

Chapter 3 Descriptive Statistics: Numerical Methods

Descriptive Summary Measures

- Ways to numerically summarize a data set
- Measures of Location (Central Tendency)
- Measures of Variability
- Measures of Relative Standing

2

Inference

- Procedures where we estimate the value of a **parameter** based on the value of a **statistic**
- Symbols

Parameter	Summary Measure	Statistic
μ	Mean	\bar{X}
σ^2	Variance	s^2
σ	Std Deviation	s
p	Proportion (RF)	\bar{p}

3



Measures of Location

- aka Measures of Central Tendency
- A summary measure indicating what is **typical**
- 5 measures of central tendency
 - Mean
 - Median
 - Mode
 - Weighted Mean (3.6)
 - Geometric Mean (nib)

4



Measures of Location: Mean

- The simple arithmetic average
- Explicitly uses every value in the data set:

$$\mu = \frac{\sum x}{N} \quad \bar{X} = \frac{\sum x}{n}$$

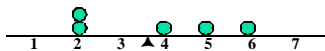
- Gas Price assignment

5



Measures of Location: Mean

- Data set 1: {2, 2, 4, 5, 6} mean = _____



- The **total (net) deviation from the mean** will always be **zero**
 - the mean represents the **balance point** of the data set
- Data set 2: {2, 2, 4, 5, 60} mean = ____
- Outlier
- Gas Price assignment

6

Measures of Location: Median

- The value located in the **middle** of the **ordered data array**

$$\text{median's location: } \frac{(n+1)}{2}$$

- Does **not** explicitly use every value
 - data set 1: 2, 2, 4, 5, 6 mean = ___ median = _____
 - data set 2: 2, 2, 4, 5, 60 mean = ___ median = _____
- When a data set has extreme values, the median is the preferred measure of central location
- Gas Price assignment

7

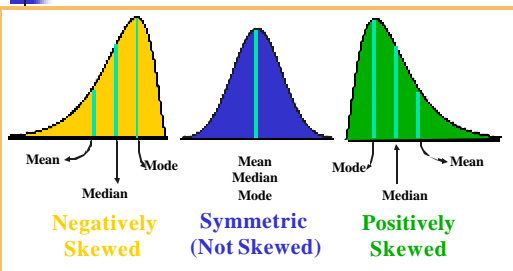
Measures of Location: Mode

- The value that occurs with the highest frequency
- If no values repeat, the mode is undefined
- Bimodal distribution
- Multimodal distribution

- Gas Price assignment

8

Relative Locations for Measures of Location



9

Measures of Location: Percentiles

- The p^{th} percentile is a value such that at least $p\%$ of the observations are less than or equal to this value and at least $(100 - p)\%$ of the observations are greater than or equal to the p^{th} percentile.
- $P_{50} = \text{median} = Q_2$
- similar to CRF
- gas price ogive: 55¢/gal \approx _____ percentile?

10

Measures of Location: Percentiles

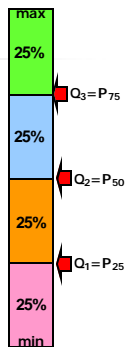
- 1 Prepare the ordered data array.
- 2 Compute i , the location of the p^{th} percentile.

$$i = \left(\frac{p}{100} \right) n$$
 - If i is not an integer, round up. The p^{th} percentile is the value in the i^{th} position.
 - If i is an integer, the p^{th} percentile is the average of the values in positions i and $i + 1$.
- 3
 - Gas Price assignment

11

Measures of Location: Quartiles

- Divide the ordered array into 4 segments, with each segment having an equal number of observations
 - $Q_3 =$ approximately 75% of observations are less than or equal to this value
 - $Q_2 =$ approximately 50% of observations are less than or equal to this value
 - $Q_1 =$ approximately 25% of observations are less than or equal to this value
- Find the equivalent percentile



12

Measures of Location: Weighted Mean

- Used when some of the data values require more weight than others

$$\mu_w \text{ or } \bar{x}_w = \frac{\sum xw}{\sum w}$$

class	grade	credits
CHM 200	A=4	5
GBS 221	A=4	3
ACC 211	B=3	3
PED 101	D=1	1

- Example: measuring academic performance
 - what is this student's average grade?
 - what is this student's grade point average?
- Law Firm assignment

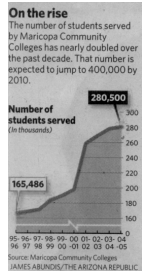
13

Measures of Location: Geometric Mean _(nib)

- Used to determine the mean **percentage rate of change** over a time series

$$GM = \sqrt[n]{\frac{\text{value at end}}{\text{value at start}} - 1}$$

- Example: what was the mean annual enrollment growth rate for MCCCDC between the 1995-6 and 2004-5 academic years?
- Vehicle Crashes assignment



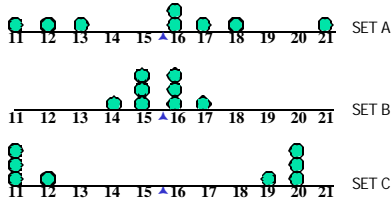
Measures of Variability

- aka Measures of Dispersion
- Measures the **dispersion** among values in a data set
- Measures of **location** alone aren't enough to adequately describe a data set
 - {5, 5, 5}
 - {4, 5, 6}
 - {1, 1, 13}

15

The Need for Measures of Variability

- Each of these data sets has a mean of 15.5
- Which has the least variation?
- Which has the most variation?

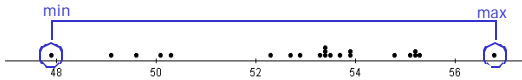


16

Measures of Variability: Range

Range = (Max - Min)

- Simplistic
 - uses only the 2 extreme values
 - ignores the location of all other values
- Outliers will make it deceptive
- Gas Price assignment

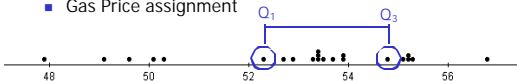


17

Measures of Variability: Interquartile Range

$$IQR = (Q_3 - Q_1) = (P_{75} - P_{25})$$

- aka Middle 50% Range
- Attempts to remove the distorting effects of outliers by trimming the data set
- Gas Price assignment



18

Measures of Variability: Variance

- The average squared deviation from the mean

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N} = \frac{\sum x^2 - N\mu^2}{N}$$

Computational formula

- Explicitly uses every value in the data set
- Sum of deviations from the mean always = 0
- **Reasons for focusing on squared deviations:**
 - squaring prevents (+) & (-) deviations from canceling
 - squaring draws added attention to large deviations
- {2, 2, 4, 5, 6} assignment

19

Measures of Variability: Variance

$$\text{sample: } s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{\sum x^2 - n\bar{x}^2}{n - 1}$$

Computational formula

- s^2 is an estimate of population's variance (σ^2) based on sample data
 - s^2 is **not** the variance within the sample
 - (n-1) is a **correction for bias**
 - degrees of freedom
 - the number of values in a data set that are **free to vary** given that you have estimated other parameters from the same sample data
 - we lose one degree of freedom whenever we use s^2 to estimate σ^2 because we first had to use the sample data to estimate μ with \bar{x}
- Gas Price assignment
- Units of measurement are squared—nonsensical

20

Measures of Variability: Standard Deviation

- The positive square root of variance

$$\text{population: } \sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}} = \sqrt{\frac{\sum x^2 - N\mu^2}{N}}$$

$$\text{sample: } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n - 1}}$$

- units of measurement are not squared
- we'll use daily
- StatCrunch calculates **sample** standard deviation & variance
- Gas Price

21

Common Uses for the Standard Deviation Coefficient of Variation

$$CV = \left(\frac{\text{standard deviation}}{\text{mean}} * 100 \right) \%$$

- A **relative** measure of dispersion
- Useful when comparing the dispersion of data sets having **different magnitudes** or **different units of measurement**
- Gas Price assignment
- Investment assignment

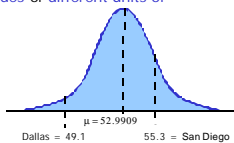
22

Common Uses for the Standard Deviation z-Score

- a.k.a. Standard Score or Standardized Value
- Tells how many standard deviations away from the mean a particular value lies
- Indicates **relative** distance from the mean
- Useful when comparing the relative position among values in data sets with **different magnitudes** or **different units of measurement**

$$Z = \frac{(x - \mu)}{\sigma} \text{ or } \frac{(x - \bar{x})}{s}$$

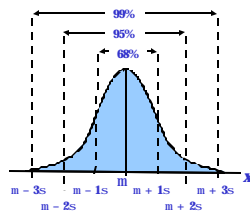
- Gas Price assignment
 - standard score for San Diego?
 - standard score for Dallas?
 - interval endpoints
- Milwaukee vs Beckonsfield assignment



23

Common Uses for the Standard Deviation Empirical Rule

- A Normal distribution has approximately:
 - **68%** of values within **± 1** standard deviation of the mean
 - **95%** of values within **± 2** standard deviations of the mean
 - **99%** of values within **± 3** standard deviations of the mean



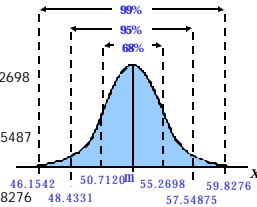
24

Common Uses for the Standard Deviation

Empirical Rule

Gas Price

- mean = 52.9909, std dev = 2.2789
- interval endpoints assignment
- $52.9909 \pm 1(2.2789) = 50.7120 - 55.2698$
contains _____% of cities
- $52.9909 \pm 2(2.2789) = 48.4331 - 57.5487$
contains _____% of cities
- $52.9909 \pm 3(2.2789) = 46.1542 - 59.8276$
contains _____% of cities



Retail Store assignment

25

Common Uses for the Standard Deviation

Empirical Rule

- The distribution of salaries at Mega Corp is approximately Normally distributed and ranges from \$15,500 to 70,100.



- Estimate this distribution's mean and standard deviation
- Approximately what proportion of salaries exceed 51,900?

26

Common Uses for the Standard Deviation

Chebyshev's Theorem

- Regardless of the shape of the distribution:

at least $1 - (1/z^2)$ % of observations within $\pm z$ std deviations of mean

- $\geq 75\%$ of observations fall within 2 standard deviations of the mean
- $\geq 89\%$ of observations fall within 3 standard deviations of the mean
- $\geq 94\%$ of observations fall within 4 standard deviations of the mean
- $\geq \underline{\hspace{1cm}}$ % of observations fall within 2.5 standard deviations of the mean

Gas Price Data Set

- ____% of observations fall within 2 standard deviations of the mean
- ____% of observations fall within 3 standard deviations of the mean

27

Common Uses for the Standard Deviation Detecting Outliers

- What are they?
- Possible causes?
- How detect them?

28

Common Uses for the Standard Deviation Pearson Skew Coefficient (rib)

- A numerical measure of a distribution's skew
 - is used to compare the **relative** skew of 2 or more distributions

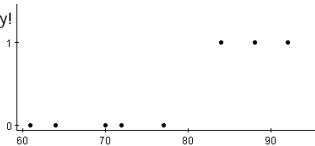
$$P = \frac{3(\text{mean} - \text{median})}{\text{std dev}}$$

- Gas Price

29

Comparative Analysis

- Based on subgroups
- StatCrunch makes it easy!
 - use the **Group by** option
- Exam Score data set



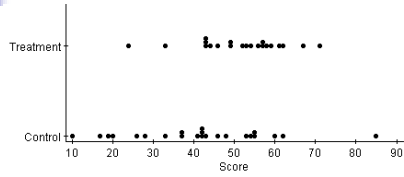
Summary statistics for Score grouped by Gender

Gender	n	Mean	Variance	Std.Dev.	Median	Range	Min	Max	Q1	Q3
0	5	68.8	40.7	6.3796554	70	16	61	77	64	72
1	3	88	16	4	88	8	84	92	84	92

30



Comparative Analysis: Reading



Summary statistics for Score grouped by Group

Group	n	Mean	Variance	Std. Dev.	Median	Range	Min	Max	Q1	Q3
Control	23	41.52174	294.07904	17.148733	42	75	10	85	28	54
Treatment	21	51.47619	121.1619	11.007357	53	47	24	71	44	58

31
