Lecture 1: Basic Concepts and Descriptive Statistics

Ye Tian

Department of Statistics, Columbia University
Calculus-based Introduction to Statistics (S1201)

July 5, 2022

Today's plan

- Know basic concepts in/about statistics
  - Data, samples, statistics, population, randomness ...
  - The loop of data analysis/statistical modeling
- Know some commonly-used descriptive statistics
  - Descriptive statistics: mean, median, quartiles, variance, standard deviation

Basic concepts

What is the data?

- Can 4.45, 4.93, 4.95 be called as the data?

- Now?

Basic concepts

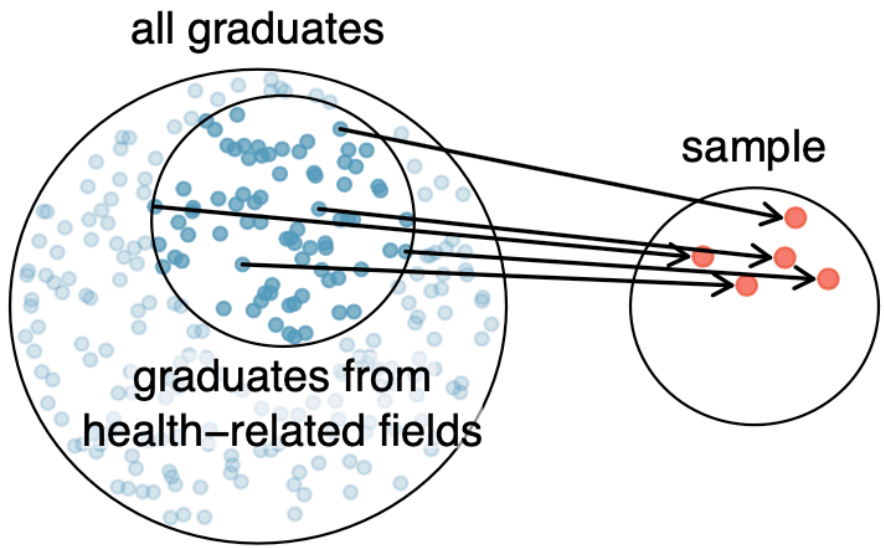
- Scientists seek to answer questions using rigorous methods and careful observations. These observations — collected from the likes of field notes, surveys, and experiments — form the backbone of a statistical investigation and are called **data**.

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93

variable	description
`name`	County name.
`state`	State where the county resides, or the District of Columbia.
`pop`	Population in 2017.
`pop_change`	Percent change in the population from 2010 to 2017. For example, the value `1.48` in the first row means the population for this county increased by 1.48% from 2010 to 2017.
`poverty`	Percent of the population in poverty.

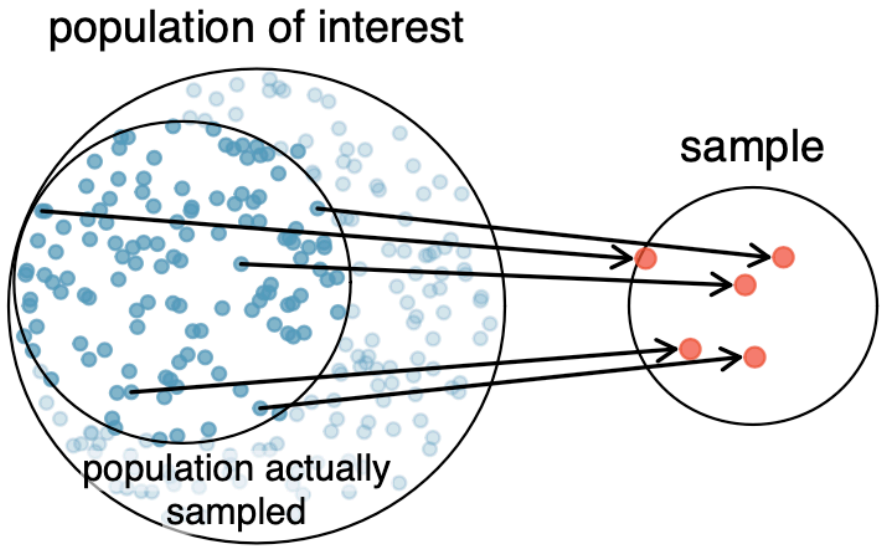
Basic concepts

- An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest.
- When desired information is available for all objects in the population, we have what is called a **census**.
- A **sample** represents a subset of the cases and is often a small fraction of the population. (Why sampling is important?)



Basic concepts

- It is possible that the sample could be skewed to that data collector's interests, which may be entirely unintentional. This introduces **(selection) bias** into a sample.
- Sampling **randomly** helps resolve this problem $\rightarrow$ simple random sample
- We want the samples to be **representative** of the entire **population of interest**.



Example: selection bias

The following is the final grades of some courses I took in college. I listed them on my CV when applying for Ph.D. programs.

Core courses:
Mathematical Analysis 1-3 (99, 87, 92), Linear Algebra 1-2 (99, 90), Real Analysis (95), Complex Analysis (95), Partial Differential Equation (95), Applied Stochastic Processes (97), Mathematical Statistics (93), Multivariate Statistical Analysis (93), Applied Statistical Software (R & Python) (100), Bayes Analysis (graduate level, 93), Regression Analysis (96), Operations Research 1-2 (98, 91)

- Are the course grades fully representative? Some grades I didn't mention:
  - Functional Analysis (85)
  - Ordinary Differential Equations (83)
  - Experimental Design (77)
  - Non-parametric statistics (79)
- Was I lying?

Basic concepts

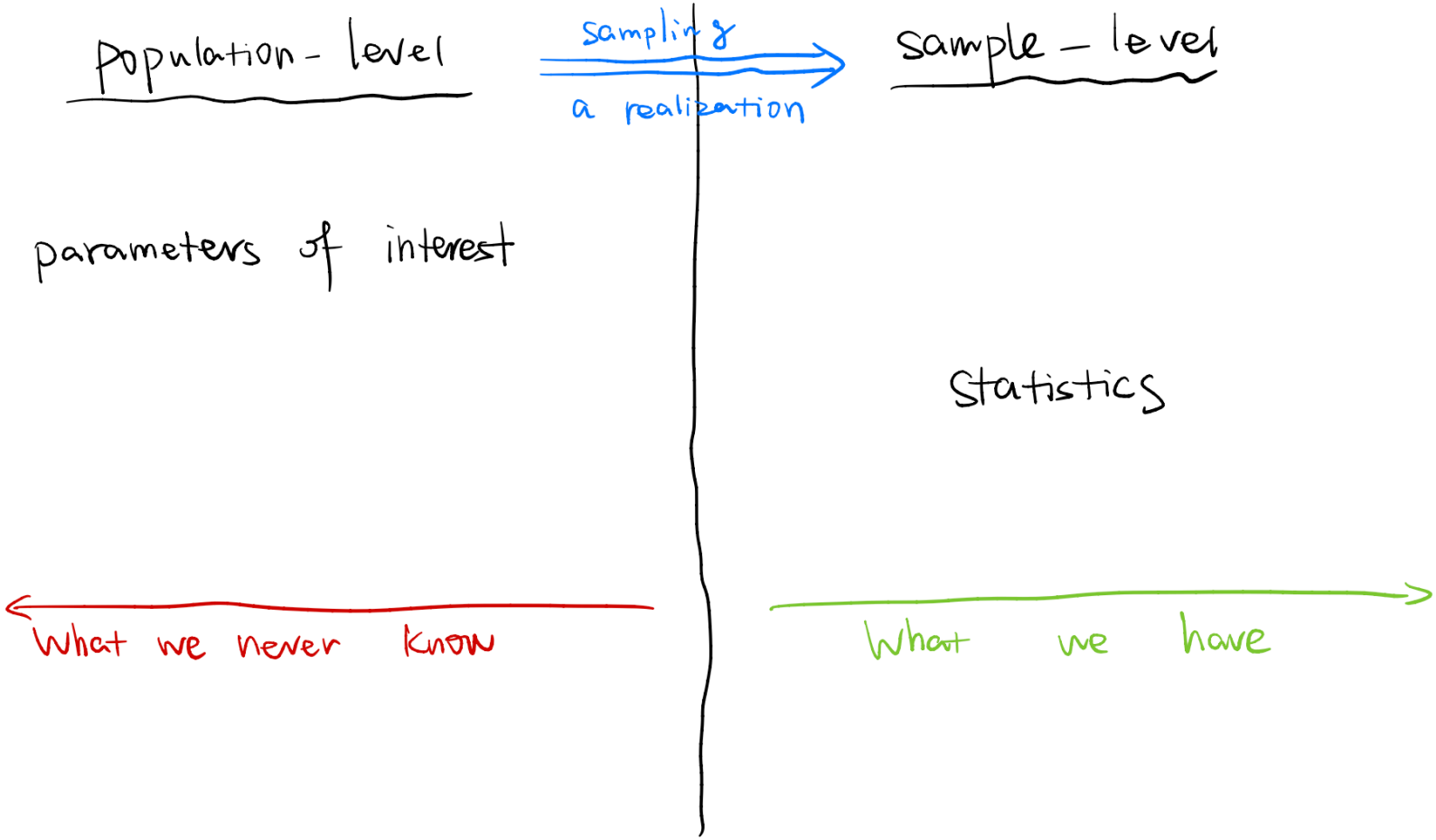
- An investigation will typically focus on a well-defined collection of objects constituting a **population** of interest.
- A **sample** represents a subset of the cases and is often a small fraction of the population.
- Sampling introduces some **randomness** into our study.
- These samples consist of a **dataset**.
- Each column is called a **(random) variable** (a very coarse definition)

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93

Basic concepts

- A (summary) (sample) **statistic** indicates some characteristic of the **dataset**:
  - Maximum of the population among all counties
  - Average of the poverty percentage among all counties
- Parameter of interest and sample statistic:
  - Usually what we are really interested in is some parameter of the population (most are impossible/hard to directly know)
    - The percentage of cat videos on YouTube
    - The average battery life of MacBooks
  - We **estimate/infer** this information from the samples (the dataset)
    - The percentage of cat videos among 100 randomly selected videos on YouTube
    - The average battery life of 100 randomly selected MacBooks
    - They are called as **sample statistics** because they indicate some characteristic of the **dataset**
    - We use **sample statistics** to estimate the true **parameter** of the **population**

Population, samples, probability and inference



- Statistical **inference** is the procedure to infer parameters of population by sample statistics
- Population (and the parameter) is deterministic, while samples are random
- We often say a sample is a **realization** of the population (distribution)

Population, samples, probability and inference

Example 1: Suppose we want to know the average number of hours every day that each American spends on things that they really enjoy. The General Social Survey asked the question to a random sample of 1,155 Americans. The average relaxing time was found to be 1.65 hours.

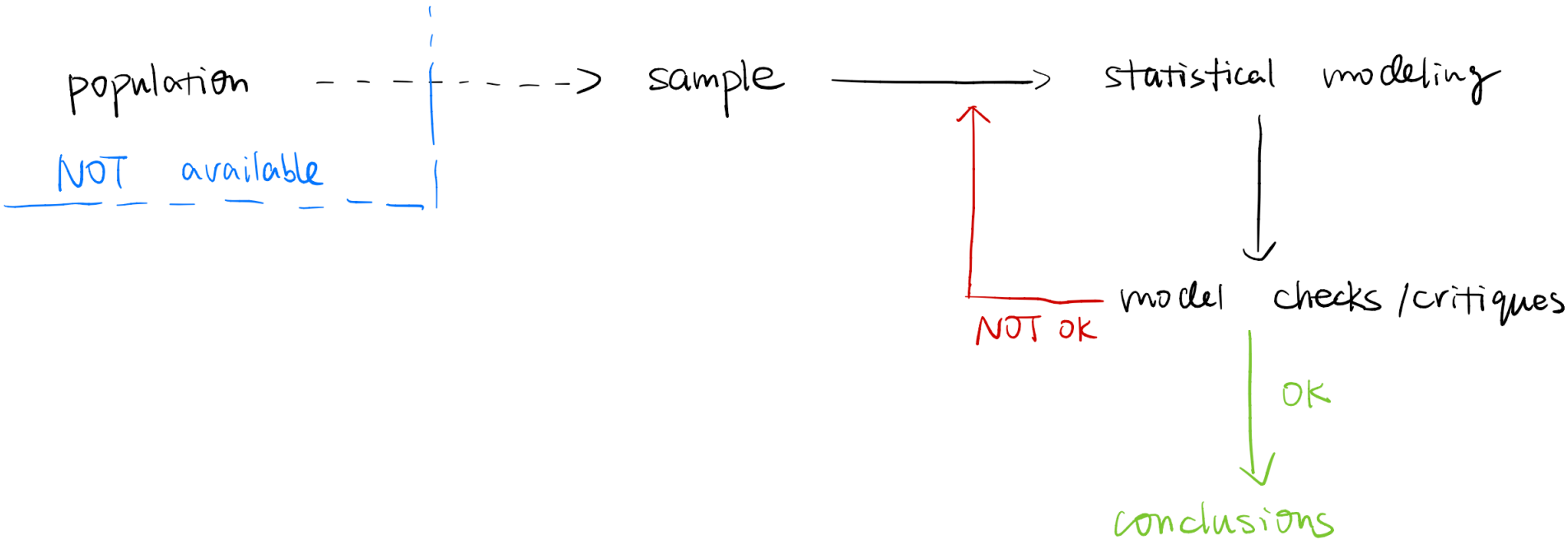
- Population of interest: all Americans
- Parameter of interest: the average relaxing time that each American has every day
- The sample:a random sample of 1,155 Americans
- A sample statistic: the average relaxing time among 1,155 samples (1.65 hours)

Population, samples, probability and inference

Example 2: Suppose you want to estimate the percentage of videos on YouTube that are cat videos. It is impossible for you to watch all videos on YouTube so you use a random video picker to select 1000 videos for you. You find that 2% of these videos are cat videos. Determine which of the following is an observation, a variable, a sample statistic (value calculated based on the observed sample), or a population parameter.

- Population of interest:all videos on YouTube
- Parameter of interest:the percentage of videos on YouTube that are cat videos
- The sample:1000 videos we select
- A sample statistic:the percentage of cat videos among 1000 samples (2%)

The loop of data analysis



Descriptive **statistics**

Two types of data/variables

- **Continuous** data: age, income, blood pressure, height, weight, etc.
- **Discrete (categorical)** data: number of hits in baseball game, number of patients who respond to treatment, etc.

	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93

For descriptive statistics, most of them make sense for both two types of data.
For visualization (to be covered in next class), many methods are only meaningful for either continuous or discrete data, not both.

What do we care about?

The grades of final exams of this course last summer:

99, 70, 74, 55, 60, 60, 80, 88, 85, 92, 98, 100, 86, 85, 74, 90, 72, 92, 88, 87, 81, 100, 79, 90, 68, 89, 91, 90, 96, 85, 100

What are you curious about this data?

- Average?
- Maximum? Minimum?
- If I got 90, where do I stand in the class?
- What's the range of "most" people's grades?
- How much difference between people's grades?

Location

Suppose we have samples $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a number.

- **Sample mean**: $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- **Sample median**:
  - First rank the samples: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ (we call them as **order statistics**)
  - Then

$$x_{\text{median}} = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ is odd} \\ \frac{1}{2}\left[x_{(n/2)} + x_{((n/2)+1)}\right], & \text{if } n \text{ is even} \end{cases}$$

Recall our data (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100, 100

- Sample mean $\bar{x} = \frac{1}{31}(55 + \ldots + 100) = 84$
- Sample median $x_{\text{median}} = 87$

Location

Suppose we have samples $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a number. Rank the samples: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

- **Sample median**:

$$x_{\text{median}} = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ is odd} \\ \frac{1}{2}[x_{(n/2)} + x_{((n/2)+1)}], & \text{if } n \text{ is even} \end{cases}$$

  Median is the midpoint of the data: 50% of the values are below it. Hence, it is also the 50th **percentile** or 50% **quantile**.

- **Sample percentiles/quantiles**: kth percentile or $k\%$ quantile is the value where $k\%$ of the values are **below**. Remember $k\% =$ the number of values below$/(n-1)$.

Recall our data (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100, 100

- 10% quantile $= 68$, $(100 \times 4/30)\%$ quantile $= 70$
- What's the 15% quantile?

Trick of interpolation

Recall our data (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

- What's the 15% quantile?
  - $100 \times 4/30\% \approx 13.3\%$ quantile $= 70$
    $100 \times 5/30\% \approx 16.7\%$ quantile $= 72$
  - $13.3\% \leq 15\% \leq 16.7\%$ and $15\% = \frac{1}{2} \times 13.3\% + \frac{1}{2} \times 16.7\%$
  - Therefore we can approximate 15% quantile as
    $\frac{1}{2} \times 70 + \frac{1}{2} \times 72 = 71$
- What about the 16% quantile?
  - $13.3\% \leq 16\% \leq 16.7\%$ and
    $16\% = 20.6\% \times 13.3\% + 79.4\% \times 16.7\%$
  - Therefore we can approximate 16% quantile as
    $20.6\% \times 70 + 79.4\% \times 72 = 71.588$

Trick of interpolation: illustration

$0.2\times$  $+0.8\times$ 

- How does it taste like?
- 0.2 $\times$ "sweet" + 0.8 $\times$ "salty"
- Recall: What's the 15% quantile?
  - $100 \times 4/30\% \approx 13.3\%$ quantile $= 70$, $100 \times 5/30\% \approx 16.7\%$ quantile $= 72$
  - $13.3\% \leq 15\% \leq 16.7\%$ and $15\% = \frac{1}{2} \times 13.3\% + \frac{1}{2} \times 16.7\%$
  - Therefore we can approximate 15% quantile as $\frac{1}{2} \times 70 + \frac{1}{2} \times 72 = 71$

Location

Suppose we have samples $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a number.
Rank the samples: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$

- **Sample quartiles** (not **quantiles**)
  - 25% quantile (25th percentile, lower forth) is also called **the 1st quartile** (Q1)
  - 50% quantile (50th percentile) is also called **the 2nd quartile** (Q2), i.e. the **median**
  - 75% quantile (75th percentile, upper forth) is also called **the 3rd quartile** (Q3)
- **Range** = maximum − minimum
- **Interquartile range** (IQR) = Q3 − Q1
- **Five number summary**: min, lower forth, median, upper forth, max

---

Recall our data (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100, 100

- Range $= 100 - 55 = 45$, 1st quartile $= 50\% \times 74 + 50\% \times 79 = 76.5$
  3rd quartile $= 50\% \times 91 + 50\% \times 92 = 91.5$
- IQR = Q3 − Q1 $= 91.5 - 76.5 = 15$

Which perspective these statistics describe

The grades of final exams of this course last summer (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100

- Sample mean (84): the average score, the average level of students's understanding of the course contents
- 1st quartile (76.5), sample median (87), 3rd quartile (91.5): how data "spread", the position of people with some score
- Range (45), IQR (15): how much difference between people

Question: Which one is more sensitive to the "extreme" values, sample mean or sample median?
Hint: What if the lowest score is 10 instead of 55? Which one is affected? Which one is invariant?

Variability

Suppose we have samples $\{x_1, x_2, \ldots, x_n\}$, where each x_i is a number.

- **Sample variance** is roughly the average squared deviation from the mean.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

  where $\bar{x}$ is the sample mean
- **Sample standard deviation** is the square root of the sample variance, i.e. s here.

Recall our data (re-ranked): 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100, 100

- Sample mean $\bar{x} = \frac{1}{31}(55 + \ldots + 100) = 84$
- Sample variance $s^2 = \frac{1}{30}[(55-84)^2 + \ldots + (100-84)^2] = 151$
- Sample standard deviation $s = \sqrt{151} = 12.29$

What do we care about? (Revisited)

The grades of final exams of this course last summer:

99, 70, 74, 55, 60, 60, 80, 88, 85, 92, 98, 100, 86, 85, 74, 90, 72, 92, 88, 87, 81, 100, 79, 90, 68, 89, 91, 90, 96, 85

What are you curious about this data?

- Average? Sample mean $\bar{x} = 84$
- Maximum? Minimum? Maximum $= 100$, minimum $= 55$
- If I got 90, where do I stand in the class?
  - Re-rank data: 55, 60, 60, 68, 70, 72, 74, 74, 79, 80, 81, 85, 85, 85, 86, 87, 88, 88, 89, 90, 90, 90, 91, 92, 92, 96, 98, 99, 100, 100, 100
  - The first "90" is 66.7% quantile, the third "90" is 73.3% quantile
  - So I am doing better than $66.7\% \sim 73.3\%$ of my classmates
- What's the range of "most" people's grades? 1st quartile (76.5), sample median (87), 3rd quartile (91.5)
- How much difference between people's grades? Sample standard deviation $s = \sqrt{151} = 12.29$

About the mode

For discrete data, for example, the ratings of a professor from 15 students: A, A, A, A, A, A, A, A, B, B, B, B, C, C, D

- The **mode** is the value that appears most often in a set of data values
- Here the mode equals A

For continuous data, instead of talking about its mode, it may be better to look at its "distribution" (via the histogram etc.) and get a sense that where the data concentrates.

Reading list (optional)

- "Probability and Statistics for Engineering and the Sciences" (9th edition):
  - Chapter 1.1, 1.3, and 1.4
- "OpenIntro statistics" (4th edition, free online, download `[here]`):
  - Chapter 1.2
  - Chapter 1.3.1-1.3.3
  - Chapter 2.1.1, 2.1.2, 2.1.4-2.1.6

Many thanks to

- Chengliang Tang
- Yang Feng
- Joyce Robbins
- Owen Ward
- Wenda Zhou
- And all my teachers in the past 25 years