Lecture 02 Part 01 - Measures of Central Tendency

Data Analysis in CJ (CJUS 6103)

Measures of Central Tendency

Outline

I. Describing a data series
II. What is central tendency?
III. Mode, median, and mean
IV. Evaluating skewness
V. "Least Squares" Property of the Mean

Describing a data series

Before we jump into our discussion of measures of central tendency, it is important to become familiar with notation that I will use throughout the course.

n is sample size (sometimes, N)

Raw data: $x_1, x_2, x_3, ...x_n$

x_i denotes a single observation, where i = 1, 2, 3,. . . ,n

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

What is central tendency?

Measures of central tendency provide a single number that represents the most common or "typical" score around which others in the distribution tend to cluster. They describe the center point in a distribution of scores. Researchers usually have a lot of raw data at their disposal, in some cases thousands or tens of thousands of observations. All of this data is numerical, so a researcher is literally looking at a huge matrix of numbers. For obvious reasons, it is a little cumbersome to refer to an entire data matrix when trying to describe it.

- Three types of central tendency measures are used:
  - Mode – most frequently occurring value.
  - Median – value in the middle of the distribution.
  - Mean – arithmetic average of a distribution.

Mode

The mode is that value or category that occurs with the highest frequency (or highest probability). It is important to keep in mind that the mode is not the highest frequency, but is the category or value associated with the highest frequency.

With qualitative data, the mode is a category; with quantitative data, the mode is a value.

Advantages: 1) Simple to compute, and requires no calculations. The analyst simply searches the data series for the most common value. 2) It can be used with all types of data, qualitative or quantitative, and is the only measure of central tendency that can be used with nominal data.

Disadvantages: 1) Fails to take into account all of the data. It ignores all categories other than the most frequent, and potentially gives a misleading picture of central tendency. 2) It is not very useful with data other than at the nominal and ordinal levels and is seldom used in practice. This is particularly true with quantitative data where the distribution may be relatively flat.

Median

The median is the value that divides a distribution of scores exactly in half. It is the middle value, or the 50th percentile. It is a "positional" measure of central tendency that can be used with data at the ordinal level or higher. Other popular positional measures are the deciles and quartiles.

Advantages: 1) Uses more information than the mode. The median takes into account the position of all of the data values. 2) Not sensitive to extreme scores, thus it is the preferred measure of central tendency with skewed data. Inclusion of outliers does not change values in the middle of the distribution. 3) Can be used with data at the ordinal level or higher. Use of the median only requires that data be rank ordered.

Disadvantages: 1) Still does not use all of the data available. You are only interested in the value that corresponds to a certain position (the 50th percentile), but you do not take into account the actual values themselves. Thus, the fact that the median is not sensitive to the values in a distribution is both an advantage and disadvantage. 2) Most frequently used statistical techniques rely on the sample mean as a measure of central tendency.

Calculating the median: 1) Arrange the data in ascending order. 2) Find the position of the median using the formula [(n + 1) / 2]. 3) Count up from the first observation to identify the value (or midpoint of two values) in this position. It is important to remember that the median is the value in the position, not the position itself.

Mean

The mean is the most common measure of central tendency. It represents the arithmetic average of a distribution of scores, which requires data at the ordinal level or higher. The mean is a "balancing" score, meaning that the positive and negative differences from the mean ( ) cancel, or balance, each other out. This important property means that .

Advantages: 1) It has desirable mathematical properties. Among these is the property of "least squares." 2) It is a stable measure of central tendency. 3) Utilizes all data in the distribution. In other words, the mean uses the values of all of the data points.

Disadvantages: 1) It is only useful with quantitative (interval, ratio) data. 2) Because it uses all of the data, the mean is sensitive to extreme scores, or outliers. This is especially true when the sample size is small.

To calculate the mean, you must simply add up all of the values of x and then divide by the number of observations:

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

Examples of Mode, Median, and Mean

Let's consider a couple of examples of central tendency with simple hypothetical data series.

Data series with n = 15:

1 1 1 2 2 3 4 4 4 4 4 5 5 6 6

The mode in this example is 4, as it has the highest frequency in the data (f = 5). The median is computed by first finding the median position, MP = (15 + 1) / 2 = 8. The value in the 8th position is 4. Note that I could have counted up from the lowest value or down from the highest value and arrived at the same answer for the median. Finally, the mean is calculated as 52 / 15 = 3.47.

Data series with n = 20:

1 1 2 2 3 3 3 3 4 4
5 6 7 7 8 8 8 8 9 9

In this example, there are actually two modes: 3 and 8, both with f = 4. We would refer to these data as bimodal. The median position is MP = (20 + 1) / 2 = 10.5. Since the value in the 10.5th position is not an integer, we must take the midpoint of the values in the 10th and 11th positions: (4 + 5) / 2 = 4.5. The mean is calculated as 101 / 20 = 5.05.

When data are in the form of a frequency distribution, we can still compute each measure of central tendency. Consider the following data (n = 50).

Score(x)	f	p	cp	$f * x$
1	3	.06	.06	3
2	4	.08	.14	8
3	5	.10	.24	15
4	10	.20	.44	40
5	7	.14	.58	35
6	6	.12	.70	36
7	6	.12	.82	42
8	5	.10	.92	40
9	3	.06	.98	27
10	1	.02	1.00	10
Total	50	1.00	100%	256

In this form, the mode is easier to determine. All we must do is find the value or values that have the highest frequency, proportion, or percent. In this case, the mode is 4. The median is also very easy to find with data in a frequency distribution. We do not need to compute the median position first. Rather we simply add an additional column by including the cumulative proportion. To find the median, we simply follow the column down until we encounter the first value that exceeds .500 (or 50%). If a value or category falls exactly at 0.50, this is the median. In our example, the median is 5. Finally, since we observe the raw x-values and their frequencies, we can easily modify the mean formula to reflect this:

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

This formula tells us that we must multiply each ¬x¬-value by its respective frequency, and then sum up the resulting values. This means that we need to add one additional column of data that is the product of the score and its frequency. When we sum up the values in this column, the mean is 256 / 50 = 5.12.

Now let's consider measures of central tendency with real data.

The following table provides a distribution of Index offenses for the State of North Carolina in 2016. These are offenses that were reported to police.

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	$\tilde{1}.00$

What measures of C.T. are appropriate?

Just the mode. Because offense type is a nominal variable, this limits the measures of central tendency that are appropriate to describe these data.

The following table provides information on family living arrangements among a nationally representative sample of youths ages 12 to 16.

	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

What measures of C.T. are appropriate?

Again, just the mode. Because family structure is a nominal variable, this limits the measures of central tendency that are appropriate to describe these data.

The following table provides educational attainment among a sample 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		

What measures of C.T. are appropriate?

Both the mode and the median are appropriate. Because educational attainment is an ordinal variable, this also allows us to make statements like "the median educational attainment among police officers is some college, approximately half of officers have less than this level of education and the other half have more than this level of education (or equal to it)."

The following table provides grades in the 8th grade for male and female youths.

	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		

What measures of C.T. are appropriate?

Both the mode and the median are appropriate. Because grades in the 8th grade is an ordinal variable, this also allows us to make statements like "the median grades among female youth is Mostly Bs & Cs, approximately half of female youth have lower grades and the other half have better (or equivalent) grades."

This table compares the number of weekdays spent doing homework among self-reported delinquents and non-delinquents.

	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

Since the data are measured at the ratio level, we can compute a mean. The means are 10,089 / 2,782 = 3.63 and 7,852 / 2,509 = 3.13 for non-delinquents and delinquents, respectively.

The following table provides a variety scale of the number of six different delinquent acts that a youth reports having ever committed. The mean is computed as 8,924 / 8,934 = 0.999.

# 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

Since the data are measured at the ratio level, we can compute a mean. The mean variety of crimes is computed by dividing the sum of the $p * x$ column by the sum of the p column: .995 / .999 = .996.

The following data represent sentence length in months for persons convicted of armed robbery (n = 40).

- 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 → Median = (68 +69) / 2 = 68. 5

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

The following data represent state-level unemployment for 2016 (n = 50).

- 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 → Median = (4.8+4.8) / 2 = 4.8

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

The following data provide homicide rates per 100,000 from 1985 to 1995 for Washington, DC and Baltimore, MD (n = 11, or n=22, depending upon the analysis).

		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

The median position is 12 / 2 = 6, and when rank ordered, the medians are 70.0 and 40.6 for Washington and Baltimore, respectively. The respective means are 669.4 / 11 = 60.85 and 415.7 / 11 = 37.79.

Evaluating Skewness

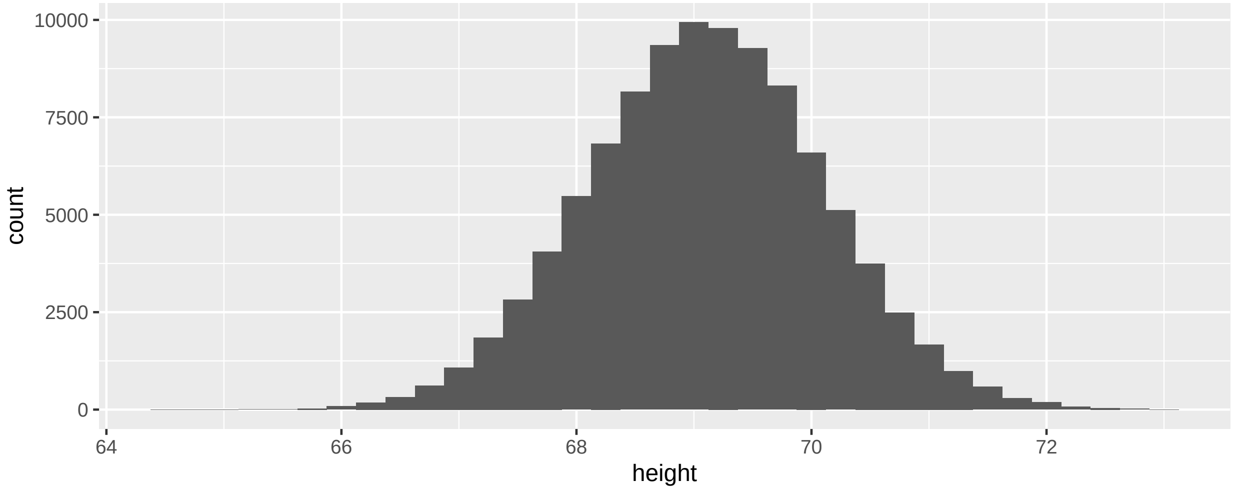
With normally distributed or mound-shaped data, the mode, median, and mean will converge on the same value. With skewed data, one of the tails in the distribution is pulled away from the center. With right or positive skew, the right tail is pulled, and with left of negative skew, the left tail is pulled. Specifically, the mean gets pulled away from the median. Recall that this is because the mean is sensitive to outliers, so it will always be pulled in the direction of those outliers.

Much of the data that we utilize in criminology and the social sciences in general are skewed (e.g., income, number of accidents, crime rates, number of arrests, sentence length). With skewed data, it is important to report both the median and the mean. This is because the median is usually a better measure of central tendency with skewed data, but the mean is more desirable for statistical inference.

A useful general rule is that if the mean is greater than the median, there is positive skew in the data. If the mean is less than the median, there is negative skew.

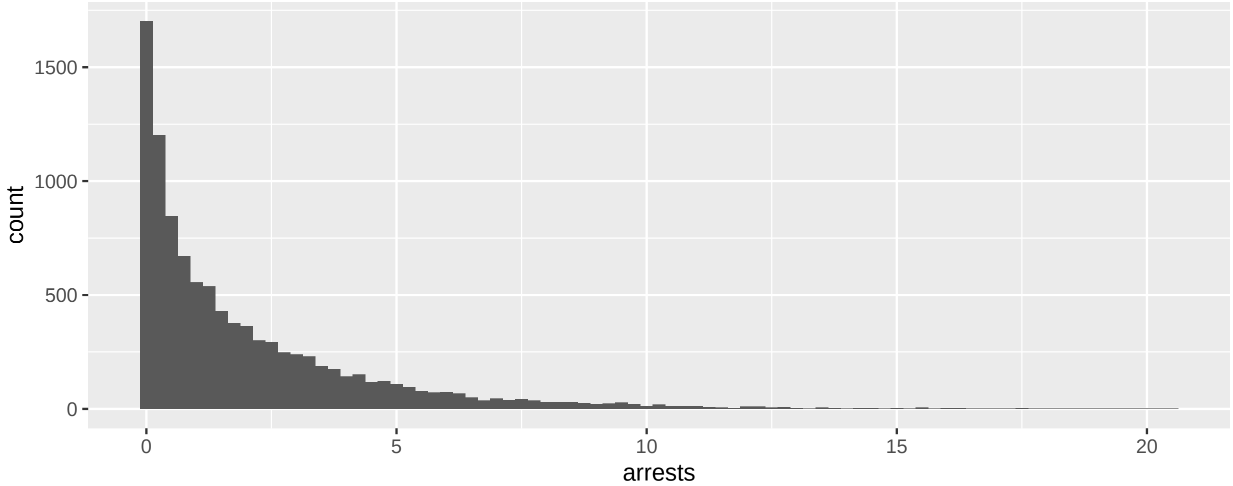
Normal Data

This is a simulated normal distribution using the rnorm function in R to create a distribution of height with 100000 observations and a mean of 69.1 inches.



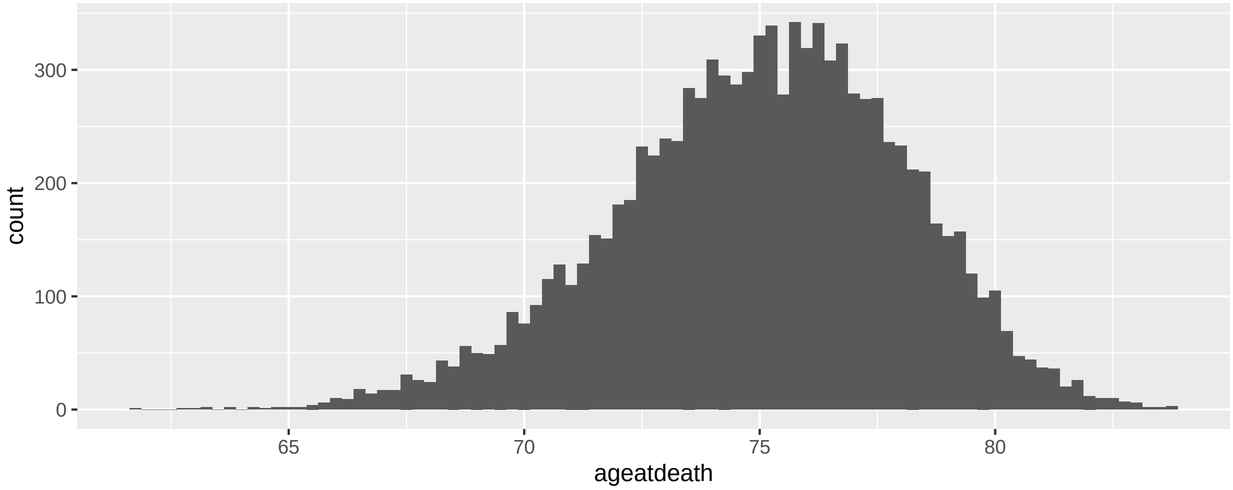
Positive (Right) Skew

Here's another distribution but this time I create one with positive skew - this can be accomplished by squaring the values of a normal distribution (because polynomial transformations are non-linear). The hallmark of a positively skewed distribution is that it has a long right tail that "pulls" the mean toward it such that the mode will typically be smaller than the median and the median will be smaller than the mean.



Negative (Left) Skew

And the final distribution with negative (left) skew - I use the rbeta() function to create a skewed distribution of age at death. The hallmark of a negatively skewed distribution is that it has a long left tail that "pulls" the mean toward it such that the mode is larger than the median and the median is larger than the mean.



Least Squares Property of the Mean

The "least squares" property of the mean indicates that 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:

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

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

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)