Statistical Data Analysis

Manuella Lech Cantuaria
Victoria Blanes-Vidal
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science

Fall 2022

1

2

2

TEACHERS

Manuella Lech Cantuaria
Assistant Professor, PhD
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science
mlca@mmmi.sdu.dk

Victoria Blanes-Vidal
Associate Professor, PhD
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science
vbv@mmmi.sdu.dk

3

3

TEACHERS

Manuella Lech Cantuaria
Assistant Professor, PhD
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science
mlca@mmmi.sdu.dk

- Engineer with a Ph.D. degree in data science and environmental epidemiology.
- Research focus on data science applied to different health-related contexts (e.g. hearing impairment, dementia, prostate cancer, health effects of air and noise pollution)
- Work mostly with application of epidemiological/statistical modelling and machine learning techniques using e.g. R and SAS.

4

4

ABOUT THE COURSE

Content

- A first introduction to probability and statistics, this course provides background to understand and produce statistical analysis, including descriptive and inferential statistics. Applicability and limitations of these methods will be illustrated with different data sets using statistical software (R).

5

5

ABOUT THE COURSE

- **Learning objectives - Knowledge**
  - explain relevant data types and their representation for statistical analysis
  - explain probabilities and random variables
  - explain distributions of random variables
  - explain inference and hypothesis testing
  - explain how data may be collected from experiments involving randomness
- **Learning objectives - Skills**
  - choose an appropriate experimental design in respect to a given task
  - perform statistical analyzes on data collected
  - use a statistical tool for analysis and visualization of data
- **Learning objectives - Competences**
  - use statistical methods and tools to interpret experimental data

6

6

RECOMMENDED LITERATURE

- OpenIntro Statistics (David Diez, Mine Cetinkaya-Rundel, Christopher Barr).
  - It can be downloaded for free here: https://www.openintro.org/book/os/
- Other recommended books:
  - Applied statistics and probability for engineers / Douglas C. Montgomery, George C. Runger —3rd ed. ISBN 0-471-20454-4
  - A Handbook of Statistical Analyses Using R / Brian S. Everitt, Torsten Hothorn , ISBN 1420079336

7

7

COURSE ORGANIZATION

- Thursdays 12-14: Main lecture (Manuella or Victoria)
  - Content will be presented
  - There will be small exercises in between: we recommend you to discuss your results with the person sitting next to you!

8

8

COURSE ORGANIZATION

- Thursdays 12-14: Main lecture (Manuella or Victoria)
- Thursdays 14-16: Exercise time (independent work on exercises' list with help of instructors). There will be one instructor responsible per class.
- Exercises' list is available in ItsLearning, at the corresponding plan.
  - Instructors are:
    - Henrik Dyrberg Egemose (U90)
    - Sofie Ørnfeldt Nedergaard (U171)
    - Lasse Schier Christiansen (U172)
    - Stine Lee Jakobsen (U176)
- You are able to know your classroom by looking at your team and the corresponding classroom in ItsLearning

9

9

LECTURE PLANNING

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics – Part I	MLC
2	36	8/sep	Descriptive statistics – Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics) **Point-giving activity (in class)**	MLC+VBV
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Regression analysis	VBV
11	47	24/Nov	Notions of experimental design and questions **Point-giving activity (in class)**	VBV+MLC
12	48	1/Dec	Multiple regression	MLC

R

10

10

EXAM AND POINTGIVING ACTIVITIES

- Multiple choice exam at the end of the course
  - Questions will involve concepts' understanding and calculations for problem solving
  - R can be used (not Excel).
  - 120 minutes
  - Probably beginning of January – dates will come later
- Two Pointgiving activities during the semester
  - Pointgiving activities serve as a **"bonus"** to your grade (10%)
  - Carried out in the classroom and individually
  - Dates: 13th of October and 24th of November (there is no recurrence of point-giving activities)
  - You can consult any material you like (**it will NOT be like that in the final exam**)
- Reexam in February

11

11

Chapter 1: Descriptive Statistics

Manuella Lech Cantuaria
Victoria Blanes-Vidal
The Maersk Mc-Kinney Moller Institute
Applied AI and Data Science

12

Chapter 1 Overview

1.1. Statistics: Descriptive and Inferential
1.2. Variables and Types of Data
1.3. Measures of:
    Central Tendency (Location)
    Variation (Dispersion)
    Position
1.4. Data representation: frequency distributions, histograms and other graphs
1.5. Shapes of frequency distributions: Skewness and kurtosis

OpenIntro book:
- Chapter 1 (Pag 22-24)
- Chapter 2

13

13

Chapter 1 Overview

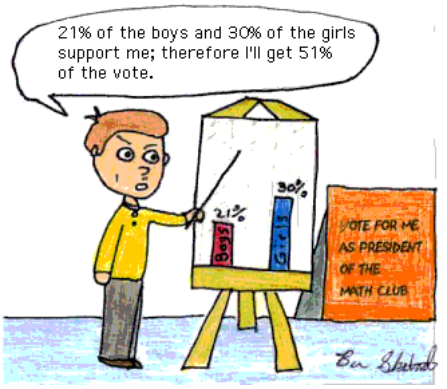
1.1. Statistics: Descriptive and Inferential
1.2. Variables and Types of Data
1.3. Measures of:
    Central Tendency (Location)
    Variation (Dispersion)
    Position
1.4. Data representation: frequency distributions, histograms and other graphs
1.5. Shapes of frequency distributions: Skewness and kurtosis

14

14

1-1 Statistics

- **Statistics** is the science of conducting studies to
  collect,
  organize,
  summarize,
  analyze, and
  draw conclusions from data.



15

15

1-1 Statistics

- Study to better understand characteristics of software developers working in Odense

16

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

16

1-1 Statistics

- A **variable** is a characteristic or attribute that can assume different values.
- The values that a variable assumes are called **data**.

17

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

17

1-1 Statistics

Study to better understand characteristics of software developers working in Odense

- A **population** consists of all subjects (human or otherwise) that we want to study.
  - All software developers working in Odense
- A **sample** is a subset of the population.
  - 15 software developers that were randomly selected for the study

18

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

18

1-1 Statistics

Descriptive statistics

Inferential statistics

19

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

19

Descriptive statistics

Study to better understand characteristics of software developers working in Odense

Used to describe the **sample** data

What is the mean age of software developers?

What is the proportion of developers who prefer Python?

Tables

- Tables are extremely useful to summarize data upon conclusions are based.
- It uses a minimum of space to communicate a large amount of information.

Graphs

- More visual than tables
- Often preferred to show variable's trends, better understand data distribution and variability and compare groups.

20

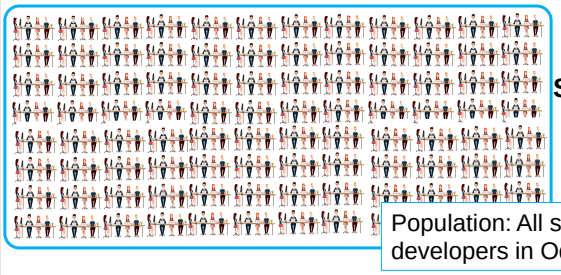
Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

20

Inferential statistics

Study to better understand characteristics of software developers working in Odense

Uses the **sample** data to draw conclusions about a **population**



Population: All software developers in Odense

SAMPLING

STUDY SAMPLE

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

INFERENCE

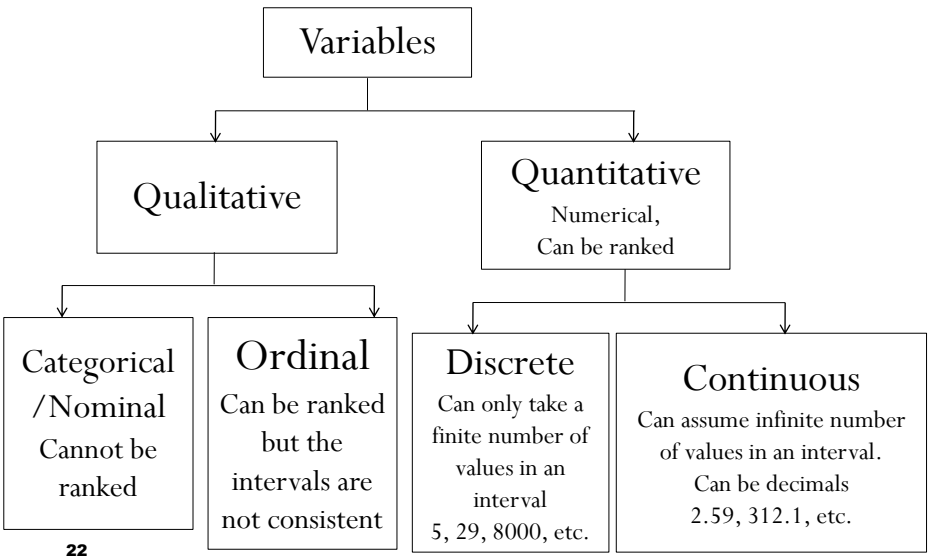
Inferential statistics are statistical techniques that **allow us to use samples to draw conclusions about the populations** from which the samples were taken.

21

21

1.2 Types of Variables and Data

- Variables can be classified as:

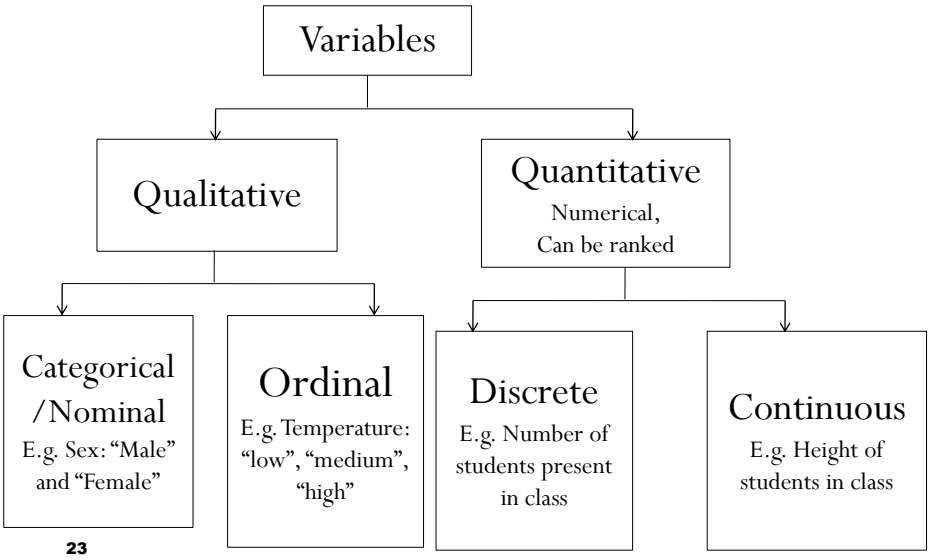


22

22

1.2 Types of Variables and Data

- Variables can be classified as:



23

23

Exercise 1: Determine the type of variable

Variable	Nominal	Ordinal	Discrete	Continuos
Hair color "blond", "red", "black"				
Temperature (in Celsius)				
School grade "1st", "2nd", "3rd", …				
Level of satisfaction "very", "intermediate", "low"				
Height (in cm)				
Age (in years)				
Inhabitants				

24

1-3 Measures of Central Tendency (location), Variation (dispersion) and Position

- The data distribution can be mainly described by three different characteristics:
  - Measures of location
  - Measures of dispersion
  - Measures of position

25

Measures of Central Tendency (Location)

- Mean
- Median
- Mode
- Midrange

26

26

Mean

- The **mean** is the division of the sum of the values and the total number of values.
- The symbol $\bar{X}$ is used for sample mean:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\sum X}{n}$$

- For a population, the Greek letter μ (mu) is used for the population mean:

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\sum X}{N}$$

Note: General Rounding Rule

The basic rounding rule is that rounding should not be done until the final answer is calculated.

27

27

Mean

- The **mean** is the division of the sum of the values and the total number of values.
- The symbol $\bar{X}$ is used for sample mean:

$$\bar{X} = \frac{32 + 41 + 23 + \cdots + 25}{15} = 32.33$$

28

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

28

Median

- The **median** is the midpoint of the data array.
- How to calculate the median:
  - Sort in ascending order.
  - Select the middle value.
- The median will be one of the data values if there is an odd number of values.
- The median will be the average of two data values if there is an even number of values.

29

29

Median

- The **median** is the midpoint of the data array.

23, 23, 25, 25, 25, 29, 29, 30, 32, 32, 34, 34, 41, 47, 56

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

30

30

Mode

- The **mode** is the value that occurs most often in a data set.
- It is sometimes said to be the most typical case.
- There may be no mode, one mode (unimodal), two modes (bimodal), or many modes (multimodal).

31

31

Mode

- The **mode** is the value that occurs most often in a data set.

23, 23, 25, 25, 25, 29, 29, 30, 32, 32, 34, 34, 41, 47, 56

32

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

32

Midrange

- The **midrange** is the average of the lowest and highest values in a data set.

$$MR = \frac{Lowest + Highest}{2}$$

33

Midrange

- The **midrange** is the average of the lowest and highest values in a data set.

23, 23, 25, 25, 25, 29, 29, 30, 32, 32, 34, 34, 41, 47, 56

$$MR = \frac{23 + 56}{2} = 39.5$$

34

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

Let's practice!

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11	Jan. 25 – 25
Jan. 19 – 11	Jan. 26 – 33
Jan. 20 – 25	Jan. 27 – 22
Jan. 21 – 29	Jan. 28 – 18
Jan. 22 – 27	Jan. 29 – 19
Jan. 23 – 14	Jan. 30 – 30
Jan. 24 – 11	Jan. 31 – 27

- For these 14 values, calculate all four measures of central tendency - the **mean**, **median**, **mode**, and **midrange**

35

Mean

- **Mean** –Most commonly used measure of central tendency
- **Procedures**
- (1) **Sum** all the values in the data set
- (2) **Divide** the sum by the number of values in the data set
- Watch for **outliers**

An **outlier** is an observation point that is very distant from other observations

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

36

Mean

- (1) **Sum** all the values in the data set

  → 11 + 11 + 11 + 14 + 18 + 19 + 22 + 25 + 25 + 27 + 27 + 29 + 30 + 33 = 302

- (2) **Divide** the sum by the number of values in the data set

  → **Mean** = 302/14 = 21.57

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

37

Median

- **Median** - **1/2** of the values **are above** it & **1/2 below**
- (1) **Sort** the data in **ascending** order
- (2) **Find** the value with an **equal number** of values above and below it
- (3) **Odd** number of observations → [(n-1)/2]+1 value from the lowest
- (4) **Even** number of observations → average (n/2) and [(n/2)+1] values

38

Median

- (1) **Sort** the data in ascending order:

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

- (2) **Find** the value with an equal number of values above and below it

  Even number of observations → average the (n/2) and [(n/2)+1] values

  → (14/2) = 7; [(14/2)+1] = 8

  → (22+25)/2 = **23.5** (°F)

39

Mode

- **Mode** – This is the most frequently occurring value in the distribution
- (1) **Sort** the data in **ascending** order
- (2) **Count** the **instances** of each value
- (3) **Find** the value that has the **most** occurrences
- If more than one value occurs an **equal number** of times and these exceed all other counts, we have **multiple** modes
- Use the mode for **multi-modal** data

40

Mode

- (1) **Sort** the data in ascending order:

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

- (2) **Count** the instances of each value:

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

  3x 1x 1x 1x 1x 2x 2x 1x 1x 1x

- (3) **Find** the value that has the most occurrences

  → **mode** = 11 (°F)

41

Midrange

- (1) **Sort** the data in ascending order:
- (2) **Select** the lowest and highest values:
- (3) **Find** the mean of those two values

42

Midrange

- (1) **Sort** the data in ascending order:

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

- (2) **Select** the lowest and highest values:

  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33

  Lowest = 11; Highest = 33

- (3) **Find** the mean of those two values

  → **midrange** = (11 + 33) / 2 = 22 (°F)

43

Discuss in pairs:

1. Which of the measures of central tendency is sensitive to extreme scores (values) on the higher or lower end of a distribution of data?
2. Can the median be negative?
3. Which measures of central tendency can have more than one value?

44

44

1-3 Measures of Central Tendency (location), Variation (dispersion) and Position

- The data distribution can be mainly described by three different characteristics:
  - Measures of location
  - Measures of dispersion
  - Measures of position

45

Measures of variation (dispertion)

Study to better understand characteristics of software developers working in Odense

Software developer	Sex	Age	Preferred language
1	M	32	Python
2	M	41	HTML
3	F	23	SQL
4	M	56	Python
5	F	32	Python
6	M	34	HTML
7	M	47	SQL
8	F	25	Python
9	F	29	JavaScript
10	F	29	Python
11	M	30	JaveScript
12	M	23	Python
13	F	34	Python
14	F	25	HTML
15	M	25	SQL

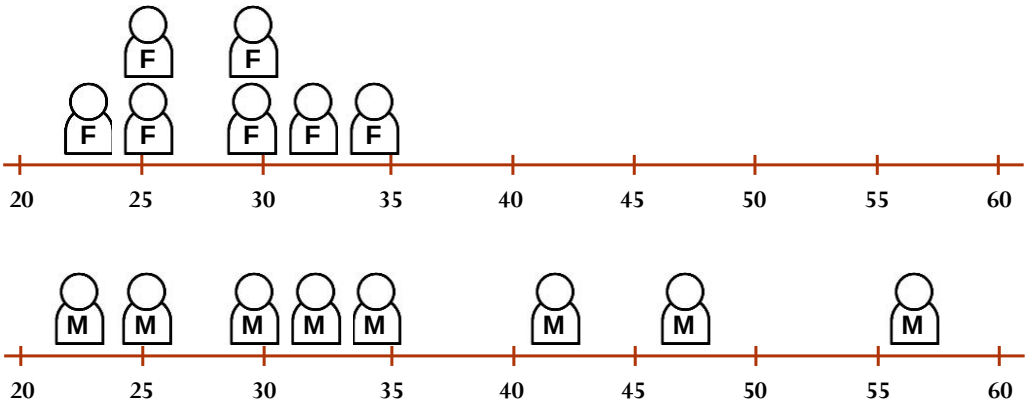
Age of male and female software developers

Male	Female
32	23
41	32
56	25
34	29
47	29
30	34
23	25
25	

46

Measures of variation (dispertion)

Study to better understand characteristics of software developers working in Odense



47

47

Measures of Variation (Dispersion)

Measures of dispersion are concerned with the distribution of values around the mean in data.

How Can We Measure **Variability**?

- Range
- Variance
- Standard Deviation

48

48

Range

- The **range** is the difference between the highest and lowest values in a data set.

$$R = Highest - Lowest$$

49

Variance & Standard Deviation

- The standard deviation and variance are measures of how spread out your data are.
- The **variance** is the average of the squares of the distance each value is from the mean.
- The **standard deviation** is the square root of the variance.

50

Measures of Variation: Variance & Standard Deviation (Sample Theoretical Model)

- The **sample variance** is

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

- The **sample standard deviation** is

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

The value of variance calculated from sample data is higher than the value that could have been found out by using population data. The logic of doing that is to compensate our lack of information about the population data.

51

51

An Example Data Set

- **Daily low temperatures** recorded in a town (01/18-01/31, 2005, °F)

Jan. 18 – 11	Jan. 25 – 25
Jan. 19 – 11	Jan. 26 – 33
Jan. 20 – 25	Jan. 27 – 22
Jan. 21 – 29	Jan. 28 – 18
Jan. 22 – 27	Jan. 29 – 19
Jan. 23 – 14	Jan. 30 – 30
Jan. 24 – 11	Jan. 31 – 27

- For these 14 values, we will calculate all measures of **dispersion**

52

Range

- **Range** – The difference between the largest and the smallest values
- (1) **Sort** the data in ascending order
  → 11, 11, 11, 14, 18, 19, 22, 25, 25, 27, 27, 29, 30, 33
- (2) **Find** the largest value
  → max = 33
- (3) **Find** the smallest value
  → min = 11
- (4) **Calculate** the **range**
  → range = 33 – 11 = 22

54

Variance

- (1) **Calculate** the **mean**
  → $\bar{x}$
- (2) **Calculate** the **deviation** for each value
  → $x_i - \bar{x}$
- (3) **Square** each of the deviations
  → $(x_i - \bar{x})^2$
- (4) **Sum** the **squared** deviations
  → $\sum (x_i - \bar{x})^2$
- (5) **Divide** the **sum of squares** by (n-1) for a sample
  → $\sum (x_i - \bar{x})^2 / (n-1)$

56

Variance

- (1) **Calculate** the **mean**

  → $\bar{x} = 21.57$

- (2) **Calculate** the **deviation** for each value

  → $x_i - \bar{x}$

Jan. 18	(11 – 21.6) = -10.57	Jan. 25	(25 – 21.6) = 3.43
Jan. 19	(11 – 21.6) = -10.57	Jan. 26	(33 – 21.6) = 11.43
Jan. 20	(25 – 21.6) = 3.43	Jan. 27	(22 – 21.6) = 0.43
Jan. 21	(29 – 21.6) = 7.43	Jan. 28	(18 – 21.6) = -3.57
Jan. 22	(27 – 21.6) = 5.43	Jan. 29	(19 – 21.6) = -2.57
Jan. 23	(14 – 21.6) = -7.57	Jan. 30	(30 – 21.6) = 8.42
Jan. 24	(11 – 21.6) = -10.57	Jan. 31	(27 – 21.6) = 5.42

57

Variance

- (3) **Square** each of the **deviations**

  → $(x_i - \bar{x})^2$

Jan. 18	(-10.57)^2 = 111.76	Jan. 25	(3.43)^2 = 11.76
Jan. 19	(-10.57)^2 = 111.76	Jan. 26	(11.43)^2 = 130.61
Jan. 20	(3.43)^2 = 11.76	Jan. 27	(0.43)^2 = 0.18
Jan. 21	(7.43)^2 = 55.18	Jan. 28	(-3.57)^2 = 12.76
Jan. 22	(5.43)^2 = 29.47	Jan. 29	(-2.57)^2 = 6.61
Jan. 23	(7.57)^2 = 57.33	Jan. 30	(8.43)^2 = 71.04
Jan. 24	(-10.57)^2 = 111.76	Jan. 31	(5.43)^2 = 29.27

- (4) **Sum** the **squared** deviations

  → $\sum (x_i - \bar{x})^2 = 751.43$

58

Variance

- (5) **Divide** the **sum of squares** by (n-1) for a sample

  →

$$\sum (x_i - \bar{x})^2 / (n-1)$$

= 751.43 / (14-1) = 57.8

- The **variance** of the Tmin (F) data set is 57.8

59

Standard Deviation

- (1) **Calculate** the **mean**

  → $\bar{x}$

- (2) **Calculate** the **deviation** for each value

  → $x_i - \bar{x}$

- (3) **Square** each of the **deviations**

  → $(x_i - \bar{x})^2$

- (4) **Sum** the **squared** deviations

  → $\sum (x_i - \bar{x})^2$

- (5) **Divide** the **sum of squares** by (n-1) for a sample

  → $\sum (x_i - \bar{x})^2 / (n-1)$

- (6) **Take the square root** of the resulting **variance**

  → $\sqrt{\sum (x_i - \bar{x})^2 / (n-1)}$

61

Standard Deviation

- (1) – (5)

  → $s^2 = 57.8$

- (6) **Take the square root** of the **variance**

  → $\sqrt{57.8} = 7.6$

- The **standard deviation** (*s*) of the Tmin data is 7.6 (°F)

62

Questions?

Now it is time for you to practice what you learned at the exercises' class!

Rooms: U90, U171, U172, U176

63

63