
  - **Family A:**
    - **Everyone watches a different # hours/week**



  - **Family B:**
    - **No one watches 0 or 20**

  - **Family C:**
    - **Everyone watches 0 or 20**

CENTRAL TENDENCY

- **Means:**
  - Family A: (0+4+8+12+16+20)/6 = 60/6 = 10
  - Family B: (4+8+8+8+12+12+16)/6 = 60/6 = 10
  - Family C: (0+0+0+20+20+20)/6 = 60/6 = 10
- **Medians:**
  - For each, average the middle two & get 10
- **But the distributions clearly differ!**
  - Same problem as start of lecture
  - Need measure of dispersion, not just C.T.
  - Ready for some math? ☺

Dispersion – Family A

Value (Y_i)	Deviation: (Y_i - Mean)	Squared deviation: (Y_i - Mean)^2
0	0- 10 = -10	100
4	4- 10 = -6	36
8	8 – 10 = -2	4
12	12 – 10 = 2	4
16	16 – 10 = 6	36
20	20 – 10 = 10	100
Sum= 60	Sum = 0	Sum = 280

- **Variance = [Sum of (Mean-Y_i)^2] / (n-1) = 280/5 = 56**
- **Std dev = Square root of variance = 7.48**

Dispersion – Family B

Value (Y_i)	Deviation: (Y_i - Mean)	Squared deviation: (Y_i - Mean)^2
4	4 – 10 = -6	36
8	8 – 10 = -2	4
8	8 – 10 = -2	4
12	12 – 10 = 2	4
12	12 – 10 = 2	4
16	16 – 10 = 6	36
Sum= 60	Sum = 0	Sum = 88

- **Variance = [Sum of (Mean-Y_i)^2] / (n-1) = 88/5 = 17.6**
- **Std dev = Square root of variance = 4.19**

Intro Dispersion Range IQR Deviation Example Std.Dev. Var.Rat. Summary

Dispersion – Family C

Value (Y_i)	Deviation: (Y_i - Mean)	Squared deviation: (Y_i - Mean) ^{2}
0	0 – 10 = -10	100
0	0 – 10 = -10	100
0	0 –10 = -10	100
20	20 – 10 = 10	100
20	20 – 10 = 10	100
20	20 – 10 = 10	100
Sum=60	Sum = 0	Sum = 600

- **Variance = [Sum of (Mean-Y_i)^2] / (n-1) = 600/5=120**
- **Std dev = Square root of variance = 10.95**

Intro Dispersion Range IQR Deviation Example Std.Dev. Var.Rat. Summary

The Idea of "Standard Deviation"

A standard deviation is the "average" (mean) deviation from the mean.

The larger the standard deviation, the more dispersed the distribution.

A "short and squat" distribution has a larger standard deviation than a "tall and thin" distribution.

Intro Dispersion Range IQR Deviation Example Std.Dev. Var.Rat. Summary

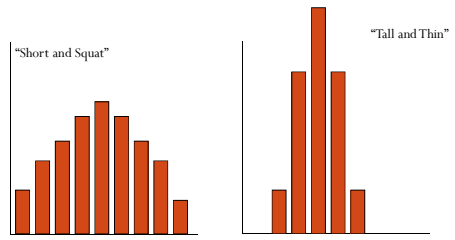
Simple Summary

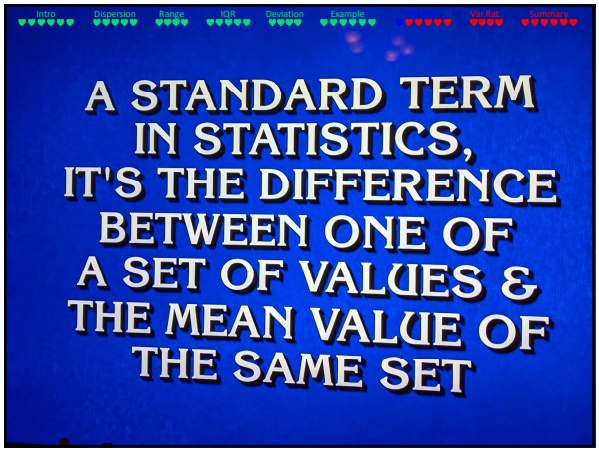
	Mean	Stddev
Family A:	10	7.50
Family B:	10	4.19
Family C:	10	10.95

Family C's TV-viewing habits are the most dispersed, Family B's the least

Intro Dispersion Range IQR Deviation Example Std.Dev. Var.Rat. Summary

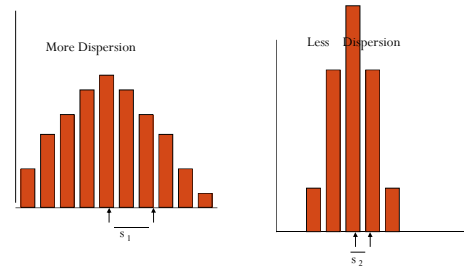
The left histogram is more dispersed





Intro Dispersion Range IQR Deviation Example Std.Dev. Var.Rat. Summary

Compare their standard deviations



$s_1 > s_2$

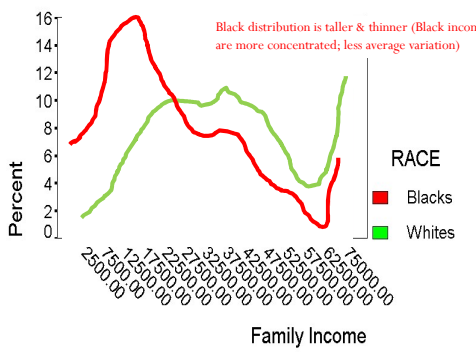
When is stddev "large" or "small"?

- No single, simple answer.
  - Does *not* mean much on its own
- Can say something about magnitude.
  - But conclusions must be based on substantive & comparative judgments.
    - Compare to mean?
    - Compare to another sample?
    - Understandable distances – dollars, years
  - Key: Adds something *beyond* central tendency

VARIATION RATIO: *Not* Variance. Really!

- **NOT same as "variance" – don't get 'em confused**
  - **Seriously…**
  1. **It was on the previous slide.**
  2. **"Variance" and "Variation" are *not the same word***
  3. **"Variance" isn't really much of a *ratio***
  4. **SPSS doesn't explicitly label the variation ratio as the variation ratio, anywhere… so don't look for it by that name… and don't pick "variance" as if that was the same thing… it's not…. I promise :)**

Race Differences in Family Income Distribution



VARIATION RATIO: Examples

- **Interpretation:**
  - As with any measure of dispersion, the higher the value of the variation ratio, the less concentrated the data are
- **Examples:**
  - 54% female -> variation ratio is 46%
  - 72% Caucasian -> variation ratio is 28%
- **Extremes:**
  - If the mode is only 1% of the sample, the variation ratio is 99% – very dispersed
  - If the mode is 100% of the sample, the variation ratio is 0% – not dispersed at all

VARIATION RATIO: Basics

- **NOT same as "variance" – don't get 'em confused**
  - Don't check the box for "variance" in SPSS. Never!
- **Uncommon**
  - **Not in most texts** (including ours)
  - **Not labeled in SPSS** (you must calculate it)
- **Q: How many cases are *not* the mode?**
  - VR = the total % of all non-modal cases
  - Add up all the others, or subtract the modal percentage from 100%
- **Use for nominal, b/cuz nothing else works**
  - Alternative (?): Index of Qualitative Variation (IQV)
    - Only assesses which values observed
    - Does not assess distribution of cases *across* those values

More VR Notes:

- If all values have the same relative frequency
  - No mode. All cases (100%) are nonmodal. VR is 100%.
  - As dispersed as possible, using this measure of dispersion (though 2/5/10 is more dispersed than 2/3/4 using std dev)
- If all cases have same value
  - They're all the mode. None (0%) are nonmodal.
  - As concentrated (& as far from dispersed) as they could be
- All variables have a variation ratio
  - Not just nominals
  - All have some % that is nonmodal (even if 0%, or 100%)
  - But VR not as useful for non-nominals
- Calculation reminder
  - Sum the percentages of nonmodal cases, not actual values

Appropriate Measures...

- **Variation Ratio** – really only for nominal
- **Range** – preferably interval
  - may use for ordinal, but that may be suspect unless sample range = full range of the measurement
- **IQR** – at least ordinal, good if interval skewed
- **Stdev** – at least interval
  - works best w/ bell-shaped distribution
- If bimodal, nothing really best; if interval, compute & report all dispersion measures

Lab Overview

- **Last lab (4) - SPSS for descriptive stats**
  - Central tendency for 3 variables
- **This lab (5) has two parts**
  - Not two labs – **there are two parts of Lab #5!**
  - Some math (yay!), then Some SPSS (double yay!)
  - *You **must** do both parts!! No credit if you skip ½ of it!* ☹
  - Seriously – If you skip half the lab, you won't get **ANY** credit.
  - This box is yellow. Please don't ignore something so explicitly highlighted and repeated. THERE ARE TWO PARTS ☺
- **Groups okay**
  - 3 minimum, for Secretary Bonus (5pts)
  - Be sure to specify the secretary *when you submit*

Summary of Preferences

- **If nominal, no options:**
  - *Center:* Can only use mode
  - *Spread:* Can only use variation ratio
- **If ordinal, a bit of option:**
  - *Center:* Use median unless mode is different
  - *Spread:* IQR is best; range *may* be ok
- **If interval, more options:**
  - *Center:* Mean unless skewed, then median
  - *Spread:* Stddev unless outliers, then IQR
- Note: When there are options, choices are contingent

Lab, Part I (just the first HALF of Lab #5):

- **Consider the sample 2, 4, 7, 8, and 9**
  - Those are the values of some variable, for 5 cases
  - Really. Those are the values. That's the sample.
- **Calculate each of the following:**
  - Range
  - Mean (show your work, please)
  - Standard deviation (use a table like Family A/B/C)
  - Variation ratio

One (more) exception...

- **Some comparisons require other choices**
  - **E.g. to compare dispersion of two variables**
  - **One with 3 values and one with 9**
    - **Very different ranges, just in terms of possible values**
    - **Comparing ranges of the actual data, probably isn't telling**
    - **Regardless of how cases are distributed, one will have smaller range**
  - **One with 2 values and one with 5**
    - **Different levels of measurement**
      - **If 2 values, it's definitely nominal and you'd use variation ratio**
      - **But if the variable with 5 is ordinal, you'd probably use the IQR**
    - **You can't compare a variation ratio to an IQR**
  - **Could make a choice you might not otherwise make**
    - **Could use the variation ratio for all four variables**
    - **Not ideal for the variables *on their own*, but**
    - **That choice *might* be the only way to compare two dispersions**

Lab Exercise, Part II (HALF of lab 5):

- **From the same dataset as Lab 4**
- **Pick 1 variable of each LOM**
  - See notes from last lecture on how to pick
  - If you did lab 4 successfully (1 variable of each LOM), then you could use the same variables and even the same output
- **Get & print from SPSS, for all 3:**
  - Frequency distribution
  - Histogram
  - All available measures of *dispersion*
- **Write a few sentences**
  - Report appropriate measure(s) of dispersion for each variable
    - Check plus? Interpret them! Don't remember the difference? See Slide 8!
  - Do *not* report all of them!!