Lecture 4: Measure of Dispersion

Donglei Du
(ddu@unb.edu)

Faculty of Business Administration, University of New Brunswick, NB Canada Fredericton
E3B 9Y2

Table of contents

Layout

1. Measure of Dispersion: scale parameter
   - Introduction
   - Range
   - Mean Absolute Deviation
   - Variance and Standard Deviation
   - Range for grouped data
   - Variance/Standard Deviation for Grouped Data
   - Range for grouped data
2. Coefficient of Variation (CV)
3. Coefficient of Skewness (optional)
   - Skewness Risk
4. Coefficient of Kurtosis (optional)
   - Kurtosis Risk
5. Chebyshev's Theorem and The Empirical rule
   - Chebyshev's Theorem
   - The Empirical rule
6. Correlation Analysis
7. Case study

Introduction

- Dispersion (a.k.a., variability, scatter, or spread)) characterizes how stretched or squeezed of the data.
- A measure of statistical dispersion is a nonnegative real number that is zero if all the data are the same and increases as the data become more diverse.
- Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.
- There are many types of dispersion measures:
    - Range
    - Mean Absolute Deviation
    - Variance/Standard Deviation

Range

- $$\text{Range} = \max - \min$$

Example

- Example: Find the range of the following data: 10, 7, 6, 2, 5, 4, 8, 1
- $$\text{Range} = 10 - 1 = 9$$
- Example: the weight example (weight.csv)
- The R code:

```
>weight <- read.csv("weight.csv")
>sec_01A<-weight$Weight.01A.2013Fall
>range(sec_01A)
 [1] 130 189
```

- $$\text{Range} = 189 - 130 = 59$$

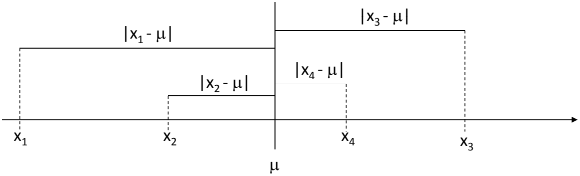
Properties of Range

- Only two values are used in its calculation.
- It is influenced by an extreme value (non-robust).
- It is easy to compute and understand.

Mean Absolute Deviation

- The Mean Absolute Deviation of a set of n numbers

$$\text{MAD} = \frac{|x_1 - \mu| + \ldots + |x_n - \mu|}{n}$$



Example

- **Example:** A sample of four executives received the following bonuses last year ($000): 14.0 15.0 17.0 16.0
- **Problem:** Determine the MAD.
- **Solution:**

$$
\begin{aligned}
\bar{x} &= \frac{14+15+17+16}{4} = \frac{62}{4} = 15.5. \\
\text{MAD} &= \frac{|14-15.5|+|15-15.5|+|17-15.5|+|16-15.5|}{4} \\
&= \frac{4}{4} = 1.
\end{aligned}
$$

Example

- Example: the weight example (weight.csv)
- The R code:

```
>weight <- read.csv("weight.csv")
>sec_01A<-weight$Weight.01A.2013Fall
>mad(sec_01A, center=mean(sec_01A),constant = 1)
 [1]  6.5
```

Properties of MAD

- All values are used in the calculation.
- It is not unduly influenced by large or small values (robust)
- The absolute values are difficult to manipulate.

Variance

- The variance of a set of n numbers as population:

$$\begin{aligned}\text{Var} := \sigma^2 &= \frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{n} \text{ (conceptual formula)}\\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n} \text{ (computational formula).}\end{aligned}$$

- The variance of a set of n numbers as sample:

$$\begin{aligned}S^2 &= \frac{(x_1-\mu)^2+\ldots+(x_n-\mu)^2}{n-1} \text{ (conceptual formula)}\\ &= \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} \text{ (computational formula).}\end{aligned}$$

Standard Deviation

- The Standard Deviation is the square root of variance
- Other notations: σ for population and S for sample.

Example: Conceptual formula:

- **Example**: The hourly wages earned by three students are: \$10, \$11, \$13.
- **Problem:** Find the mean, variance, and Standard Deviation.
- **Solution**: Mean and variance

$$
\begin{aligned}
\mu &= \frac{10+11+13}{3} = \frac{34}{3} \approx 11.3333333333 \\
\sigma^2 &\approx \frac{(10-11.33)^2 + (11-11.33)^2 + (13-11.33)^2}{3} \\
&= \frac{1.769 + 0.1089 + 2.7889}{3} = \frac{4.6668}{3} = 1.5556.
\end{aligned}
$$

- Standard Deviation

$$\sigma \approx 1.247237.$$

Example: Computational formula

- **Example**: The hourly wages earned by three students are: \$10, \$11, \$13.
- **Problem:** Find the variance, and Standard Deviation.
- **Solution**:
  - Variance

$$
\begin{aligned}
\sigma^2 &= \frac{(10^2+11^2+13^2)-\frac{(10+11+13)^2}{3}}{3} \\
&= \frac{390-\frac{1156}{3}}{3} = \frac{390-385.33}{3} = \frac{4.67}{3} = 1.555667.
\end{aligned}
$$

  - Standard Deviation

$$\sigma \approx 1.247665.$$

- If the above is sample, then $\sigma^2 \approx 2.33335$ and $\sigma \approx 1.527531$.

Example

- Example: the weight example (weight.csv)
- The R code:

```
>weight <- read.csv("weight.csv")
>sec_01A<-weight$Weight.01A.2013Fall
# Sample Variance
>var(sec_01A)
[1] 146.3228
# if you want the population variance
>var(sec_01A)*(length(sec_01A)-1)/(length(sec_01A))
[1] 143.9628
# Sample standard deviation
>sd(sec_01A)
  [1] 12.0964
# if you want the population SD
>sd(sec_01A)*sqrt((length(sec_01A)-1)/(length(sec_01A)))
[1] 11.99845
```

Note

- Conceptual formula may have accumulated rounding error.
- Computational formula only has rounding error towards the end!

Properties of Variance/standard deviation

- All values are used in the calculation.
- It is not extremely influenced by outliers (non-robust).
- The units of variance are awkward: the square of the original units. Therefore standard deviation is more natural since it recovers he original units.

Range for grouped data

- The range of a sample of data organized in a frequency distribution is computed by the following formula:

$$\text{Range} = \text{upper limit of the last class - lower limit of the first class}$$

Variance/Standard Deviation for Grouped Data

- The variance of a sample of data organized in a frequency distribution is computed by the following formula:

$$S^2 = \frac{\sum_{i=1}^{k} f_i x_i^2 - \frac{\left(\sum_{i1}^{k} f_i x_i\right)^2}{n}}{n-1}$$

- where f_i is the class frequency and x_i is the class midpoint for Class $i = 1, \ldots, k$.

Example

- **Example**: Recall the weight example from Chapter 2:

class	freq. (f_i)	mid point (x_i)	$f_i x_i$	$f_i x_i^2$
$[130, 140)$	3	135	405	54675
$[140, 150)$	12	145	1740	252300
$[150, 160)$	23	155	3565	552575
$[160, 170)$	14	165	2310	381150
$[170, 180)$	6	175	1050	183750
$[180, 190]$	4	185	740	136900
	62		9,810	1,561,350

- **Solution:** The Variance/Standard Deviation are:

$$\begin{aligned} S^2 &= \frac{1,561,350 - \frac{9{,}810^2}{62}}{62 - 1} \approx 150.0793. \\ S &\approx 12.25069 \end{aligned}$$

- The real sample variance/SD for the raw data is 146.3228/12.0964.

Range for grouped data

- The range of a sample of data organized in a frequency distribution is computed by the following formula:

$$\text{Range} = \text{upper limit of the last class - lower limit of the first class}$$

Layout

Coefficient of Variation (CV)

- The Coefficient of Variation (CV) for a data set defined as the ratio of the standard deviation to the mean:

$$CV = \begin{cases} \frac{\sigma}{\mu} & \text{for population} \\ \frac{S}{\bar{x}} & \text{for sample.} \end{cases}$$

- It is also known as unitized risk or the variation coefficient.
- The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage.
- It shows the extent of variability in relation to mean of the population.
- It is a normalized measure of dispersion of a probability distribution or frequency distribution.
- CV is closely related to some important concepts in other disciplines
  - The reciprocal of CV is related to the Sharpe ratio in Finance
    - William Forsyth Sharpe is the winner of the 1990 Nobel Memorial Prize in Economic Sciences.
  - The reciprocal of CV is one definition of the signal-to-noise ratio in image processing.

Example

- **Example**: Rates of return over the past 6 years for two mutual funds are shown below.

Fund A:	8.3	-6.0	18.9	-5.7	23.6	20
Fund B:	12	-4.8	6.4	10.2	25.3	1.4

- **Problem:** Which fund has higher risk?
- **Solution:** The CV's are given as:

$$
\begin{aligned}
CV_A &= \frac{13.19}{9.85} = 1.34 \\
CV_B &= \frac{10.29}{8.42} = 1.22
\end{aligned}
$$

- There is more variability in Fund A as compared to Fund B. Therefore, Fund A is riskier.
- Alternatively, if we look at the reciprocals of CVs, which are the Sharpe ratios. Then we always prefer the fund with higher Sharpe ratio, which says that for every unit of risk taken, how much return you can obtain.

Properties of Coefficient of Variation

- All values are used in the calculation.
- It is only defined for ratio level of data.
- The actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number.
    - For comparison between data sets with <u>different units</u> or <u>widely different means</u>, one should use the coefficient of variation instead of the standard deviation.

Layout

Coefficient of Skewness

- Skewness is a measure of the extent to which a probability distribution of a real-valued random variable "leans" to one side of the mean. The skewness value can be positive or negative, or even undefined.
- The Coefficient of Skewness for a data set:

$$\text{Skew} = \mathrm{E}\left[\left(\tfrac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

where μ_3 is the third moment about the mean μ, σ is the standard deviation, and E is the expectation operator.

Example

- Example: the weight example (weight.csv)
- The R code:

```
>weight <- read.csv("weight.csv")
>sec_01A<-weight$Weight.01A.2013Fall
# Skewness
>skewness(sec_01A)
[1] 0.4267608
```

Skewness Risk

- Skewness risk is the risk that results when observations are not spread symmetrically around an average value, but instead have a skewed distribution.
- As a result, the mean and the median can be different.
- Skewness risk can arise in any quantitative model that assumes a symmetric distribution (such as the normal distribution) but is applied to skewed data.
- Ignoring skewness risk, by assuming that variables are symmetrically distributed when they are not, will cause any model to understate the risk of variables with high skewness.
- Reference: Mandelbrot, Benoit B., and Hudson, Richard L., The (mis)behaviour of markets : a fractal view of risk, ruin and reward, 2004.

Layout

Coefficient of Excess Kurtosis: Kurt

- Kurtosis (meaning curved, arching) is a measure of the "peakedness" of the probability distribution.
- The Excess Kurtosis for a data set:

$$\text{Kurt} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$$

  where μ_4 is the fourth moment about the mean and σ is the standard deviation.
- The Coefficient of excess kurtosis provides a comparison of the shape of a given distribution to that of the normal distribution.
  - **Negative excess kurtosis**: platykurtic/sub-Gaussian: A low kurtosis implies a more rounded peak and shorter/thinner tails: such as uniform, Bernoulli distributions.
  - **Positive excess kurtosis**: leptokurtic/super-Gaussian: A high kurtosis implies a sharper peak and longer/fatter tails: such as the Student's t-distribution, exponential, Poisson, and the logistic distribution.
  - **Zero excess kurtosis** are called mesokurtic, or mesokurtotic: such as Normal distribution

Example

- Example: the weight example (weight.csv)
- The R code:

```
>weight <- read.csv("weight.csv")
>sec_01A<-weight$Weight.01A.2013Fall
# kurtosis
>kurtosis(sec_01A)
[1] -0.06660501
```

Kurtosis Risk

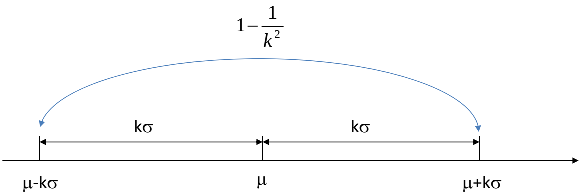
- Kurtosis risk is the risk that results when a statistical model assumes the normal distribution, but is applied to observations that do not cluster as much near the average but rather have more of a tendency to populate the extremes either far above or far below the average.
- Kurtosis risk is commonly referred to as "fat tail" risk. The "fat tail" metaphor explicitly describes the situation of having more observations at either extreme than the tails of the normal distribution would suggest; therefore, the tails are "fatter".
- Ignoring kurtosis risk will cause any model to understate the risk of variables with high kurtosis.

Layout

Chebyshev's Theorem

- Given a data set with mean μ and standard deviation σ, for any constant k, the minimum proportion of the values that lie within k standard deviations of the mean is at least

$$1 - \frac{1}{k^2}$$



Example

- Given a data set of five values: 14.0 15.0 17.0 16.0 15.0
- Calculate mean and standard deviation: $\mu = 15.4$ and $\sigma \approx 1.0198$.
- For $k = 1.1$: there should be at least

$$1 - \frac{1}{1.1^2} \approx 17\%$$

within the interval

$$\begin{aligned}[\mu - k\sigma, \mu + k\sigma] &= [15.4 - 1.1(1.0198), 15.4 + 1.1(1.0198)] \\ &= [14.28, 16.52].\end{aligned}$$

- Let us count the real percentage: Among the five, three of which (namely 15.0, 15.0, 16.0) are in the above interval. Therefore there are three out of five, i.e., 60 percent, which is greater than 17 percent. So the theorem is correct, though not very accurate!

Example

- **Example:** The average of the marks of 65 students in a test was 70.4. The standard deviation was 2.4
- **Problem 1.** About what percent of students' marks were between 65.6 and 75.2?
    - **Solution.** Based on Chebychev's Theorem, 75%, since $65.6 = 70.4 - (2)(2.4)$, and $75.2 = 70.4 + (2)(2.4)$.
- **Problem 2.** About what percent of students' marks were between 63.2 and 77.6?
    - **Solution.** Based on Chebychev's Theorem, 88.9%, since $63.2 = 70.4 - (3)(2.4)$, and $77.6 = 70.4 + (3)(2.4)$.

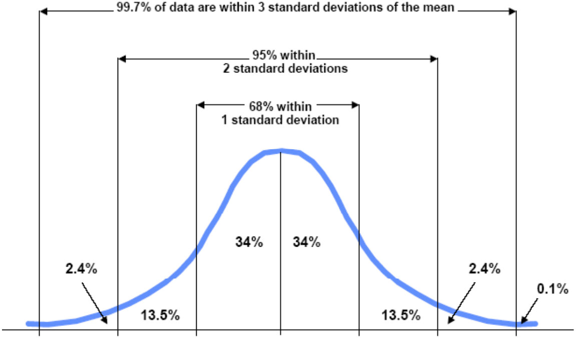
Example

- **Example:** The average of the marks of 65 students in a test was 70.4. The standard deviation was 2.4
- **Problem 1.** About what percent of students' marks were between 65.6 and 75.2?
  - **Solution.** Based on Chebychev's Theorem, 75%, since $65.6 = 70.4 - (2)(2.4)$, and $75.2 = 70.4 + (2)(2.4)$.
- **Problem 2.** About what percent of students' marks were between 63.2 and 77.6?
  - **Solution.** Based on Chebychev's Theorem, 88.9%, since $63.2 = 70.4 - (3)(2.4)$, and $77.6 = 70.4 + (3)(2.4)$.

The Empirical rule

- We can obtain better estimation if we know more...
- For any symmetrical, bell-shaped distribution (i.e., Normal distribution):
  - About 68% of the observations will lie within 1 standard deviation of the mean;
  - About 95% of the observations will lie within 2 standard deviation of the mean;
  - Virtually all the observations (99.7%) will be within 3 standard deviation of the mean.

The Empirical rule



Comparing Chebyshev's Theorem with The Empirical rule

k	Chebyshev's Theorem	The Empirical rule
1	0%	68%
2	75%	95%
3	88.9%	99.7%

An Approximation for Range

- Based on the last claim in the empirical rule, we can approximate the range using the following formula

$$\text{Range} \approx 6\sigma.$$

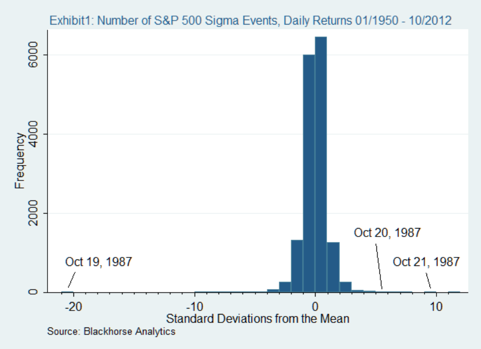
- Or more often we can use the above to estimate standard deviation from range (such as in Project Management, or Quality Management):

$$\sigma \approx \frac{\text{Range}}{6}.$$

- **Example:** If the range for a normally distributed data is 60, given a approximation of the standard deviation.
  - Solution: $\sigma \approx 60/6 = 10$.

Why you should check the normal assumption first?

- Model risk is everywhere!
- Black Swans are more than you think!

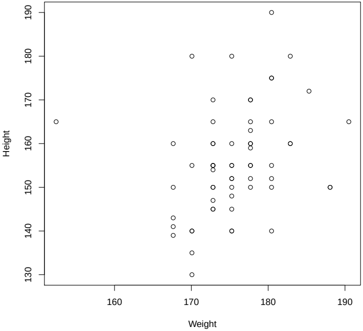


Layout

Coefficient of Correlation: Scatterplot for the weight vs height collected form the guessing game in the first class

- Coefficient of Correlation is a measure of the linear correlation (dependence) between two sets of data $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$:
- A scatter plot pairs up values of two quantitative variables in a data set and display them as geometric points inside a Cartesian diagram.

Scatterplot: Example



- The R code:

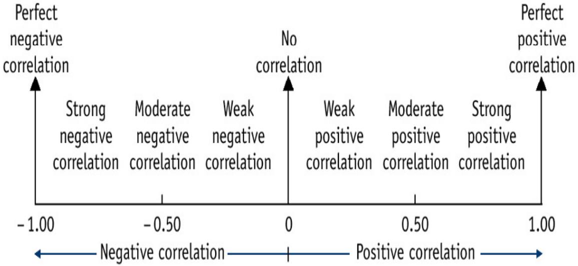
```
>weight_height <- read.csv("weight_height_01A.csv")
>plot(weight_height$weight_01A, weight_height$height_01A,
 xlab="Weight", ylab="Height")
```

Coefficient of Correlation: r

- Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations:

$$
\begin{aligned}
r_{xy} &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \\
&= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i - \left(n\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i - \left(n\sum_{i=1}^{n} y_i\right)^2}}
\end{aligned}
$$

Coefficient of Correlation: Interpretation



Coefficient of Correlation: Example

- For the the weight vs height collected form the guessing game in the first class.
- The Coefficient of Correlation is

$$r \approx 0.2856416$$

- Indicated weak positive linear correlation.
- The R code:

```
> weight_height <- read.csv("weight_height_01A.csv")
> cor(weight_height$weight_01A, weight_height$height_01A,
 use = "everything", method = "pearson")
 [1] 0.2856416
```

Layout

Case 1: Wisdom of the Crowd

The aggregation of information in groups results in decisions that are often better than could have been made by any single member of the group.

- We collect some data from the guessing game in the first class. The impression is that the crowd can be very wise.
- The question is when the crowd is wise? We given a mathematical explanation based on the concepts of mean and standard deviation.
- For more explanations, please refer to: Surowiecki, James (2005). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.

Case 1: Wisdom of the Crowd: Mathematical explanation

- Let x_i $(i = 1, \ldots, n)$ be the individual predictions and let θ be the true value.
- The crowd's overall prediction:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- The crowd's (squared) error:

$$\begin{aligned}(\mu - \theta)^2 = \left(\frac{\sum_{i=1}^{n} x_i}{n} - \theta \right)^2 &= \frac{\sum_{i=1}^{n} (x_i - \theta)^2}{n} - \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \\ &= \text{Bias} - \sigma^2 = \text{Bias - Variance.}\end{aligned}$$

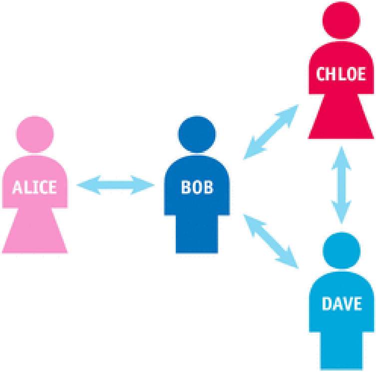
- Diversity of opinion: Each person should have private information, independent of each other.

Case 2: Friendship paradox: Why are your friends more popular than you

On average, most people have fewer friends than their friends have. Scott L. Feld (1991).

- References: Feld, Scott L. (1991), "Why your friends have more friends than you do", American Journal of Sociology 96 (6): 1464-147.

Case 2: Friendship paradox: An example



- References: [hyphen]http://www.economist.com/blogs/economist-explains/2013/04/economist-explains-why-friends-more-popular-paradox.

Case 2: Friendship paradox: An example

Person	degree
Alice	1
Bob	3
Chole	2
Dave	2

- Individual's average number of friends is

$$\mu = \frac{8}{4} = 2.$$

- Individual's friends' average number of friends is:

$$\frac{1^2 + 2^2 + 2^2 + 3^2}{1 + 2 + 2 + 3} = \frac{18}{8} = 2.25.$$

Case 2: Friendship paradox: Mathematical explanation

- Let x_i $(i = 1, \ldots, n)$ be the degrees of a network.
- Individual's average number of friends is

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Individual's friends' average number of friends is:

$$\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} = \mu + \frac{\sigma^2}{\mu}.$$

Case 2: Friendship paradox: Mathematical explanation

- Let x_i $(i = 1, \ldots, n)$ be the degrees of a network.
- Individual's average number of friends is

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Individual's friends' average number of friends is:

$$\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} = \mu + \frac{\sigma^2}{\mu}.$$

Case 2: A video by Nicholas Christakis: How social networks predict epidemics?

- Here is the youtube link