Measures of Variability

Descriptive Statistics Part 2

Cal State Northridge

Ψ320

Andrew Ainsworth PhD

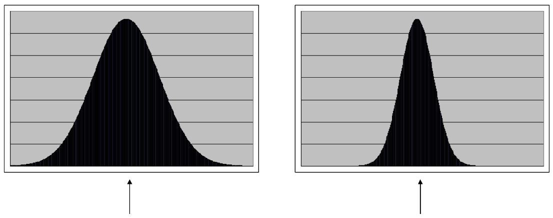
Reducing Distributions

- Regardless of numbers of scores, distributions can be described with three pieces of info:
  - Shape (Normal, Skewed, etc.)
  - Central Tendency
  - Variability

How do scores spread out?

- Variability
  - Tell us how far scores spread out
  - Tells us how the degree to which scores deviate from the central tendency

How are these different?



Measure of Variability

Measure	Definition	Related to:
Range	Largest - Smallest	Mode
Interquartile Range	X_{75} - X_{25}	Median
Semi-Interquartile Range	(X_{75} - X_{25})/2	
Average Absolute Deviation	$\frac{\sum \lvert X_i - \bar{X} \rvert}{N}$	Mean
Variance	$\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}$	
Standard Deviation	$\sqrt{\frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N-1}}$	

The Range

- The simplest measure of variability
  - **Range (R) = $X_{highest}$ – X_{lowest}**
  - Advantage – Easy to Calculate
  - Disadvantages
    - Like Median, only dependent on two scores → unstable

      {0, 8, 9, 9, 11, 53} Range = 53

      {0, 8, 9, 9, 11, 11} Range = 11
    - Does not reflect all scores

Detour: Percentile

- A **percentile** is the score at which a specified percentage of scores in a distribution fall below
  - To say a score 53 is in the 75th percentile is to say that 75% of all scores are less than 53
- The **percentile rank** of a score indicates the percentage of scores in the distribution that fall at or below that score.
  - Thus, for example, to say that the percentile rank of 53 is 75, is to say that 75% of the scores on the exam are less than 53.

Detour: Percentile

- **Scores** which divide distributions into specific proportions
  - Percentiles = hundredths

    P1, P2, P3, … P97, P98, P99
  - Quartiles = quarters

    Q1, Q2, Q3
  - Deciles = tenths

    D1, D2, D3, D4, D5, D6, D7, D8, D9
- **Percentiles are the SCORES**

Detour: Percentile Ranks

- What percent of the scores fall below a particular score?

$$PR = \frac{(Rank - .5)}{\text{N}} \times 100$$

- **Percentile Ranks are the Ranks not the scores**

Detour: Percentile Rank

- **Ranking no ties** – just number them

Score:	1	3	4	5	6	7	8	10
Rank:	1	2	3	4	5	6	7	8

- **Ranking with ties** - assign midpoint to ties

Score:	1	3	4	6	6	8	8	8
Rank:	1	2	3	4.5	4.5	7	7	7

	Step 1	Step 2	Step 3	Step 4
Data	Order	Number	Assign Midpoint to Ties	Percentile Rank (Apply Formula)
9	1	1	1	2.381
5	2	2	2	7.143
2	3	3	4	16.667
3	3	4	4	16.667
3	3	5	4	16.667
4	4	6	7	30.952
8	4	7	7	30.952
9	4	8	7	30.952
1	5	9	10	45.238
7	5	10	10	45.238
4	5	11	10	45.238
8	6	12	12	54.762
3	7	13	14	64.286
7	7	14	14	64.286
6	7	15	14	64.286
5	8	16	17.5	80.952
7	8	17	17.5	80.952
4	8	18	17.5	80.952
5	8	19	17.5	80.952
8	9	20	20.5	95.238
8	9	21	20.5	95.238

- Steps to Calculating Percentile Ranks
- Example:

$$PR_3 = \frac{(Rank_3 - .5)}{\text{N}} \times 100 =$$

$$\frac{(4 - .5)}{21} \times 100 = 16.667$$

Detour: Finding a Percentile in a Distribution

$$X_P = (p)(n+1)$$

- Where X_P is the score at the desired percentile, p is the desired percentile (a number between 0 and 1) and n is the number of scores)
- If the number is an integer, than the desired percentile is that number
- If the number is not an integer than you can either round or interpolate

Detour: Interpolation Method Steps

- Apply the formula $X_P = (p)(n+1)$
  1. You'll get a number like 7.5 (think of it as *place1.proportion)*
  2. Start with the value indicated by *place1* (e.g. 7.5, start with the value in the 7th place)
  3. Find *place2* which is the next highest place number (e.g. the 8th place) and subtract the value in *place1* from the *value* in *place2*, this *distance1*
  4. Multiple the *proportion* number by the *distance1* value, this is *distance2*
  5. Add *distance2* to the value in *place1* and that is the *interpolated value*

13

Detour: Finding a Percentile in a Distribution

- Interpolation Method Example:
- 25th percentile:
  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- X_{25} = (.25)(9+1) = 2.5
  - *place1* = 2, *proportion* = .5
  - Value in *place1* = 4
  - Value in *place2* = 9
  - *distance1* = 9 – 4 = 5
  - *distance2* = 5 * .5 = *2.5*
  - *Interpolated value* = 4 + 2.5 = 6.5
  - 6.5 is the 25th percentile

14

Detour: Finding a Percentile in a Distribution

- Interpolation Method Example 2:
- 75th percentile
  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- X_{75} = (.75)(9+1) = 7.5
  - *place1* = 7, *proportion* = .5
  - Value in *place1* = 49
  - Value in *place2* = 64
  - *distance1* = 64 – 49 = 15
  - *distance2* = 15 * .5 = 7.5
  - *Interpolated value* = 49 + 7.5 = 56.5
  - 56.5 is the 75th percentile

15

Detour: Rounding Method Steps

- Apply the formula $X_P = (p)(n+1)$
  1. You'll get a number like 7.5 (think of it as *place1.proportion)*
  2. If the *proportion* value is any value other than exactly .5 round normally
  3. If the *proportion* value is exactly .5
     - And the *p* value you're looking for is above .5 round down (e.g. if *p* is .75 and X_p = 7.5 round down to 7)
     - And the *p* value you're looking for is below .5 round up (e.g. if *p* is .25 and X_p = 2.5 round up to 3)

Detour: Finding a Percentile in a Distribution

- Rounding Method Example:
- 25^{th} percentile

  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- X_{25} = (.25)(9+1) = 2.5 (which becomes 3 after rounding up),
- The 3^{rd} score is 9, so 9 is the 25^{th} percentile

Detour: Finding a Percentile in a Distribution

- Rounding Method Example 2:
- 75^{th} percentile

  {1, 4, 9, 16, 25, 36, 49, 64, 81}
- X_{75} = (.75)(9+1) = 7.5 which becomes 7 after rounding down
- The 7^{th} score is 49 so 49 is the 75^{th} percentile

Detour: Quartiles

- To calculate Quartiles you simply find the scores the correspond to the 25, 50 and 75 percentiles.
- $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$

Back to Variability: IQR

- Interquartile Range
  - $= P_{75} - P_{25}$ or $Q_3 - Q_1$
  - This helps to get a range that is not influenced by the extreme high and low scores
  - Where the range is the spread across 100% of the scores, the IQR is the spread across the middle 50%

Variability: SIQR

- Semi-interquartile range
  - $=(P_{75} - P_{25})/2$ or $(Q_3 - Q_1)/2$
  - IQR/2
  - This is the spread of the middle 25% of the data
  - The average distance of Q1 and Q3 from the median
  - Better for skewed data

Variability: SIQR

- Semi-Interquartile range

Average Absolute Deviation

- Average distance of *all* scores from the mean disregarding direction.

$$AAD = \frac{\sum \left| X_i - \bar{X} \right|}{N}$$

Average Absolute Deviation

	Score	$(X_i - \bar{X})$	$\left\| X_i - \bar{X} \right\|$
	8	3	3
	6	1	1
	4	-1	1
	2	-3	3
Sum=		0	8
Mean=	5		2

Average Absolute Deviation

- Advantages
  - Uses all scores
  - Calculations based on a measure of central tendency - the mean.
- Disadvantages
  - Uses absolute values, disregards direction
  - Discards information
- Cannot be used for further calculations

Variance

- The average squared distance of each score from the mean
- Also known as the **mean square**
- Variance of a sample: s^2
- Variance of a population: σ^2

Variance

- When calculated for a sample

$$s^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{N-1}$$

- When calculated for the entire population

$$\sigma^2 = \frac{\sum\left(X_i - \bar{X}\right)^2}{N}$$

Variance

- Variance Example
  - Data set = {8, 6, 4, 2}
  - Step 1: Find the Mean

$$\overline{X} = \frac{\_\_ + \_\_ + \_\_ + \_\_}{\_\_} = \_\_\_\_$$

Psy 320 - Cal State Northridge 28

Variance

- Variance Example
  - Data set = {8, 6, 4, 2}
  - Step 2: Subtract mean from each value

Score	Deviation
8	(8 - ____) = _____
6	(6 - ____) = _____
4	(4 - ____) = _____
2	(2 - ____) = _____

29

Variance

- Variance Example
  - Data set = {8, 6, 4, 2}
  - Step 3: Square each deviation

Score	Deviation	Squared
8	____	____
6	____	____
4	____	____
2	____	____

30

Variance

- Variance Example
  - Data set = {8, 6, 4, 2}
  - Step 4: Add the squared deviations and divide by N - 1

$$s^2 = \frac{\_\_ + \_\_ + \_\_ + \_\_}{\_\_ - 1} = \_\_\_\_$$

31

Standard Deviation

- Variance is in squared units
- What about regular old units
- Standard Deviation = Square root of the variance

$$s = \sqrt{\frac{\sum \left(X_i - \bar{X} \right)^2}{N - 1}}$$

Standard Deviation

- Uses measure of central tendency (i.e. mean)
- Uses all data points
- Has a special relationship with the normal curve (we'll see this soon)
- Can be used in further calculations
- Standard Deviation of Sample = *SD* or *s*
- Standard Deviation of Population = σ

Why N-1?

- When using a sample (which we always do) we want a statistic that is the best estimate of the parameter

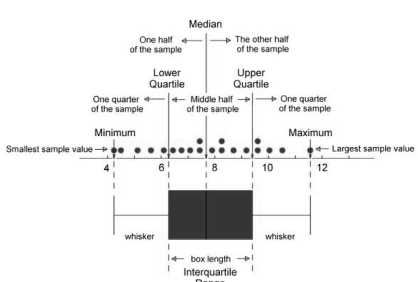
$$E\left(\frac{\sum(X_i - \bar{X})^2}{N-1}\right) = \sigma^2 \qquad E\left(\sqrt{\frac{\sum(X_i - \bar{X})^2}{N-1}}\right) = \sigma$$

Degrees of Freedom

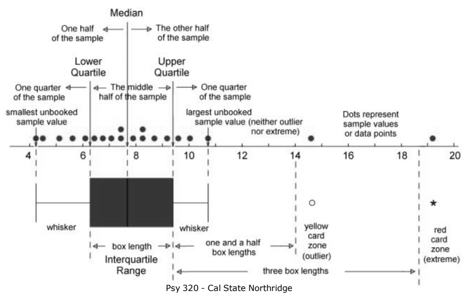
- Usually referred to as *df*
- Number of observations minus the number of restrictions

__+__+__+__=10 - 4 free spaces
2 +__+__+__=10 - 3 free spaces
2 + 4 +__+__=10 - 2 free spaces
2 + 4 + 3 +__=10
Last space is not free!! Only 3 dfs.

Boxplots



Boxplots with Outliers



Computational Formulas

- Algebraic Equivalents that are easier to calculate

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

$$s = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N-1}} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}}$$