

Central Tendency

Chapter 3

+ Class Outline – 7-09-08

- Central tendency – Chapter 3
- Variability – Chapter 4
- Problem Set #1

+ Central Tendency

- Some of this may be a review – but it is important!
- Most other topics we cover in this class will be based on central tendency
- Central Tendency:
 - a statistical measure to determine a single score that defines the center of a distribution.
 - Goal is to find the single score that is most typical or most representative of the entire group.

+ Central Tendency

- Want to find a single number to represent the entire distribution
- Typical, average
- Ex. Describe the summer temperature in NJ
 - Average temp is 80 degrees
 - Is every summer day 80 degrees?
 - No, but the average gives you a general idea of what most days will be like

+ Three distributions demonstrating the difficulty of defining central tendency. In each case, try to locate the "center" of the distribution.

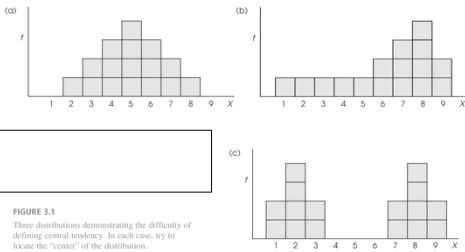


FIGURE 3.1
Three distributions demonstrating the difficulty of defining central tendency. In each case, try to locate the "center" of the distribution.

+ Three Main Measures of Central Tendency

- Mean
- Median
- Mode

+ The Mean

- The Mean - most common descriptor of central tendency (also known as arithmetic average)
- Mean = sum of scores/ number of scores
- population mean = μ (mu)
 - $\mu = \sum X / N$
- sample mean = \bar{X} (X-bar) or M
 - $M = \sum X / n$
- Work backwards - you can find the total (sum of X) by multiplying the mean by N (or n)

+ The Weighted Mean

- Weighted Mean - combine 2 sets of scores with unequal n and find the mean for the 2 sets.
- Because sample sizes are not the same, cannot just take the mean of the two means
- Need 4 pieces of information
 - $\sum X_1, \sum X_2, n_1, n_2$
 - Overall/ weighted mean = $(\sum X_1 + \sum X_2) / (n_1 + n_2)$
 - The weighted mean will be closer to the mean of the **larger sample**
 - Let's try one!

+ Computing mean from frequency table

- Can easily figure out info:
 - $\sum X = \sum (Xf)$
 - $n = \sum f$
 - $M = \sum X / n$

X	f
5	3
4	4
3	2
2	1
1	2

+
The Mean

- Characteristics of the mean
 - affected by every score in distribution (not true of median and mode)
 - changing any score will affect the mean unless the new score or removed score is exactly equal to the mean
- Adding or subtracting a constant to each score will affect the mean by that score
- Multiplying or dividing a constant to each score will change the mean in the same way
- (EXAMPLES)

+
The Median

- The Median - score that divides distribution exactly in half
 - no symbols or notation (sometimes Mdn.)
 - computation the same for population as sample
- When N is odd, use # in middle
- When N is even, put scores in order from low to high and then find mean of the two middle scores
- Examples

+
The Median

- Median may be used when we want to divide sample into two groups (low vs high on test scores)
 - Median split - example
- Mean is not always the *middle* but median is always the middle in terms of frequency of scores ($\frac{1}{2}$ of scores will be on one side of the median and $\frac{1}{2}$ will be on the other)
 - Set of scores - 0, 2, 2, 4, 10, 12

+ The Mode

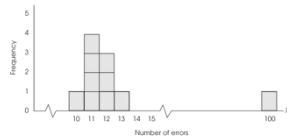
- The Mode - most common observation in a group of scores
 - in frequency distribution, it is tallest point
 - no symbols or notation
 - good for finding info. in nominal scale
 - It is possible to have more than one mode
 - 2 modes - bimodal
 - more than 2 modes - multimodal
 - or, if there are a lot, say there is no mode
 - can have major mode and minor mode though one is slightly higher

+ Which one to use?

- Sometimes mean, median, and mode will be the same or similar - other times they will be different
- Mean uses every score in distribution, which is good and bad
 - good: uses every score so it's a representative value
 - bad: outliers will throw it off
- Mean is most commonly used measure of central tendency

+ Which one to use?

- If a distribution has extreme scores, mean will not be very representative
 - the more scores there are, the less that the extreme scores will influence the mean
 - Example:
 - $M = \sum X / n = 203 / 10$
 - $M = 20.3$
 - Median = 11.5
 - Mode = 11



+ Which one to use?

- Because median typically unaffected by extreme scores, the median is typically used with skewed distributions (i.e. median income)
- Sometimes in research, extreme scores (called outliers) may be removed from the data set if they are atypical
- Book provides other examples of when to use median (missing data, open-ended distributions, ordinal data)

+ Which one to use?

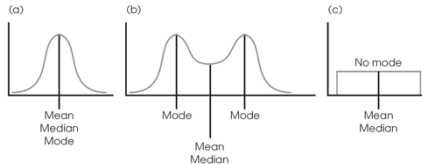
- Use the mode for:
 - Nominal scales
 - discrete variables (# of children in a family)
 - Ex. average family has 2.33 children vs. modal family has 2 children
 - describing shape - add mode to median or mean to give indication of distribution shape

+ Relationship between measures

- Relationship between mean, median, and mode is determined by the shape of the distribution
- For a symmetrical distribution, the mean and median will be exactly the same
 - in symmetrical distribution with only one mode, mode will be the same as the median and mean

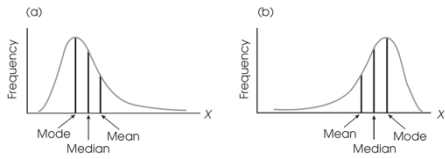
+ Relationship between measures

- bimodal distribution that is symmetrical will have mean and median in the center and modes on each side



+ Relationship between measures

- Positive Skew - mode will be at top, median on right of mode, mean on right of median (mean influenced by extreme scores in the tail). Opposite for negative skew



+ Relationship between measures

- Knowing relationship can help you determine shape
- If mean is 10, and mode is 20, what is the shape of the distribution?
 - Negatively skewed

+

Variability

Chapter 4

+ Chapter 4 - Variability

- ❖ **Variability:** a quantitative measure of the degree to which scores in a distribution are spread out or clustered together
- ❖ If all the scores in a distribution were the same, we would say that there is no variability
- ❖ Central tendency gives a limited view of scores - variability adds another dimension

+ Purpose of Variability

- Two distributions can have the same mean, but look very different
 - Example on board – which has more variability?
- 1) Variability tells you how scores are spread throughout a distribution (the distance to expect between scores)
- 2) Variability measures how well an individual score (or group of scores) represents the entire distribution
 - Important for inferential statistics (sampling from true population)

+ Variability and Inferential Statistics

- If we randomly select one person to represent the population:
 - If variability is small, that person will be fairly close to the population mean
 - If variability is large, we would probably not obtain someone close to the population mean
- Variability provides information about how much error to expect if you are using a sample to represent a population

+ Types of Variability

- Range
- Interquartile Range (skim)
- Standard Deviation

+ The Range

- The range is the difference between the largest (highest) score and the smallest (lowest) score in a distribution
- IGNORE WHAT THE BOOK SAYS ABOUT REAL LIMITS
- For this class, range = $X_{max} - X_{min}$

+ The Range

- Highest X - Lowest X = range
 - If you collect the scores 4, 7, 2, 6, 3, 8 - range = ?
 - If you collect the scores 13, 1, 2, 9, 4, 1 - range = ?
 - Find the mean of both samples
 - Which one has more variability?
- Range is easiest and most obvious way of describing how spread out the scores are
- Problem - range solely determined by two most extreme values
 - Ignores all other scores
 - What about outliers and skewed distributions?

+ The Interquartile Range

- Interquartile range is used to avoid excessive influence of extreme scores
- Interquartile range - the range covered by the middle 50% of the distribution

+ Standard Deviation & Variance

- ◇ Standard deviation is the most common measure of variability
- ◇ Standard deviation uses the mean as a reference point and APPROXIMATES the average distance of each score from the mean
- ◇ Deviation = distance and direction from the mean
 - ◇ $X - \mu$
 - ◇ Example on board

+ Standard Deviation & Variance

- Deviation score tells you 2 things:
 - Distance from mean
 - Direction (+ or -): whether the score is above or below the mean
- Goal is to find the standard distance from the mean
 - Average distance?
 - Impossible to take the mean of deviation scores, because they sum to zero
 - This is because mean is balance point for the distribution - total of deviations above the mean will always equal total of deviations below the mean
 - (example)

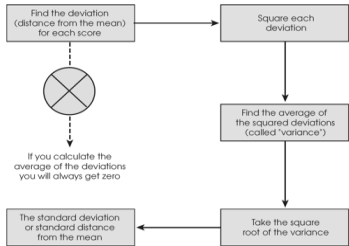
+ Standard Deviation & Variance

- Problem = + and - signs add up to zero
- Solution = get rid of + and - signs by squaring each deviation
- If we then calculate the mean of these squared deviations (add up each squared deviation and divide by N), we get what is called VARIANCE or the mean squared deviation

+ Standard Deviation & Variance

- Variance is based on squared distances, but we want a measure of the standard distance from the mean
- To get the STANDARD DEVIATION, we must correct for the fact that everything was squared
- Take the square root of the variance:
 - standard deviation = $\sqrt{\text{variance}}$

+ Computing the Standard Deviation



+ Formulas

❖ Calculations differ slightly from population to sample though underlying concepts are the same

❖ Variance = mean squared deviation =

$$\frac{\text{sum of squared deviations}}{\# \text{ of scores}}$$

❖ Sum of squared deviations is basic component of variability know as the SUM OF SQUARES or SS for short

+ Formulas

❖ There are two types of formulas that your book will present, **definitional** and **computational**.

❖ Definitional formulas make more sense and enable you to see what you will be calculating in an easier way, but they can be difficult to calculate and lead to rounding error

❖ Computational formulas are not as straight forward to read but are easier to use when doing the math

+ Formulas for Population

- Definitional formula for SS:
 - $SS = \sum (X - \mu)^2$
 - Which is telling you to:
 - Find each deviation score
 - Square each deviation score
 - Add them together
- Computational formula for SS:
 - $SS = \sum X^2 - \frac{(\sum X)^2}{N}$
- Population Variance = SS / N
- Standard deviation = $\sqrt{SS / N}$

+ Notation

- These are all population parameters so they are identified by Greek letter
 - Lower case sigma: σ
- Population standard deviation = σ
- Population variance = σ^2

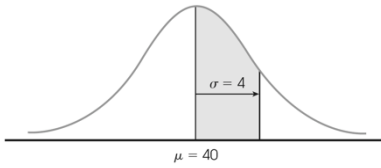
+ Your turn!

- Given the following set of scores, compute the SS using both the definitional and computational formulas:

X	$X - \mu$	$(X - \mu)^2$
4		
2		
7		
9		
3		
$\mu =$		$\sum (X - \mu)^2 =$

X	X^2
4	
2	
7	
9	
3	
$\sum X =$	$\sum X^2 =$
$(\sum X)^2 / N =$	SS =

+ Graphical Representation



+ Sample σ^2 and σ

- Inferential statistics uses limited information from samples to draw general conclusions about a population
- Samples should be representative of the populations from which they come
- This is a bit of a problem because samples are consistently less variable than the population from which they come – example (increasing sample size improves this)

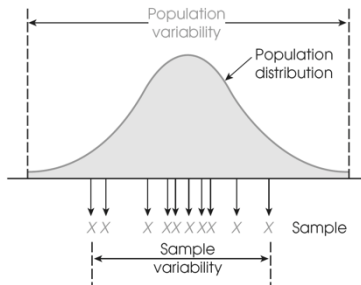


Figure 4-6
The population of adult heights forms a normal distribution. If you select a sample from the population, you are most likely to obtain individuals who are near average in height. As a result, the scores in the sample will be less variable (spread out) than the scores of the population.

+ Samples

- This basically means that the variability of a sample gives a biased estimate of the population variability (a sample will naturally draw more from the center of a normal distribution)
- We must therefore make a small adjustment in our calculation of sample variance and standard deviation to make it a more accurate estimate of population variability

+ Samples

- To do this, our SS calculations are exactly the same as population. But now we divide the SS by $n - 1$ instead of N .
 - This will make the variance and standard deviation slightly larger - EXAMPLE
 - As sample size increases, this $n - 1$ adjustment has less of an impact
- Also, notation changes
 - Use M instead of μ
 - Use n instead of N

+ Formulas for Samples

- ◇ Definitional formula: $SS = \sum (X - M)^2$
- ◇ Computational formula: $SS = \sum X^2 - \frac{(\sum X)^2}{n}$
- ◇ Note that is exactly the same as with population

◇ Sample variance = $s^2 = \frac{SS}{n - 1}$

◇ Sample std. deviation = $s = \sqrt{\frac{SS}{n - 1}}$

- ◇ $n - 1$ is also called degrees of freedom or df

+ Degrees of Freedom

- With a population, you find the deviation of each score from μ , the true population mean
- With a sample, μ is unknown, so must approximate (or estimate) with M
- But, when you know what the mean is for a sample (M), it restricts how the scores can vary in your sample

+ Degrees of Freedom

- ◆ For example, if $n = 5$ and $M = 10$, the first 4 scores can have any value (they are free to vary)
- ◆ The fifth score, however, must equal a certain value in order to make the mean 10
- ◆ Example: randomly draw scores - first score is 8, second is 15, then 3, then 6
 - ◆ Since we know that the mean is 10, the last score has to equal 18 ($8+15+3+6+18 = 50/5 = 10$), it is restricted.
 - ◆ We therefore say that the first $n - 1$ scores are free to vary, or that the sample has $n - 1$ degrees of freedom; $df = n - 1$

+ Unbiased estimates

- Sample variance should provide an unbiased estimate of population variance
- This means that the average of all sample variances will produce accurate estimate of population variance
 - Does not mean that any one sample variance will be equal to population variance
- A statistic is said to be biased if average value of the statistic underestimates or overestimates the population parameter CONSISTENTLY

+ Use of Standard Deviation

- ◆ What does the standard deviation tell us about individual scores?
 - ◆ Can tell whether the score is extreme or not

- ◆ If the mean of a distribution is 25 and the standard deviation is 5, what does a score of 33 tell you? What if the standard deviation is 10? What if the standard deviation is 2?

- ◆ Mean and standard deviation are not merely abstract concepts, they are meaningful in the context of a set of scores – they tell you a fair amount about how a set of scores are distributed

+ Use of Sample Variance

Sample variance tells us two important things about a population:

1. Sample variance gives indication of how accurately a score or a sample represents the population (if variance is small, can select any score and it should reflect the population fairly accurately)
2. A large sample variance means that a pattern is more difficult to detect (because scores are all over the place, there is less of a predictable pattern) **EXAMPLE** – give one group of rats drug and another placebo – if variance is small can conclude more than if variance is large. (figure 4.9 p. 106)

+ Error Variance

- For inferential statistics, the variance that exists in a set of sample data is often termed *error variance*
 - Error variance = unexplained and uncontrolled differences between scores

- This is distinguished from treatment effects, which is the variance **BETWEEN** samples – over and above the variance within samples.

+ In sum

- Looking at mean differences is not enough
- Must consider variability (spread) of scores in determining whether a sample adequately represents a population, or whether two samples differ

+ Transformations

- How does transforming a data set (adding, subtracting, multiplying or dividing by a constant) affect the standard deviation and variance?
 - Adding (subtracting) a constant does not change the SD
 - Example: curving an exam
 - Multiplying (dividing) by a constant causes the standard deviation to be multiplied (or divided) by the same amount
 - Example: feet to inches

Problem Set #1
Due 7/14/08
Look it over tonight to see if you have questions.
