Topic 6a Measures of Center

Overview

1. Three characteristics of quantitative variables
    - Shape - Visually via graphs
    - Center - Numeric summery
    - Spread - Numeric summery
2. Analyzing population versus analyzing samples
    - For populations:
        - we know all data
        - parameters
        - using greek letters μ
    - For samples:
        - we know part of data
        - statistics
        - using roman letters $\overline{x}$

Center of the data

1. Center of the data: Numeric values that represent the $\underline{average}$ or typical value of a quantitative variable.

- example of measure central tendency:
    - mean and median
- Properties of mean
    - non-resistant measure of center (easily influenced by outlier)
- Properties of median
    - resistant measure

2. Note: a fairly standard convention is to round the average to **one more decimal place than the data**
3. Note： One interpretation of the mean: The arithmetic mean can be thought of as the center of gravity.
4. Comparing Mean and Median
    1. Skewed left– mean will usually be smaller than the median.
    2. Symmetric– mean will usually be close to the median.
    3. Skewed right– mean will usually be larger than the median.

Variation or Meansures of Dispersion

Numeric values that represent the degree to which the values are $\underline{spreadout}$.

- examples of Meansures of Dispersion:
    - range, quaties, variance, standard deviation

Topic 6b Measure of Spread

Overview

- Measure of Spread: quartiles, standard deviation
- Five-number summary and boxplot
- Choosing among summary statistics
- Changing the unit of measurement

1. Spread or Variability
   Variability exists when some values are different from the mean
2. Numerical values that represent the degree to which the values are spread out
   - range: Non-resistant measure of spread
   - quartiles： Resistant measure of spread
     note: The interquartile range(IQR) is the difference between the third and first quartiles
   - variance, standard deviation: Noe-resistant meansure of spread
3. The five-number summary
   It contains Min, Q_1, M, Q_3, Max
4. Boxplot
   - Central box spans Q_1 and Q_3
   - The line marks the median M
5. Outliers:
   - Outliers could be:
     - Chance occurrences
     - Measurement errors
     - Data entry errors
     - Sampling errorsP
   - Outliers are not necessary invalid data
   - Fence Rule:
     1. Calculate lower and upper fences:
        Lower fence = Q_1-(1.5xIQR)
        Upper fence = $Q3$+(1.5xIQR)
        Values (strictly) less than or larger than the fence could be considered outliers
6. Variance and Standard Deviation
   1. Population Variance / Sample Variance:
      Population Variance:

$$\sigma^2 = \frac{1}{N} * \sum_{i=1}^{N} (x_i - \mu)^2$$

      Sample Variance: **an Unbiased Estimator**

$$s^2 = \frac{1}{N} * \sum_{i=1}^{n} (x_i - \mu)^2$$

7. Which measure to use

- The five-number summary should be used to describe center and spread for skewed distributions, or when outliers are present.
- Use the mean and standard deviation for reasonably symmetric distributions that are free of outliers.

Topic 6c Describing Data Using SAS

1. proc means

```
proc means data = example1 n mean std
```

"proc mean" provides basic descriptive statistics. we can add options to request more information

- n: sample size
- nmiss: number of observations with missing values
- mean: sample mean
- median: median
- std: standard deviation
- stderr: standard error
- clm: lower and upper two-sided 95% confidence interval for the mean
- lclm/uclm: lower/upper one-sided 95% CI for mean
- min: minimum
- max: maximum
- sum: sum
- var: variance
- q1: first quartile
- q3: third quartile
- qrange: interquartile range (IQR)
- cv: coefficient of variation
- skewness: skewness
- kurtosis: kurtosis
- T: student t-test, testing the null hypothesis that the population mean is 0
- PRT: Probability of obtaining a larger absolute value of t under the null hypothesis
- MAXDEC=n : specifies the number of decimal places for printed statistics

2. proc univariate

This command can compute: 1. The number of observations (nonmissing) 2. Mean 3. Standard deviation 4. Variance 5. Skewness 6. Kurtosis 7. Uncorrected and corrected sum of squares 8. Coefficient of Variation 9. Standard error of the mean 10.A t-test comparing the variable's value against zero 11.Maximum (largest value)12.Minimum (smallest value) 13.Range 14.Median, 1st, and 3rdquartiles 15.Interquartile range 16. Mode 17. 1st, 5th,10th, 90th,95th, and 99thpercentiles 18. The five highest and five lowest values (useful for data checking) 19. W or D statistic to test whether data are normally distributed20. Stem-and-leaf plot 21. Boxplot 22. Normal probability plot, comparing our cumulative frequency distribution to a normal distribution.

```
proc univariate DATA=my_data NORMAL PLOT;
title “title”;
```

```
VAR HEIGHT WEIGHT;
RUN;
```

options:

- Skewness: measure of the symmetry of asymmetry of the distribution
- Kurtosis: measure of the flatness of the distribution
- USS: Uncorrected sum of squares (the sum of the scores squared – each score is squared and the squares are added together)
- CSS: Corrected sum of squares (the sum of squared deviations from the mean, usually more useful than USS)
- CV: Coefficient of variation (the standard deviation divided by the square root of n)
- Std Mean: Standard error of the mean (the standard deviation divided by the square root of n)
- T:Mean=0: Student's t-test for testing the hypothesis that the population mean is zero
- Prob>|T| : The p-value for the t-test
- Num^=0: Number of nonzero observations
- Prob<W/Prob>D : P-value testing the null hypothesis that the population is normally distributed (when theD:Normal test is done, the statistic is Prob>D)