Lecture 02 - Measures of Central Tendency & Dispersion

Data Analysis (CJUS 6103)

Review of Order of Operations

P.E.M.D.A.S.

Parentheses, exponents, multiplication & division, addition & subtraction

1) $4 + 2 * 3 \neq (4 + 2) * 3$
2) $4 + 2^2 \neq (4 + 2)^2$

For multiplication & division (as well as addition & subtraction), just work from left to right.

1) $15/3 * 4 \neq 15/(3 * 4)$

Measures of Central Tendency

Outline

1) Describing a data series
2) What is central tendency?
3) Mode, median, and mean
4) A note on rounding
5) Evaluating skewness
6) Least squares property of the mean

Decribing a Data Series

n is sample size (sometimes, N)

Raw data - $x_1, x_2, x_3, \ldots x_n$

x_i denotes a single observation, where $i = 1, 2, 3, \ldots, n$

Summation operator: $\sum_{i=1}^{n} x_i$ - This tells you to add all the values of x from obs. 1 to obs. n.

What is Central Tendency?

Single number that represents the most common or *typical* score around which others in the distribution tend to cluster.

Three different measures:

1) Mode: The most frequently occurring value
2) Median: Value in the middle of the distribution
3) Mean: Arithmetic average of a distribution

Mode

Qualitative data: Mode is a category

Quantitative data: Mode is a value

Advantages:

1) Simple to compute
2) Used with all types of data (only measure that works with nominal data)

Disadvantages:

1) Fails to take into account data values
2) Not very useful with interval/ratio level data

Median

Value or score that divides a rank-ordered distribution exactly in half

1) 50th percentile

“Positional” measure of central tendency

1) The middle score

Requires ordinal data or higher

Calculating the Median from Raw Data

Steps

1) Arrange the data in ascending order
2) Find the median position (MP) using the formula (n + 1) / 2
3) Count up to identify the value (or midpoint between two values) in this position

Important: The median is the value in the position, not the position itself.

Calculating the Median from a Frequency Distribution

Add a cumulative frequency column to the table and proceed as usual.

1) Compute MP = (n + 1) / 2
2) Identify the value (or midpoint) in this position

Or, use cumulative percent/proportion

1) Identify the first value with a cumulative percent greater than 50%
2) If exactly 50%, take the midpoint of that value and the next one

Mean

Most common measure of central tendency

Requires interval data or higher

Balancing score

1) Positive differences cancel out negative differences
2) Known as the **least squares** property:

$$\sum(x - \overline{x}) = 0$$

Mean

Advantages

- Desirable mathematical properties
  - Least squares, minimum variance
- Stable measure of central tendency
- Utilizes all information provided in the data
  - In other words, it takes into account the values

Disadvantages

- Only useful for quantitative (interval-ratio) data
- Is not always a value that exists in the data
- Sensitive to extreme scores, or “outliers”

Calculating the Mean

- Add up all the value of x
- Divide by the number of observations:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} = \frac{x_1 + x_2 + x_3 + \ldots x_n}{n} = \frac{\sum x}{n}$$

A Note On Rounding

- Round your final answers to at least one additional decimal unit than contained in the original values.
  - # of drinks consumed: 4,2,3,8,6,6
  - Mean = 29 / 6 = 4.8333 = 4.8
- Do not round off your calculations at each intermediate step; wait to round until your final answer if possible
  - Cumulative rounding error can be problematic

A Note on Rounding

- Decide how many decimal places you want to round your answer to
- Look at the digit to the right of the last decimal place that you want to keep
  - If the digits are larger than .5, round up
  - If the digits are smaller than .5, round down
- If the digits are exactly .5, look at the interval place that you want to round
  - If the digit is even, round up ($2.85 \rightarrow 2.9$)
  - If the digit is odd, round down ($6.35 \rightarrow 6.3$)

A Simple Example - Measures of Central Tendency

- Data series ($n = 10$):
  - 1 2 3 4 5 6 7 8 9 10
- Mode = not defined
- Median
  - Median position = $(10 + 1)/2 = 5.5$th position
  - Median = $(5 + 6)/2 = 5.5$
- Mean = $55/10 = 5.5$

Another Simple Example

- Data series ($n = 15$):
  - 1 1 1 2 2 3 4 4 4 4 4 5 5 6 6
- Mode = 4
- Median
  - Median position = $(15 + 1)/2 = 8$th position
  - Median = 4
- Mean = $52/15 = 3.466667 = 3.5$

Frequency Distributions

Mode & Median?

Score(x)	f	p	%	cf	cp	$c\%$
1	3	.06	6%	3	.06	6%
2	4	.08	8%	7	.14	14%
3	5	.10	10%	12	.24	24%
4	10	.20	20%	22	.44	44%
5	7	.14	14%	29	.58	58%
6	6	.12	12%	35	.70	70%
7	6	.12	12%	41	.82	82%
8	5	.10	10%	46	.92	92%
9	3	.06	6%	49	.98	98%
10	1	.02	2%	50	1.00	100%

Frequency Distributions

Mode & Median?

Score(x)	f	p	%	cf	cp	$c\%$
1	2	.02	2%	2	.02	2%
2	5	.05	5%	7	.07	7%
3	9	.09	9%	16	.16	16%
4	11	.11	11%	27	.27	27%
5	13	.13	13%	40	.40	40%
6	10	.10	10%	50	.50	50%
7	12	.12	12%	62	.62	62%
8	8	.08	8%	70	.70	70%
9	5	.05	5%	75	.75	75%
10	25	.25	25%	100	1.00	100%

Computing a Mean from a Frequency Distribution

Weighted Mean formula:

$$\overline{x} = \frac{\sum(w * x)}{\sum(w)}$$

Where w is the **weight** assigned to each observation.

With raw (i.e., **unweighted**) data, each observation is assigned a weight of one, so...

$$\overline{x} = \frac{\sum(w * x)}{\sum w} = \frac{\sum(1 * x)}{\sum 1} = \frac{\sum x}{n}$$

Frequency Distributions

Mean using frequencies as weights:

Score(x)	f	p	%	$f * x$
1	3	.06	6%	3
2	4	.08	8%	8
3	5	.10	10%	15
4	10	.20	20%	40
5	7	.14	14%	35
6	6	.12	12%	36
7	6	.12	12%	42
8	5	.10	10%	40
9	3	.06	6%	27
10	1	.02	2%	10
Total	50	1.00	100%	256

$$\overline{x} = \frac{\sum(f * x)}{\sum f}$$

$$= \frac{256}{50} \qquad (1)$$

$$= 5.1$$

Frequency Distributions

Mean using proportions as weights:

Score(x)	f	p	%	$p * x$
1	3	.06	6%	0.06
2	4	.08	8%	0.16
3	5	.10	10%	0.3
4	10	.20	20%	0.8
5	7	.14	14%	0.7
6	6	.12	12%	0.72
7	6	.12	12%	0.84
8	5	.10	10%	0.8
9	3	.06	6%	0.54
10	1	.02	2%	0.2
Total	50	1.00	100%	5.12

$$\overline{x} = \frac{\sum(p * x)}{\sum p}$$

$$= \frac{5.12}{1.00} \qquad (2)$$

$$= 5.1$$

North Carolina Index Offenses

- Offenses Reported to the Police, 2016
  - What measures of C.T. are appropriate?

Type of Offense	f	p
Murder	678	0.002
Forcible Rape	2849	0.009
Robbery	9336	0.030
Aggravated Assault	24906	0.079
Burglary	72082	0.228
Larceny/Theft	190377	0.603
Motor Vehicle Theft	15306	0.048
	315534	$\tilde{1}.00$

Family Structure

Relationship to parent figure(s):

	White Youth		**Black Youth**	
Family Structure	*f*	*p*	*f*	*p*
Both Bio Parents	2743	.594	607	.263
One Bio/One Step	706	.153	296	.128
Bio Mom only	864	.187	1108	.480
Bio Dad only	172	.037	61	.026
Other Family	136	.029	236	.102
	4621	1.00	2308	.999

Educational Attainment

Educational attainment of police officers:

Educational Attainment	f	p	cf	cp
Less than HS	16	0.048	16	0.048
HS Grad	67	0.199	83	0.247
Some College	117	0.348	200	0.595
College Grad	72	0.214	272	0.809
Post Grad	64	0.190	336	0.999
	336	0.99		

Scholastic Performance

Grades in the 8th grade:

	Males				Females			
Grades	f	p	cf	cp	f	p	cf	cp
Mostly As & Bs	1268	.293	1268	.293	1898	.452	1898	.452
Mostly Bs & Cs	1687	.389	2955	.682	1524	.363	3422	.815
Mostly Cs & Ds	1160	.268	4115	.950	671	.160	4093	.975
Mostly Ds & Fs	217	.050	4332	1.000	109	.026	4202	1.001
	4332	1.000			4202	1.001		

Time Spent on Homework

of weekdays spent doing homework:

	Non-Delinquents				Delinquents			
# Days	f	p	cp	$f * x$	f	p	cp	$f * x$
0	233	.084	.084	0	371	.148	.148	0
1	93	.033	.117	93	127	.051	.199	127
2	210	.075	.192	420	260	.104	.303	520
3	488	.175	.367	1464	526	.210	.513	1578
4	678	.244	.611	2712	498	.198	.711	1992
5	1080	.388	.999	5400	727	.290	1.001	3635
	2782	.999		10089	2509	1.001		7852

$$\overline{x}_{ND} = \frac{\sum(f * x)}{\sum f} = \frac{10089}{2782} = 3.6; \overline{x}_{D} = \frac{\sum(f * x)}{\sum f} = \frac{7852}{2509} = 3.1$$

Juvenile Delinquency

Variety scale of six different delinquent acts youths have committed:

# Delinquent Acts	f	p	cp	$p * x$
0	4597	.515	.515	.000
1	1916	.214	.729	.214
2	1216	.136	.865	.272
3	638	.071	.936	.213
4	279	.031	.967	.124
5	182	.020	.987	.100
6	106	.012	.999	.072
	8934	.999		.995

$$\overline{x} = \frac{\sum(p * x)}{\sum p}$$

$$= \frac{.995}{.999} \qquad (3)$$

$$= 1.0$$

Sentence Length

- Sentence length in months for armed robbery (n=40)
  - 36 38 39 47 50 51 51 53 55 55 ($\sum = 475$)
  - 56 57 60 62 63 64 64 66 67 68 ($\sum = 627$)
  - 69 70 70 70 71 75 78 79 80 80 ($\sum = 742$)
  - 81 83 85 86 87 89 95 98 99 99 ($\sum = 902$)

Mode = 70 (but not very informative)

MP = (40 + 1) / 2 = 20.5 $\rightarrow$ Median = (68 +69) / 2 = 68. 5

Mean = (475+627+742+902) / 40 = 2746 / 40 = 68.7

Unemployment Rate

- State-level unemployment rates, 2016
  - 2.8 2.8 3.0 3.2 3.2 3.3 3.3 3.4 3.7 3.7 ($\sum = 32.4$)
  - 3.8 3.9 3.9 4.0 4.0 4.1 4.1 4.2 4.3 4.4 ($\sum = 40.7$)
  - 4.4 4.5 4.6 4.8 4.8 4.8 4.9 4.9 4.9 4.9 ($\sum = 47.5$)
  - 4.9 5.0 5.0 5.1 5.1 5.3 5.3 5.3 5.4 5.4 ($\sum = 53.8$)
  - 5.4 5.4 5.7 5.8 5.9 6.0 6.0 6.1 6.6 6.7 ($\sum = 59.6$)

Mode = 4.9

MP = (50+1) / 2 = 25.5 $\rightarrow$ Median = (4.8+4.8) / 2 = 4.8

Mean = (32.4+40.7+47.5+53.8+59.6) / 50 = 234 / 50 = 4.68

City Homicide Rates

Washington & Baltimore, 1985 - 1995 ($n = 11$)

		Rank Ordered	
Washington, DC	Baltimore, MD	Wash.	Balt.
23.5	27.6	23.5	27.6
31.0	30.6	31.0	29.5
36.2	29.5	36.2	30.6
59.5	30.6	59.5	30.6
71.9	34.3	65.2	34.3
77.8	41.4	70.0	40.6
80.6	40.6	71.9	41.4
75.2	44.3	75.2	43.4
78.5	48.2	77.8	44.3
70.0	43.4	78.5	45.2
65.2	45.2	80.6	48.2
		669.4	415.7

$$\begin{aligned} \overline{x}_{Wash} &= \frac{669.4}{11} \\ &= 60.85 \\ \overline{x}_{Balt} &= \frac{415.7}{11} \\ &= 37.79 \end{aligned} \tag{4}$$

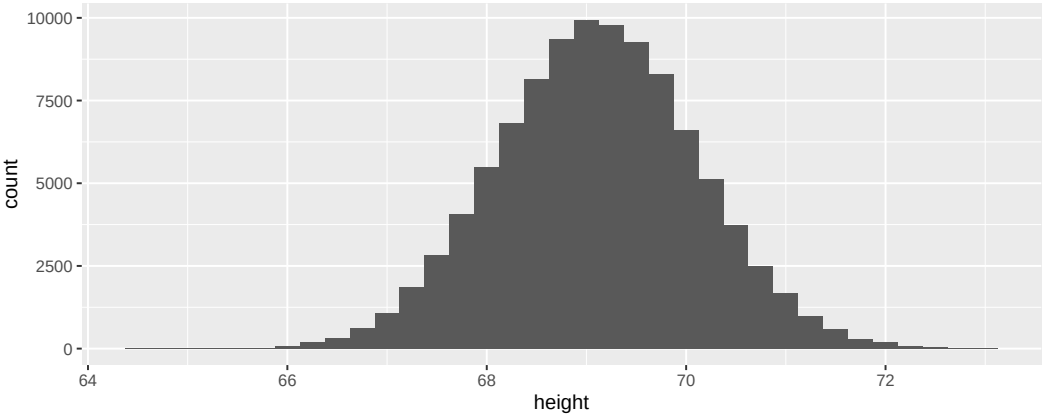
Washington's mean homicide rate for this period was 61.0% higher than Baltimore's:

$$\frac{60.85 - 37.79}{37.79} = .610$$

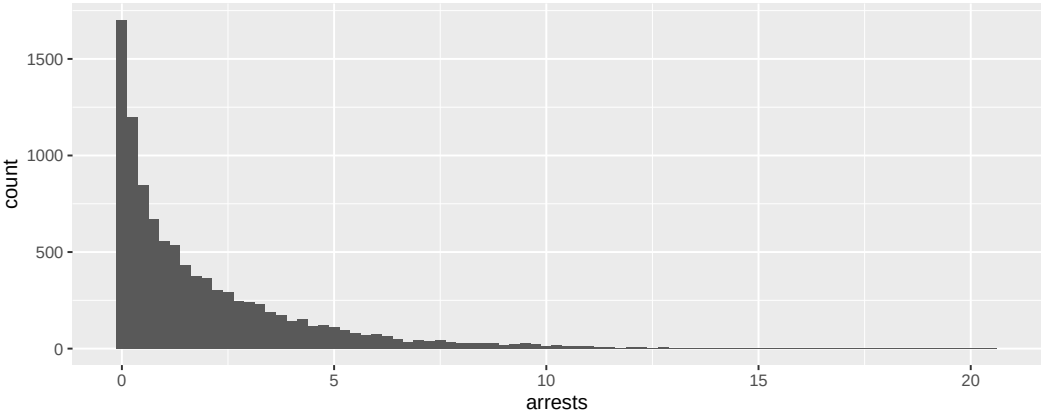
Skewed Data in the Social Sciences

- Examples of skewed variables:
  - Income
  - Number of arrests
  - Crime rate
  - Sentence length
- General rule
  - If mean $>$ median, then positive (right) skew
  - If mean $<$ median, then negative (left) skew

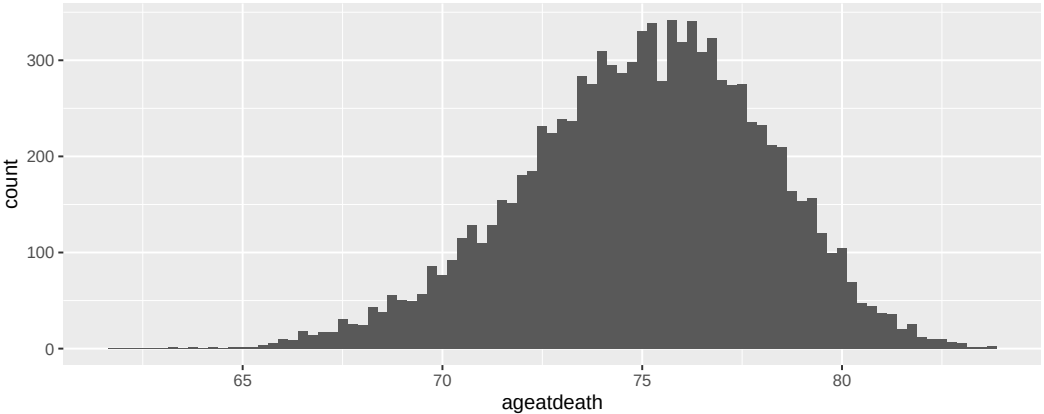
Evaluating Skewness - A Normal Distribution



Evaluating Skewness - Positive (Right) Skew



Evaluating Skewness - Negative (Left) Skew



Desirable Property of the Mean as a Measure of CT

Least squares or minimum variance

I.e., positive deviations from the mean cancel out negative deviations:

$$\sum(x - \overline{x}) = 0$$

Squared deviations are smallest around the mean as compared to any other fixed value:

$$\sum(x - \overline{x})^2 = minimum$$

$$\sum(x - \overline{x})^2 < \sum(x - median)^2$$

A Simple Example of Least Squares

Hypothetical data series ($n = 5$); mode/median=2; mean = $8/5 = 16$

x	$x - median$	$x - \overline{x}$	$(x - median)^2$	$(x - \overline{x})^2$
1	-1.0	-0.6	1.00	0.36
1	-1.0	-0.6	1.00	0.36
2	0.0	0.4	0.00	0.16
2	0.0	0.4	0.00	0.16
2	0.0	0.4	0.00	0.16
	-2.0	0.0	2.00	1.20

Sum of squared deviations from the mean is smaller (1.20) than the sum of squared deviations from the median (2.00)

Measures of Dispersion

Now we will begin to talk about measures of dispersion, which reflect the variation of scores around a central cluster of values.

- 1) Overview of Dispersion
- 2) Dispersion in qualitative (NO) data
- 3) Dispersion in quantitative (IR) data
- 4) Computational formula for variance
- 5) Parameters v. statistics

Overview of Dispersion

- Central tendency: Value around which others tend to cluster
  - Provides a **best guess** of the single value that is most reflective of the data
- Dispersion: How widely scores are scattered about the central score
  - Reflects the degree of uncertainty
  - How good of a *guess* is our **best guess**?

Overview of Dispersion

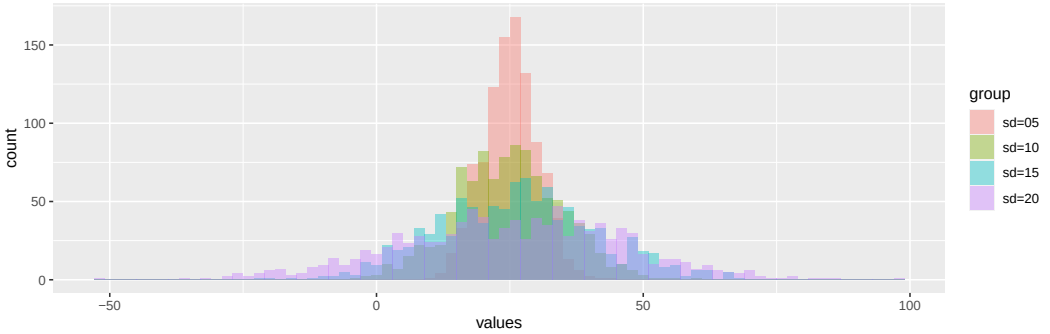
- Central tendency v. dispersion
  - Measures of central tendency are useful as *stand-alone* or **absolute** measures
- Dispersion is easier to interpret when two or more groups are compared
  - Measures of dispersion are only useful as **relative** measures

Importance of Dispersion

- Crossing a river but not a good swimmer
  - You know that the mean depth is 3 feet
- Do you cross?
  - Mean says nothing about the depth of the river **at any particular point**
- Measurement of depth at five-foot intervals
  - Scenario #1: 3 3 3 3 3 3 3 3 3 3
  - Scenario #2: 1 2 2 3 3 3 4 4 4 4
  - Scenario #3: 1 1 1 1 2 2 2 2 9 9
  - Scenario #4: 1 1 1 1 1 1 1 1 1 21

Comparing distributions

The distribution of a variable with the same mean can look very different depending on the degree of variability



Measures of Dispersion

- Qualitative data (nominal & ordinal)
  - Variation ratio (VR)
- Quantitative data
  - Range
  - Interquartile range (IQR)
  - Mean deviation
  - Variance (s^2) and standard deviation (s)

Variation Ratio

Proportion of cases that lie *outside* the **modal** category.

Counterpart to the mode; bound between 0.0 and 1.0

Can multiply by 100 to turn into a percentage

$$VR = 1 - \left(\frac{f_{mode}}{n} \right) = 1 - p_{mode}$$

Note - not defined with bimodal distributions

Variation Ratio

Relationship to parent figure(s):

Family Structure	f	p
Both Bio	3350	.483
One Bio/One Step	1002	.145
Bio Mom only	1972	.285
Bio Dad only	233	.034
Other Family	372	.054
	6929	1.001

$$VR = 1 - .483 = .517$$

51.7% of cases are *outside* of the **modal** category.

Variation Ratio

Relationship to parent figure(s) by race:

	White Youth		**Black Youth**	
Family Structure	f	p	f	p
Both Bio Parents	2743	.594	607	.263
One Bio/One Step	706	.153	296	.128
Bio Mom only	864	.187	1108	.480
Bio Dad only	172	.037	61	.026
Other Family	136	.029	236	.102
	4621	1.00	2308	.999

Which group has more variability?

$VR_{Whites} = 1 - .594 = 0.406$; $VR_{Blacks} = 1 - .480 = 0.52$

Whites are more homogeneous than blacks with respect to family structure.

Interquartile Range

- The range of the middle 50% of data
  - Counterpart to median as measure of C.T.
  - 25th to 75th percentile
  - Less sensitive to outliers than range (x_{max}˘x_{min})
- Calculating the IQR
  - Arrange the data in ascending order
  - Compute the MP & truncate it to get TMP
  - Find the quartile position $QP = (TMP + 1)/2$
  - Count up from the lowest & down from the highest

$$IQR = Q_3 - Q_1$$

Mean Deviation

Average deviation of scores about the mean

- Calculating the mean deviation
  - Calculate the sample mean
  - Subtract mean from each score (the **deviation**)
  - Sum of the **absolute** deviations and divide by **sample size**

$$MD = \frac{\sum |x - \overline{x}|}{n}$$

Variance and Standard Deviation

Variance (s^2) = Average **squared** deviation of scores about the mean (s = standard deviation).

Counterpart to the mean as a measure of CT

Calculating the variance:

1) Calculate the sample mean
2) Subtract the mean from each score
3) Square the deviation score for each case
4) Sum up squared deviations and divide by n

$$s^2 = \frac{\sum(x - \overline{x})^2}{n}; \; s = \sqrt{\frac{\sum(x - \overline{x})^2}{n}}$$

More on the Variance & Standard Deviation

- Relies on the **least squares** property of the mean
  - Variance is at a minimum when taking squared deviations from the mean v. any other value
- Squaring deviations makes other statistical analyses easier
  - But, this changes the unit of measurement, giving it a cumbersome interpretation
  - We report standard deviation (s) more often

Note on Mean Deviation & Variance

Mean deviation - the average **absolute** deviation from the mean:

$$MD = \frac{\sum |x - \overline{x}|}{n}$$

Variance - the average **squared** deviation:

$$s^2 = \frac{\sum (x - \overline{x})^2}{n}$$

Mean deviation applies equal weight to deviations, whereas the variance weights large deviations more.

River Crossing Example Revisited

Scenario# 1		**Scenario# 2**		**Scenario# 3**		**Scenario# 4**	
x	$(x-\overline{x})^2$	x	$(x-\overline{x})^2$	x	$(x-\overline{x})^2$	x	$(x-\overline{x})^2$
3	0	1	4	1	4	1	4
3	0	2	1	1	4	1	4
3	0	2	1	1	4	1	4
3	0	3	0	1	4	1	4
3	0	3	0	2	1	1	4
3	0	3	0	2	1	1	4
3	0	4	1	2	1	1	4
3	0	4	1	2	1	1	4
3	0	4	1	9	36	1	4
3	0	4	1	9	36	21	324
	0		10		92		360
$VR = 0$		$VR = 0.6$		$VR = undefined$		$VR = 0.1$	
$IQR = 3 - 3 = 0$		$IQR = 4 - 2 = 2$		$IQR = 2 - 1 = 1$		$IQR = 1 - 1 = 0$	
$s = \sqrt{\frac{0}{10}} = 0$		$s = \sqrt{\frac{10}{10}} = 1.0$		$s = \sqrt{\frac{92}{10}} = 3.0$		$s = \sqrt{\frac{360}{10}} = 6.0$	

Frequency Distributions

Calculate variation ratio and interquartile range:

Score(x)	f	p	cf	rcf	cp
1	3	.06	3	50	.06
2	4	.08	7	47	.14
3	5	.10	12	43	.24
4	10	.20	22	38	.44
5	7	.14	29	28	.58
6	6	.12	35	21	.70
7	6	.12	41	15	.82
8	5	.10	46	9	.92
9	3	.06	49	4	.98
10	1	.02	50	1	1.00
	50	1.00			

Frequency Distributions

Calculate variation ratio and interquartile range:

Score(x)	f	p	cf	rcf	cp
1	3	.06	3	50	.06
2	4	.08	7	47	.14
3	5	.10	12	43	.24
4	10	.20	22	38	.44
5	7	.14	29	28	.58
6	6	.12	35	21	.70
7	6	.12	41	15	.82
8	5	.10	46	9	.92
9	3	.06	49	4	.98
10	1	.02	50	1	1.00
	50	1.00			

$$VR = 1 - 0.20 = 0.80$$

Frequency Distributions

Calculate variation ratio and interquartile range:

Score(x)	f	p	cf	rcf	cp
1	3	.06	3	50	.06
2	4	.08	7	47	.14
3	5	.10	12	43	.24
4	10	.20	22	38	.44
5	7	.14	29	28	.58
6	6	.12	35	21	.70
7	6	.12	41	15	.82
8	5	.10	46	9	.92
9	3	.06	49	4	.98
10	1	.02	50	1	1.00
	50	1.00			

$VR = 1 - 0.20 = 0.80$

IQR using frequencies:
1) $MP = (50 + 1)/2 = 25.5$
2) $QP = (25 + 1)/2 = 13$
3) $IQR = 7 - 4 = 3$

Frequency Distributions

Calculate variation ratio and interquartile range:

Score(x)	f	p	cf	rcf	cp
1	3	.06	3	50	.06
2	4	.08	7	47	.14
3	5	.10	12	43	.24
4	10	.20	22	38	.44
5	7	.14	29	28	.58
6	6	.12	35	21	.70
7	6	.12	41	15	.82
8	5	.10	46	9	.92
9	3	.06	49	4	.98
10	1	.02	50	1	1.00
	50	1.00			

$VR = 1 - 0.20 = 0.80$

IQR using frequencies:
1) $MP = (50 + 1)/2 = 25.5$
2) $QP = (25 + 1)/2 = 13$
3) $IQR = 7 - 4 = 3$

IQR using proportions:
1) $IQR = 7 - 4 = 3$

Frequency Distributions

How can we calculate variance & standard deviation with frequency distributions?

Must modify formulas to use frequencies (or proportions/percents) as weights.

$$s^2 = \frac{\sum f * (x - \overline{x})^2}{\sum f}; \text{ or } s^2 = \sum p * (x - \overline{x})^2$$

$$s = \sqrt{\frac{\sum f * (x - \overline{x})^2}{\sum f}}; \text{ or } s = \sqrt{\sum p * (x - \overline{x})^2}$$

Frequency Distributions

Calculate s using f ($\overline{x} = 5.1$):

Score (x)	f	$x - \overline{x}$	$(x - \overline{x})^2$	$f * (x - \overline{x})^2$
1	3	$1 - 5.1 = -4.1$	16.81	50.43
2	4	$2 - 5.1 = -3.1$	9.61	38.44
3	5	$3 - 5.1 = -2.1$	4.41	22.05
4	10	$4 - 5.1 = -1.1$	1.21	12.1
5	7	$5 - 5.1 = -0.1$	0.01	0.07
6	6	$6 - 5.1 = 0.9$	0.81	4.86
7	6	$7 - 5.1 = 1.9$	3.61	21.66
8	5	$8 - 5.1 = 2.9$	8.41	42.05
9	3	$9 - 5.1 = 3.9$	15.21	45.63
10	1	$10 - 5.1 = 4.9$	24.01	24.01
	50			261.30

Frequency Distributions

Calculate s using f ($\overline{x} = 5.1$):

Score (x)	f	$x - \overline{x}$	$(x - \overline{x})^2$	$f * (x - \overline{x})^2$
1	3	$1 - 5.1 = -4.1$	16.81	50.43
2	4	$2 - 5.1 = -3.1$	9.61	38.44
3	5	$3 - 5.1 = -2.1$	4.41	22.05
4	10	$4 - 5.1 = -1.1$	1.21	12.1
5	7	$5 - 5.1 = -0.1$	0.01	0.07
6	6	$6 - 5.1 = 0.9$	0.81	4.86
7	6	$7 - 5.1 = 1.9$	3.61	21.66
8	5	$8 - 5.1 = 2.9$	8.41	42.05
9	3	$9 - 5.1 = 3.9$	15.21	45.63
10	1	$10 - 5.1 = 4.9$	24.01	24.01
	50			261.30

$$s = \sqrt{\frac{\sum f * (x - \overline{x})^2}{\sum f}}$$

$$= \sqrt{\frac{261.30}{50}} \quad (5)$$

$$= \sqrt{5.226}$$

$$= 2.29$$

Frequency Distributions

Calculate s using p ($\overline{x} = 5.1$):

Score (x)	p	$x - \overline{x}$	$(x - \overline{x})^2$	$p * (x - \overline{x})^2$
1	.06	$1 - 5.1 = -4.1$	16.81	1.0086
2	.08	$2 - 5.1 = -3.1$	9.61	0.7688
3	.10	$3 - 5.1 = -2.1$	4.41	0.441
4	.20	$4 - 5.1 = -1.1$	1.21	0.242
5	.14	$5 - 5.1 = -0.1$	0.01	0.0014
6	.12	$6 - 5.1 = 0.9$	0.81	0.0972
7	.12	$7 - 5.1 = 1.9$	3.61	0.4332
8	.10	$8 - 5.1 = 2.9$	8.41	0.841
9	.06	$9 - 5.1 = 3.9$	15.21	0.9126
10	.02	$10 - 5.1 = 4.9$	24.01	0.4802
	1.00			5.226

Frequency Distributions

Calculate s using p ($\overline{x} = 5.1$):

Score (x)	p	$x - \overline{x}$	$(x - \overline{x})^2$	$p * (x - \overline{x})^2$
1	.06	$1 - 5.1 = -4.1$	16.81	1.0086
2	.08	$2 - 5.1 = -3.1$	9.61	0.7688
3	.10	$3 - 5.1 = -2.1$	4.41	0.441
4	.20	$4 - 5.1 = -1.1$	1.21	0.242
5	.14	$5 - 5.1 = -0.1$	0.01	0.0014
6	.12	$6 - 5.1 = 0.9$	0.81	0.0972
7	.12	$7 - 5.1 = 1.9$	3.61	0.4332
8	.10	$8 - 5.1 = 2.9$	8.41	0.841
9	.06	$9 - 5.1 = 3.9$	15.21	0.9126
10	.02	$10 - 5.1 = 4.9$	24.01	0.4802
	1.00			5.226

$$
\begin{aligned}
s &= \sqrt{\frac{\sum p * (x - \overline{x})^2}{\sum p}} \\
&= \sqrt{\frac{5.226}{1}} \qquad (6) \\
&= \sqrt{5.226} \\
&= 2.29
\end{aligned}
$$

Computational Formula for Variance & Standard Deviation

With a small # of values, the **definitional** formula is fine.

With a large # of values, the **computational**, or shortcut formula, is better.

Requires less information (only x & x^2):

$$s^2 = \frac{\sum(x^2) - (\sum x)^2}{n} = \frac{\sum(x^2)}{n} - \overline{x}^2; \text{ where } \overline{x} = \frac{\sum x}{n}$$

$$s^2 = \frac{\sum(w * x^2) - (\sum w * x)^2}{\sum w} = \frac{\sum(w * x^2)}{\sum w} - \overline{x}^2; \text{ where } \overline{x} = \frac{\sum w * x}{\sum w}$$

Note: *This is helpful when doing calculations by hand, less so when having R run them for you (and generally off by a small amount).*

Sentence Length (Again)

- Sentence length in months for armed robbery (n=40)
  - 36 38 39 47 50 51 51 53 55 55
  - 56 57 60 62 63 64 64 66 67 68
  - 69 70 70 70 71 75 78 79 80 80
  - 81 83 85 86 87 89 95 98 99 99

Mode = 70

Median = 68.5

Mean = 68.7

VR = 1-(3/40) = 1-0.75 = .925

QP = (20+1)/2 = 10.5 $\rightarrow$ IQR = 80.5-55.5 = 25.0

Sentence Length

For s, first **square** all raw values:

- Sentence length in months for armed robbery (n=40)
  - 1296 1444 1521 2209 2500 2601 2601 2809 3025 3025
  - 3136 3249 3600 3844 3969 4096 4096 4356 4489 4624
  - 4761 4900 4900 4900 5041 5625 6084 6241 6400 6400
  - 6561 6889 7225 7396 7569 7921 9025 9604 9801 9801
  - Sum = 199,534

Then, plug the relevant numbers into the formulas:

$$s^2 = \frac{\sum(x^2)}{n} - \overline{x}^2 = \frac{199534}{40} - 68.7^2 = 268.66$$

$$s = \sqrt{268.66} = 16.39$$

City Homicide Rate

Washington, DC	Wash^2	Baltimore, MD	Balt^2
23.5	552.25	27.6	761.76
31.0	961	30.6	936.36
36.2	1310.44	29.5	870.25
59.5	3540.25	30.6	936.36
71.9	5169.61	34.3	1176.49
77.8	6052.84	41.4	1713.96
80.6	6496.36	40.6	1648.36
75.2	5655.04	44.3	1962.49
78.5	6162.25	48.2	2323.24
70.0	4900	43.4	1883.56
65.2	4251.04	45.2	2043.04
	45051.08		16255.87

$$s_{Wash} = \sqrt{\frac{45051.08}{11} - 60.85^2}$$

$$= \sqrt{392.830} = 19.82$$

$$s_{Balt} = \sqrt{\frac{16255.87}{11} - 37.79^2}$$

$$= \sqrt{49.722} = 7.05 \tag{7}$$

Reviews of Formulas for Standard Deviation

Unweighted data (i.e. raw numbers):

Definitional formula: $s = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$

Computational formula: $s = \sqrt{\frac{\sum(x^2)}{n} - \overline{x}^2}$

Weighted data (i.e., frequency distribution):

Definitional formula: $s = \sqrt{\frac{\sum w*(x-\overline{x})^2}{\sum w}}$

Computational formula: $s = \sqrt{\frac{\sum(w*x^2)}{\sum w} - \overline{x}^2}$

Populations v. Sample Means & Standard Deviations

- Population mean & standard deviation
  - μ = mean; σ = standard deviation
  - $\mu = \frac{\sum x}{N}$; $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$
  - Where N represents the size of the population
  - Known as population **parameters**
- Sample mean & standard deviation
  - $\overline{x}$ = mean; s = standard deviation
  - $\overline{x} = \frac{\sum x}{n}$; $s = \sqrt{\frac{\sum (x-\overline{x})^2}{n}}$
  - Where n represents the size of the sample
  - Known as sample **statistics**

Populations v. Sample Means & Standard Deviations

- For inferential purposes, we want to use sample **statistics** as estimates of population **parameters**.
  - Can we use $\overline{x}$ as a valid estimate of μ?
  - Can we use s as a valid estimate of σ?
- Fortunately, $\overline{x}$ provides an **unbiased** estimate of μ:

$$\overline{x} = \hat{\mu}$$

  - We say *mu hat*, where the hat signifies an estimate of the true quantity of interest.

Populations v. Sample Means & Standard Deviations

- However, for reasons we will discuss later in the semester s is a biased estimate σ:

$$s \neq \hat{\sigma}$$

  - We call the term on the right-hand side of the equality **sigma hat**.
  - An unbiased estimate of σ substitutes $n - 1$ in the denominator rather than n.

$$s = \sqrt{\frac{\sum(x - \overline{x})^2}{n}} \qquad \hat{\sigma} = \sqrt{\frac{\sum(x - \overline{x})^2}{n - 1}}$$

R Tutorial - Sentence Length Example

All of this is made to seem pretty easy and it is if you're doing manual hand calculations (apart from human error - i.e., writing down the wrong values). When it comes to entering these data into R it's likely a bit more intimidating at first - but it's also pretty simple.

The following slides will break down how to add the sentence length data from earlier in this lecture into R and then compute measures of central tendency and dispersion for that sample.

For each measure I will show you manual and automatic computation methods. You won't be permitted to use the automatic methods for a while, but it's helpful to know them so you can check your answers.

R Tutorial - Sentence Length Example

Here's the list of values again, for reference:

- Sentence length in months for armed robbery (n=40)
  - 36 38 39 47 50 51 51 53 55 55
  - 56 57 60 62 63 64 64 66 67 68
  - 69 70 70 70 71 75 78 79 80 80
  - 81 83 85 86 87 89 95 98 99 99

R Tutorial - Sentence Length Example

Step 1. Manually enter the data into R.

This can be accomplished using the c() function, as so:

```
sent_length<-c(36,38,39,47,50,51,51,53,55,55,56,57,60,62,63,64,64,66,
               67,68,69,70,70,70,71,75,78,79,80,80,81,83,85,86,87,89,
               95,98,99,99)
length(sent_length)
```

```
## [1] 40
```

I add the length() function at the end to make sure I entered 40 separate values. Next we will want to compute the mean.

R Tutorial - Sentence Length Example

Step 2. Compute the mean.

```
## Manual Computation
xbar1<-(sum(sent_length)/length(sent_length))
## Sum() function adds together all values in the vector
xbar1
```

```
## [1] 68.65
```

```
## Automatic Computation
xbar2<-mean(sent_length)
xbar2
```

```
## [1] 68.65
```

I want you to use the manual method for this class until you are told otherwise. Using the mean() function is kind of cheating until you properly understand what it is doing. You'll hopefully have noticed that I stored the value of the mean in a separate object - that will be helpful for a later step.

R Tutorial - Sentence Length Example

Step 3. Identify the median value.

Note: I first have to randomize the order for the purpose of the tutorial.

```
sent_length<-sample(sent_length,40,replace=FALSE)
sent_length ## No longer in order!
```

```
##  [1] 95 67 66 64 55 75 70 87 38 62 83 51 70 98 47 71 64 78 89 86 51 36 79 69 39
## [26] 63 99 55 57 53 80 80 81 70 50 60 85 56 68 99
```

R Tutorial - Sentence Length Example

Step 3. Identify the median value.

Now, onto regularly scheduled programming...

```
## Manual Computation
sent_length<-sort(sent_length) ## Have to re-sort or I will get the wrong value
sent_length
```

```
##  [1] 36 38 39 47 50 51 51 53 55 55 56 57 60 62 63 64 64 66 67 68 69 70 70 70 71
## [26] 75 78 79 80 80 81 83 85 86 87 89 95 98 99 99
```

```
(length(sent_length)+1)/2 ## Obtain the median position
```

```
## [1] 20.5
```

```
median1<-(sent_length[20]+sent_length[21])/2
median1
```

```
## [1] 68.5
```

R Tutorial - Sentence Length Example

Step 3. Identify the median value.

```
## Automatic Computation
median2<-median(sent_length)
median2
```

```
## [1] 68.5
```

Again, I want you to use the manual method until I tell you otherwise (mostly so I can see your work and know you understand how to calculate these statistics), but it's helpful to use the automatic computation to check your math at first.

R Tutorial - Sentence Length Example

Step 4. Identify the mode.

Perhaps quite astoundingly, there is not a built-in function to compute the mode in R. Instead, you can do so by examining a frequency table of values for a variable (or build a function to compute the mode - there's a lot of example code out there to do this and it's pretty straightforward).

```
table(sent_length)
```

```
## sent_length
## 36 38 39 47 50 51 53 55 56 57 60 62 63 64 66 67 68 69 70 71 75 78 79 80 81 83
##  1  1  1  1  1  2  1  2  1  1  1  1  1  2  1  1  1  1  3  1  1  1  1  2  1  1
## 85 86 87 89 95 98 99
##  1  1  1  1  1  1  2
```

R Tutorial - Sentence Length Example

Step 4. Identify the mode.

An example of building a function to get the mode:

```
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(sent_length)
```

```
## [1] 70
```

R Tutorial - Sentence Length Example

Step 5. Compute the range.

Computing the range is very straightforward and you can use two built-in functions to help you do so.

```
max(sent_length)-min(sent_length)
```

```
## [1] 63
```

And there you have it - the difference between the maximum and minimum sentence lengths is 63 months.

R Tutorial - Sentence Length Example

Step 5. Compute the interquartile range.

I'll show you two ways to do this, as well.

```
(trunc((length(sent_length)+1)/2)+1)/2
```

```
## [1] 10.5
```

The trunc() function truncates the decimal value - i.e., it replaces a positive decimal value with 0. I then compute the quartile position by dividing the truncated median position (+1) in half. I get 10.5, so that means I count up and down 10.5 values to the the 25th and 75th percentile values, respectively. Or, I can manually compute those values.

R Tutorial - Sentence Length Example

Step 5. Compute the interquartile range.

```
## Manual Computation
## 25th Percentile Value
q1<-(sent_length[10]+sent_length[11])/2
## 75th percentile Value
q3<-(sent_length[30]+sent_length[29])/2
## IQR Value
q3-q1
```

```
## [1] 24.5
```

Now we know that the difference between the values defining the middle 50% of the data is 24.5 months.

In the next slide I will show you how to compute this value automatically to check your math.

R Tutorial - Sentence Length Example

Step 5. Compute the interquartile range.

```
## Automatic Computation
IQR(sent_length)
```

```
## [1] 24.5
```

Whelp, that was a lot less verbose - but again, you have to estimate these statistics manually until I tell you otherwise. The automatic functions are just to help you check your math!

NOTE: The way this function calculates the IQR is different from the manual method. In certain scenarios the answer from this function may be slightly different than the one you get using the manual calculation.

R Tutorial - Sentence Length Example

Step 6. Compute the mean deviation.

This one is a bit more complex and is partway toward calculating the standard deviation.

```
## Manual Computation
sent_dev<-(sent_length-xbar1) ## Individual deviations from means.
sent_dev_abs<-abs(sent_dev) ## Stores absolute values of deviations
mad_sentlength<-(sum(sent_dev_abs)/length(sent_dev_abs))
mad_sentlength
```

```
## [1] 13.55
```

The abs() function calculates absolute values of the sent_dev vector.

R Tutorial - Sentence Length Example

Step 6. Compute the mean deviation.

I installed a package in the setup code chunk that has a function to automatically compute the mean absolute deviation - MeanAD(). I use it to check my math above.

```
MeanAD(sent_length)
```

```
## [1] 13.55
```

Note: the MeanAD() function is a part of the DescTools package. The package must be installed and loaded before you can use it.

R Tutorial - Sentence Length Example

Step 7. Compute the standard deviation.

This one is the most complicated - it can be done in a single step, or multiple. I suggest the latter when you first start out then use a single step only when you are more comfortable using R. Here goes...

```
sent_dev_sq<-sent_dev^2 ## Square the individual deviations
sum_dev_sq<-sum(sent_dev_sq) ## Sum the squared deviations
variance<-sum_dev_sq/(length(sent_length)-1) ## Compute variance with sample correction
variance ## Display variance value
```

```
## [1] 282.5923
```

```
standard_deviation<-sqrt(variance) ## Compute the standard deviation
standard_deviation ## Display standard deviation value
```

```
## [1] 16.81048
```

R Tutorial - Sentence Length Example

Step 7. Compute the standard deviation.

Now to do it automatically...

```
var(sent_length) ## Compute and display variance value
```

```
## [1] 282.5923
```

```
sd(sent_length) ## Compute and display standard deviation value
```

```
## [1] 16.81048
```

Note - the var() and sd() functions in R automatically estimate these values with the sample correction (n-1). For our purposes we can assume that we have population data at first, then use these corrections when I tell you to do so.

R Tutorial - Sentence Length Example

Step 8. Now for the real kick in the pants. . . .

The summarize_all function can provide you with almost all of this information OR it provides the intermediate calculations you need to easily compute the final values. **But** you do not get to use it! You have to walk before you can run, and computing these statistics manually is an important step towards understanding what these values actually mean.

R Tutorial - Sentence Length Example

Step 8. Now for the real kick in the pants.... (on this slide)

```
summarize_all(data.frame(sent_length), list(mean=mean, median=median,
                                            sd=sd, IQR=IQR, var=var,
                                            range=range, MeanAD=MeanAD))
```

```
##    mean median       sd  IQR      var range MeanAD
## 1 68.65   68.5 16.81048 24.5 282.5923    36  13.55
## 2 68.65   68.5 16.81048 24.5 282.5923    99  13.55
```

R Tutorial - Sentence Length Example

Step 8. Now for the real kick in the pants.... (an explanation of the function)

summarize_all() belongs to the dplyr package (in tidyverse). It allows you to easily create a table containing all of the descriptive statistics you want (defined by the list() option within the function). The "mean=" part names the column and the "mean" part tells R which function to use to calculate that descriptive statistic. I wrap the sent_length variable within a data.frame function because the summarize_all() function does not work with vectors outside of data frames (to my knowledge!).

Note - the range returns the maximum and minimum values, you still need to do the last step for that one.

The End

Time for your Two Questions!