Descriptive Statistics

Week 08 Day 01

DS 3000 – Foundations of Data Science

1

Reminders

FP2: Datasets

Due yesterday

Schedule Updated

To reflect changes in HW schedule

2

Outline

Statistical Preliminaries

Measures of Central Tendency

Measures of Variability

Normal Distribution

3

Data Science

The interdisciplinary study and practice of computationally extracting meaningful insights from data

Three components:

Exploration ➔ identifying patterns in data (messing around)

Prediction ➔ making informed guesses

Inference ➔ quantifying our degree of certainty

4

Data Scientist (n.):

A data wizard who extracts meaningful insights from data using computation and statistics.

5

Statistical Preliminaries

6

Why care about Statistics?

Uses of statistics (statistical procedures)

- Organize and summarize information
- Determine exactly what conclusions are justified based on the results that were obtained

Goals of statistical procedures

- Accurate and meaningful interpretation
- Provide standardized evaluation procedures

7

Statistical Terminology

Variable

Characteristic or condition that changes or has different values for different individuals

Data (plural)

Measurements or observations of a variable

Data set

A collection of measurements or observations

8

Statistical Terminology

Population

The large group of interest from which a subset is selected

Sample

The small set of data points/individuals selected from the population

Goal of data analysis

To generalize from the sample data to the larger population

9

Statistical Terminology

Parameter

A value, usually a numerical value, that describes a **population**

Derived from measurements of the individuals in the population

Statistic

A value, usually a numerical value, that describes a **sample**

Derived from measurements of the individuals in the sample

10

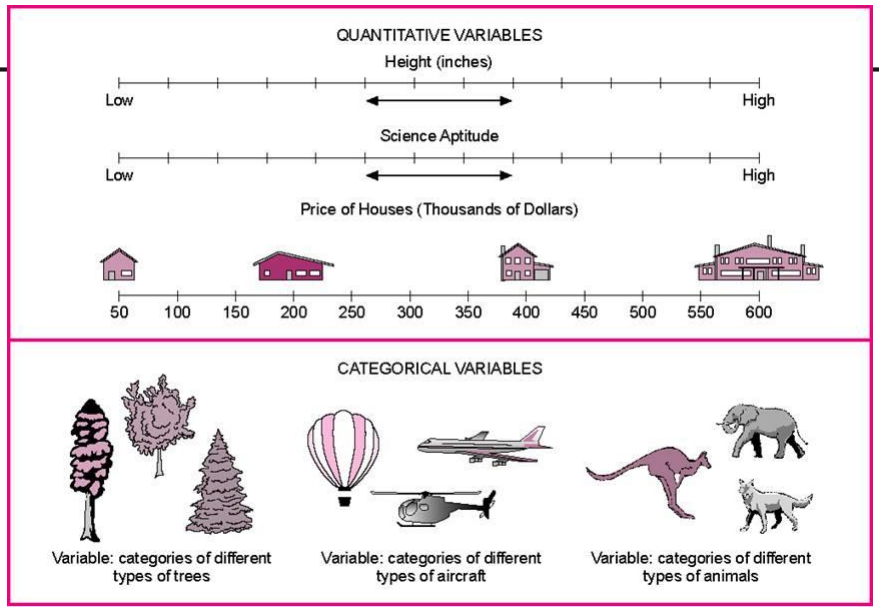
Types of Variables

Variables are classified as either quantitative or categorical

A **quantitative** variable describes differences in quantity or numeric terms (i.e., height).

A **categorical** variable does not vary in degree, amount, or quantity, but are qualitatively different (i.e., gender).

11



12

Scales of Measurement

13

Scales of Measurement

Used to categorize variables

Also referred to as Levels of Measurement

Four Scales of Measurement

Nominal, Ordinal, Interval and Ratio

14

The Nominal Scale

Involves placing data into categories without any order or structure

Represents qualitative differences in the variable measured

Qualitative comparisons are possible, but no quantitative comparisons

Examples: major, gender, occupation, and OS user

Windows vs. Mac users

Two individuals are different but is Windows "more than" Mac? – No

15

The Ordinal Scale

Represents differences in a series of ranks

Categories form an ordered sequence

the numbers have a meaningful order

We can determine the direction of the difference, but not the magnitude of the difference between two individuals

Examples: socioeconomic class (lower, middle, upper), letter grade

16

The Interval Scale

Organized sequentially and all categories are the same size

Consist of a series of equal intervals like the inches on a ruler

We can determine the direction and magnitude of the difference

The zero point is arbitrary; no absolute zero

Examples: Temperature in Fahrenheit or Celsius (no such thing as “no temperature”), score on a Likert scale (self-reported ratings)

17

The Ratio Scale

Organized sequentially and all categories are the same size

Consist of a series of equal intervals like the inches on a ruler

We can determine the direction and magnitude of the difference

Has an absolute zero point, a meaningful point representing a complete absence of a variable

Examples: Height, weight, and time (time to task completion)

18

Scales of Measurement: Summary

SCALES OF MEASUREMENT				
	NOMINAL	**ORDINAL**	**INTERVAL**	**RATIO**
Examples	Ethnicity Religion Sex	Class rank Letter grade	Temperature (Fahrenheit and Celsius) Many psychological tests	Weight Height Time
Properties	Identity	Identity Magnitude	Identity Magnitude Equal unit size	Identity Magnitude Equal unit size Absolute zero
Mathematical Operations	Determine whether $=$ or $\neq$	Determine whether $=$ or $\neq$ Determine whether $<$ or $>$	Determine whether $=$ or $\neq$ Determine whether $<$ or $>$ Add Subtract	Determine whether $=$ or $\neq$ Determine whether $<$ or $>$ Add Subtract Multiply Divide

19

The Role of Statistics in Data Science

Descriptive statistics

Methods used to organize, summarize, and simplify the data

Familiar examples: Tables, graphs, average values

Inferential statistics

Methods that use the results obtained from samples to help make generalizations about populations

Used to infer causality

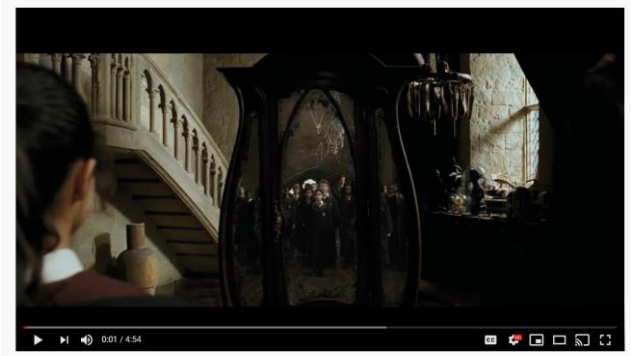
Common terms: “margin of error”, “statistically significant”

20

Descriptive Statistics

21

DADA – Dealing with a Boggart



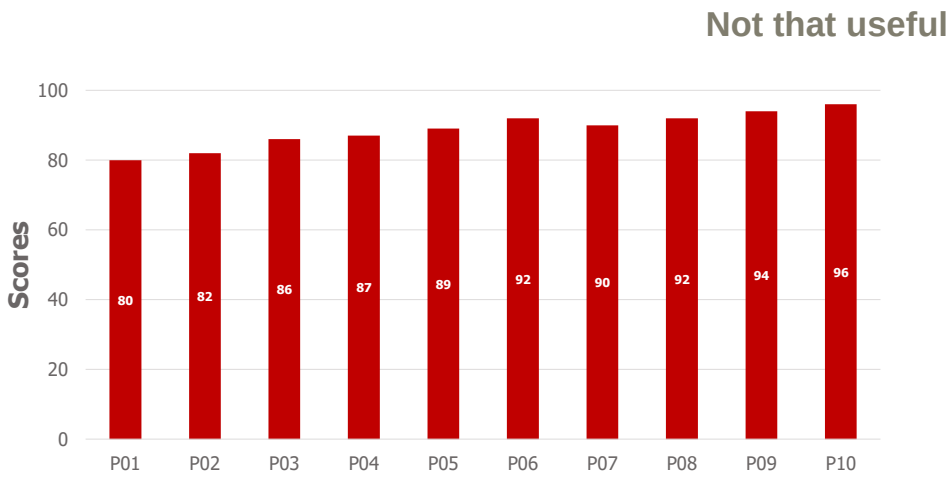
Boggart | Harry Potter and the Prisoner of Azkaban

DADA Scores:

87, 82, 89, 90, 94, 92, 86, 92, 80, 96

22

Reporting Results: Individual Scores



23

Distribution of DADA Scores

A distribution of scores is obtained when you take the scores on one variable and arrange them in order from lowest to highest

DADA Scores:

87, 82, 89, 90, 94, 92, 86, 92, 80, 96

The distribution of DADA scores is:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

We want to know about the characteristics of this distribution

24

Measures of Central Tendency

25

Measures of Central Tendency

Central tendency is a statistical measure that tells us about the characteristics of a distribution

Identifies a single score that defines the center of a distribution

Goal: to identify the value that is most typical or most representative of the entire group

Three measures of central tendency:

Mean, Median, and Mode

26

The Mean

Arithmetic average of a distribution of scores

Provides a single, simple number that gives a rough summary of the distribution

Calculated as the sum of all the scores divided by the number of scores in the data

Sample Mean: $$M = \frac{\sum X}{n}$$

27

The Mean: Example

The distribution of DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

M = (80 + 82 + 86 + 87 +89 +90 +92 +92 +94 +96)/10

M = 88.8

28

The Mean: Three Definitions

Sum of the scores divided by the number of scores in the data

The amount each individual receives when the total is divided equally among all the individuals in the distribution

The balance point for the distribution

29

The Median

The median is the midpoint of the scores in a distribution when they are listed in order from **smallest** to **largest**

The median divides the scores into two groups of equal size

Often used for data sets in which the mean does not provide a good representative value

- In a distribution with a few extreme scores, the median provides a better measure of central tendency than the mean

30

Finding The Median

Put scores in order from smallest to largest

Identify the "middle" score to find median

If there are an **odd** number of scores, the median is the middle score

If there are an **even** number of scores in the distribution, the median is the average of the two scores in the middle of the distribution

31

The Median: Example (Odd Scores)

If we had **nine** DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94

There are 9 scores in the distribution

So identify the middle number

The middle number is 89, so

Median = 89

32

The Median: Example (Even Scores)

The distribution of DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

There are 10 scores in the distribution

So average the two scores in the middle

Median = (89 + 90)/2

Median = 89.5

33

The Mode

The score or category that has the greatest frequency in the distribution

Corresponds to an actual score in the data

It is possible to have more than one mode

34

The Mode: Example

The distribution of DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

Look for the score that occurs most frequently

The mode is: 92

35

Mean, Median, and Mode

Mean is the balance point of a distribution
- Defined by distances
- Often is not the midpoint of the scores

Median is the midpoint of a distribution
- Defined by number of scores
- Often is not the balance point of the scores

Most informative when used with
- Mean ➔ Interval or ratio scale
- Median ➔ Ordinal scale
- Mode ➔ Nominal scale

36

37

Measures of Variability

38

Measures of Variability

Variability can be defined several ways

- A quantitative distance measure based on the differences between scores
- Describes distance of the spread of scores or distance of a score from the mean

Purposes of Measure of Variability

- Describe the distribution
- Measure how well an individual score represents the distribution

Three Measures of Variability

- Range, Variance, and Standard Deviation

39

The Range

The distance covered by the scores in a distribution

- From smallest value to highest value

Range = Max Value – Min Value

Based on two scores, not all the data

- An imprecise, unreliable measure of variability

40

The Range: Example

The distribution of DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

Range = 96 – 80 = 16

41

The Variance

Statistical average of the amount of dispersion in a distribution of scores

Measures the average **squared** distance from the mean

Average deviation of the scores from the mean in square units

Not used commonly as a standalone statistic, but involved in the calculation of other statistics

42

Calculating the Sample Variance

$\bar{X}$ = the sample mean (M)

X = a score in the distribution

n = the number of cases in the sample

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

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of scores - 1}}$$

43

The Variance: Example

Step One: Determine the Deviation of Each Score ($X - \bar{X}$)

Score	**Deviation ($X - \bar{X}$)**
80	
82	
86	
87	
89	
90	
92	
92	
94	
96	

44

The Variance: Example

Step One: Determine the Deviation of Each Score ($X - \overline{X}$)

Score	$(X - \overline{X})$	Deviation
80	80 – 88.8 =	-8.8
82	82 – 88.8 =	-6.8
86	86 – 88.8 =	-2.8
87	87 – 88.8 =	-1.8
89	89 – 88.8 =	0.2
90	90 – 88.8 =	1.2
92	92 – 88.8 =	3.2
92	92 – 88.8 =	3.2
94	94 – 88.8 =	5.2
96	96 – 88.8 =	7.2

45

The Variance: Example

Step Two: Calculate the **Squared** Deviations

$$(X - \overline{X})^2$$

Score	$(X - \overline{X})$	Deviation	Squared Deviation
80	80 – 88.8 =	-8.8	$(-8.8)^2 = 77.44$
82	82 – 88.8 =	-6.8	$(-6.8)^2 = 46.24$
86	86 – 88.8 =	-2.8	$(-2.8)^2 = 7.84$
87	87 – 88.8 =	-1.8	$(-1.8)^2 = 3.24$
89	89 – 88.8 =	0.2	$(0.2)^2 = 0.04$
90	90 – 88.8 =	1.2	$(1.2)^2 = 1.44$
92	92 – 88.8 =	3.2	$(3.2)^2 = 10.24$
92	92 – 88.8 =	3.2	$(3.2)^2 = 10.24$
94	94 – 88.8 =	5.2	$(5.2)^2 = 27.04$
96	96 – 88.8 =	7.2	$(7.2)^2 = 51.84$

46

The Variance: Example

Step Three: Find the **Sum of Squared** Deviations $\Sigma(X - \overline{X})^2$

Sum of Squared Deviations = SS

SS = 77.44 + 46.24 + 7.84 + 3.24 + 0.04 + 1.44 + 10.24 + 10.24 + 27.04 + 51.84

SS = 235.6

Sum of Squared Deviations (SS) = 235.6

47

The Variance: Example

Step Four: Divide Sum of Squares by n-1 $\dfrac{\Sigma(X - \overline{X})^2}{n - 1}$

Sum of Squared Deviations (SS) = 235.6

The number of scores (n) = 10

n – 1 = 10 – 1 = 9

Variance = $s^2 = \dfrac{\Sigma(X - \overline{X})^2}{n - 1}$ = 235.6 / 9 = 26.18

48

The Standard Deviation

Most common and most important measure of variability

A measure of the standard, or average, distance from the mean

Describes whether the scores are clustered closely around the mean or are widely scattered

Standard deviation (SD) is the squared root of variance

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

49

The Standard Deviation: Example

Variance = $s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$ = 26.18

SD = $s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$ = $\sqrt{26.18}$ = 5.1164

SD = 5.12

50

Descriptive Statistics of DADA Scores

The distribution of DADA scores:

80, 82, 86, 87, 89, 90, 92, 92, 94, 96

Measures of Central Tendency

Mean = 88.8

Median = 89.5

Mode = 92

Measures of Variability

Range = 96 – 80 = 16

Variance = 26.18

Standard Deviation = 5.12

Most useful

51

The Mean and Standard Deviation

Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions

- A mean provides a rough summary of the distribution
- A standard deviation describes scores in terms of average distance from the mean

When the scores are clustered close to the mean, the standard deviation is small

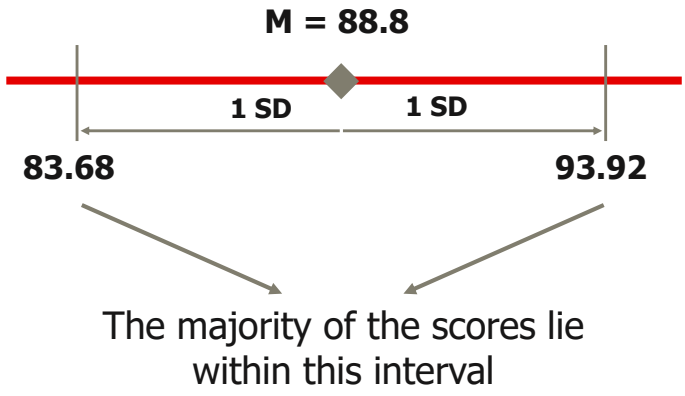
- When the scores are scattered widely around the mean, the standard deviation is large

52

The Mean and Standard Deviation

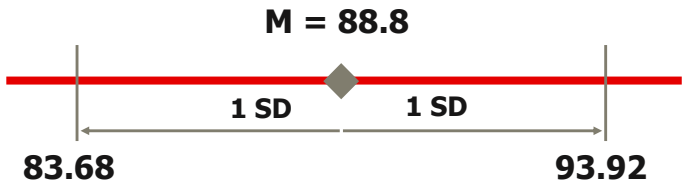
Describe an entire distribution with just two numbers (M and SD)

The mean DADA scores was 88.8 (SD = 5.12)



53

The Mean and Standard Deviation



The distribution of DADA scores:

80, 82, **86**, **87**, **89**, **90**, **92**, **92**, 94, 96

54

55

Normal Distribution

56

Frequency Distributions

A **frequency distribution** is

- An organized tabulation
- Showing the number of data points located in each category on the scale of measurement

Can be either a **table** or a **graph**

Always shows

- The categories that make up the scale
- The frequency, or number of data points, in each category

57

Frequency Distributions

Suppose we had the following scores:

1, 2, 2, 3, 4, 5, 5, 4, 4, 3, 3, 3

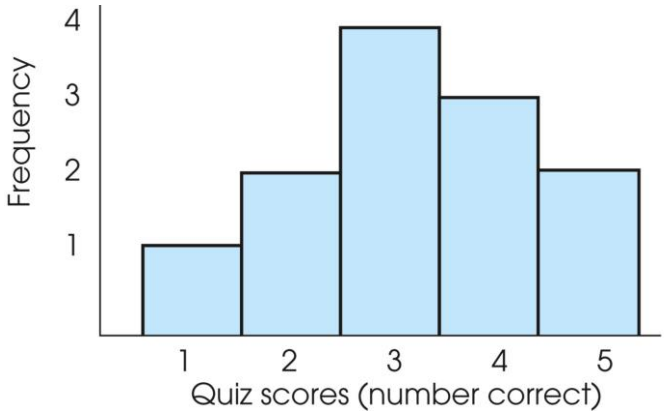
Frequency distribution:

X	f
5	2
4	3
3	4
2	2
1	1

58

Histogram

A graphical representation of a frequency distribution



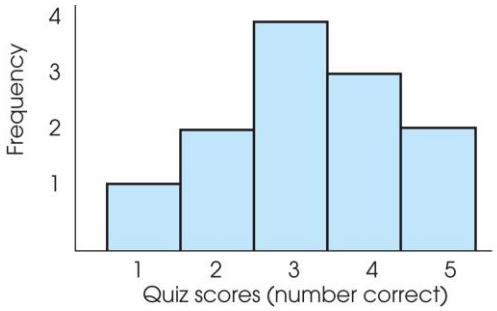
X	f
5	2
4	3
3	4
2	2
1	1

59

Histogram

X-axis represents all scores in the distribution (from min thru max)

Y-axis represents frequency of each score



Draw bars above each score

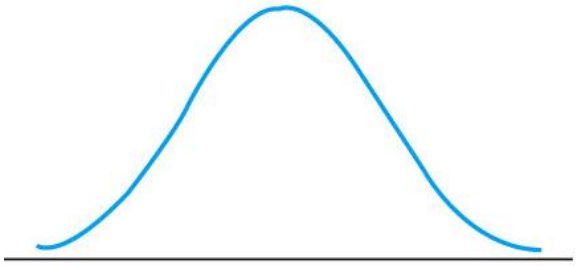
- Height of bar corresponds to frequency
- Width of bar corresponds to score real limits (or one-half score unit above/below discrete scores)

60

The Normal Distribution

In an ideal world our data would be distributed symmetrically around the center of all scores

This is known as a normal distribution and is characterized by the bell-shaped curve



61

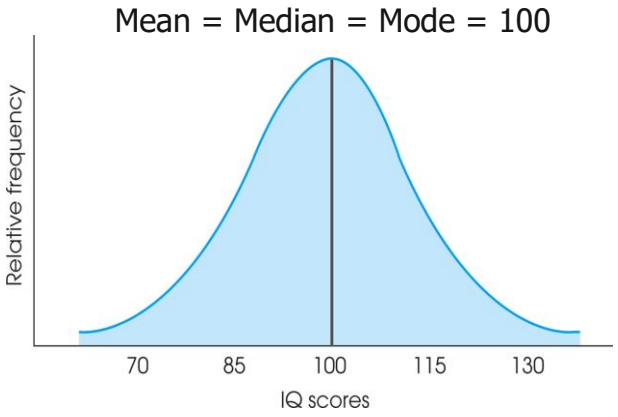
The Normal Distribution

The majority of scores lie around the center of the distribution

As we get further away from the center, the frequency of scores decreases

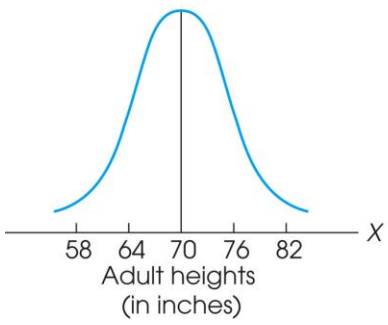
Many naturally occurring things have this shape of distribution

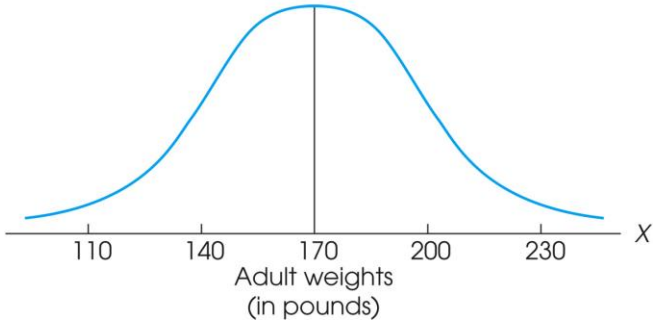
e.g., IQ scores, height, weight



62

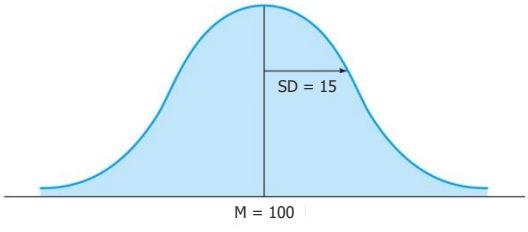
Normally Distributed Variables

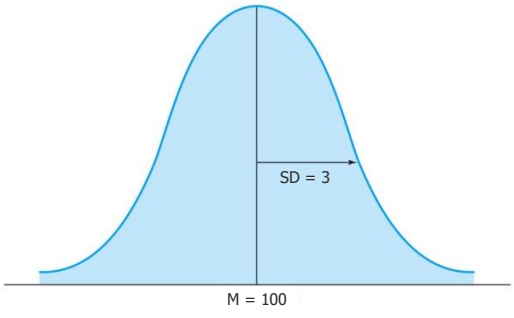




63

Standard Deviation and Distribution Shape





M = 100, SD = 15

Larger SD ➔ Flatter distribution

Scores are more spread out

M = 100, SD = 3

Smaller SD ➔ More pointy distribution

Scores close the mean are very frequent

64

65