50.034 - Introduction to Probability and Statistics

Week 1 – Lecture 1

January–May Term, 2019

Lecture 1 Assigned Readings: 1.1, 1.2

About this course

This course covers basic concepts of probability and statistics.

Course Textbook:

- Probability and Statistics (4th Edition) by Morris H. DeGroot and Mark J. Schervish; Publisher: Pearson

We will follow the course textbook closely. There will be **assigned textbook readings** for all lectures (and some cohort classes too), so please get a copy of the course textbook!

Other Reference Books:

- J. L. Devore, Probability and Statistics for Engineering and the Sciences, 8th ed. Boston, 2011
- R. Walpole, R. H. Myers, S. L. Myers, and K. Ye, Essentials of Probability and Statistics for Engineers and Scientists

Course materials, such as lecture slides, cohort class notes, and homework, will be available on eDimension.

People

Instructors:

Prof. Ernest Chong	ernest_chong@sutd.edu.sg
Office hours	Weeks 1,6–9, TBD
Prof. Tony Quek	tonyquek@sutd.edu.sg
Office hours	Weeks 2-5, TBD
Prof. Gemma Roig	gemma_roig@sutd.edu.sg
Office hours	Weeks 10–13, TBD

Teaching assistants:

Ha Thi Phuong Thao (PhD) thiphuongthao_ha@mymail.sutd.edu.sg

Huang Xia (PhD) xia_huang@mymail.sutd.edu.sg

Koh Jing Yu (Undergraduate) 1002045@mymail.sutd.edu.sg

Course Evaluation

Class participation:	5%
Homework:	15%
4 Mini-quizzes:	15% (Weeks 3, 6, 9, 12)
Midterm Exam:	30% (Week 8)
Final Exam:	35% (Week 14)

(Best 3 of 4 mini-quiz scores will be counted towards final grade.)
Each mini-quiz is 15mins long (held during your cohort class).

Homework is assigned on every Monday and is due in the cohort class of the following week.

Graded homework will be returned one week after submission.

Homework solutions will be uploaded onto eDimension every Wednesday (1–2 days before graded homework is returned)

Homework Grading

Each homework set has five to six questions; only **two** will be graded. The grading policy for each selected question is:

Score	Criteria
5	correct answer and procedure
2–4	incorrect answer, partly correct procedure correct answer, missing critical procedure
1	selected question unanswered
0	homework not submitted before due time

Maximum total score for each homework set: 10

Issues with homework grading must be raised within one week upon receiving the graded homework.

Outline of Lecture

- Population, Sample, Variable
- Probability vs. Statistics
- Descriptive and Inferential statistics
- Frequency and Relative frequency
- Range, Mean, Median
- Percentile, Histogram

Population

Population – a well-defined collection of objects

All registered students in this course:

All aircrafts sold by Lockheed Martin last year:

All positive integers strictly below 10:

$$1,\quad 2,\quad 3,\quad 4,\quad 5,\quad 6,\quad 7,\quad 8,\quad 9$$

Notation for Collections

Given some objects, if we want to put them into a collection, we have to place them in between curly brackets $\{\}$.
The collection of all registered students in this course:

The collection of all aircrafts sold by Lockheed Martin last year:

The collection of all positive integers strictly below 10:

$$\{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9\}$$

Question: What is the difference between 0 and $\{0\}$?
Answer: 0 is a number, while $\{0\}$ is a collection that contains a single number, the number 0.

Sample, Variable

Sample – a subcollection of the population selected in some prescribed manner.
E.g. "Students currently present in this lecture" is a sample of "registered students in this course".
E.g. "Aircrafts with damaged rudders" is a sample of "aircrafts with any kind of defect".
E.g. $\{1, 3, 5, 7, 9\}$ is a sample of $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Variable – any characteristic whose value may change from one object to another in the population.
E.g. $x =$ gender of a student.
E.g. $y =$ number of courses offered this semester.
E.g. $z =$ price of a 3-bedroom apartment.

Example 1: Population vs Sample

A team of Boeing engineers has created a new material for airplane wings. To test the new material, the team created 10 prototypes of the wings using the material, and tested these wings under an extreme condition. The team then recorded the break-down times of these 10 wings.

Is each statement below a characteristic of 'population' or 'sample'?

1. The average break-down time of the 10 wings:
(a) Population, (b) Sample.

2. The average break-down time of any wing:
(a) Population, (b) Sample.

Population vs Sample

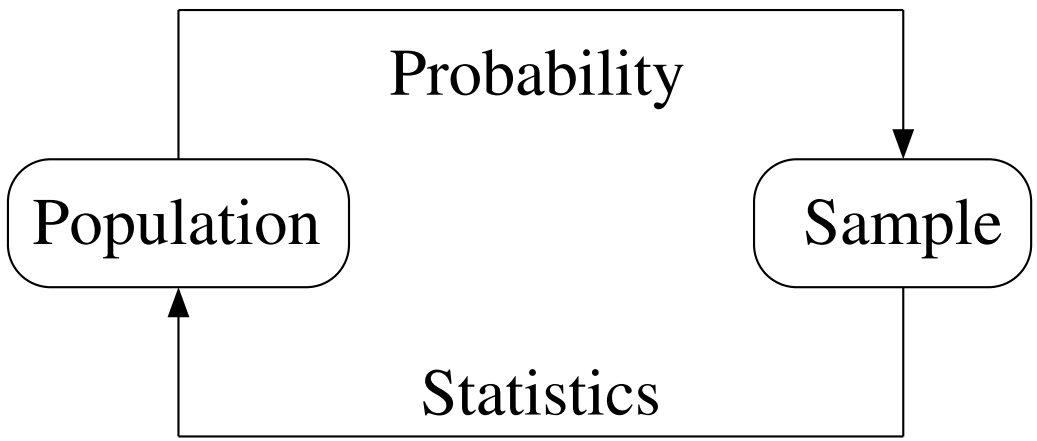
1. The average break-down time of the 10 wings:
(a) Population, (b) Sample.

2. The average break-down time of any wing:
(a) Population, (b) Sample.

Probability and Statistics

Probability: Properties of a population are known, and questions regarding a sample taken from the population are investigated (deductive reasoning).

Statistics: Characteristics of a sample are known from the experiment, and conclusions regarding the population are made (inductive reasoning).



Inferential Statistics

Inferential statistics involves making predictions or inferences about a population from observations and analyses of a sample.

Example 2: Inferential Statistics

Consider the population of all students enrolled in Course 50.034. There are 150 students enrolled in Course 50.034, so the population size is 150. We want to estimate the number of male students and female students in the population.

(Hint: Choose a sample with observable data to estimate the properties of the population.)

Solution: Consider all students in the front two rows of the lecture hall as a sample, and count the number of males n_M and females n_F, respectively. The proportions of males and females in the sample are $\frac{n_M}{n_M+n_F}$ and $\frac{n_F}{n_M+n_F}$ respectively, so the estimated numbers of males and females in the population are:

$$\frac{n_M}{n_M+n_F} \times 150 \text{ and } \frac{n_F}{n_M+n_F} \times 150, \text{ respectively.}$$

However, the answers we get may not be accurate, as we have only looked at a sample instead of the whole population.

Descriptive Statistics

Descriptive statistics involves obtaining descriptive summaries of a sample or a population. For example, such descriptive summaries could be in the form of frequency distributions, measures of central tendency (mean and median), or graphs such as histograms, pie charts and bar charts.

Some examples of techniques in descriptive statistics:

- Computation of means, medians, percentile, etc.
- Graphical visualization using histograms.

Descriptive Statistics: Range, Mean

Usually, your population or sample is a collection of numbers.

Range: the difference between the largest and smallest values.

Mean: the average of all values.

The population mean is often denoted by μ.

If the sample is a collection of numbers $\{x_1, \ldots, x_n\}$, then the sample mean, which is often denoted by $\bar{x}$, is defined as the value

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big).$$

Example 3: Linear transformation of a sample

Let $\{x_1, x_2, \ldots, x_n\}$ be a sample, and let a and b be constants. Let $y_i = ax_i + b$ be a linear transformation of x_i, for each $i = 1, \ldots, n$.

Then we have the following property:

$$\bar{y} = a\bar{x} + b$$

Proof:

$$\begin{aligned}\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\ &= a\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}nb = a\bar{x} + b\end{aligned}$$

Median

Suppose we have a population of numbers $\{x_1, \ldots, x_N\}$.
Let $\{x_{i_1}, \ldots, x_{i_n}\}$ be a sample of this population.
Assume that $x_1 \leq \cdots \leq x_N$ and $x_{i_1} \leq \cdots \leq x_{i_n}$.

[Here, the indices of the n sample values are $i_1, i_2, \ldots, i_n$. For example, if $\{x_1, \ldots, x_{10}\}$ is our population, then $\{x_2, x_3, x_5, x_7\}$ is one possible sample, where $i_1 = 2, i_2 = 3, i_3 = 5, i_4 = 7$, and $n = 4$.]

The median refers to the "middle" value after ordering the values.
The population median is defined as:

$$\tilde{\mu} = \begin{cases} x_m, & \text{if } N \text{ is odd of the form } N = 2m - 1; \\ \frac{x_m + x_{m+1}}{2} & \text{if } N \text{ is even of the form } N = 2m. \end{cases}$$

The sample median is defined as:

$$\tilde{x} = \begin{cases} x_{i_m}, & \text{if } n \text{ is odd of the form } n = 2m - 1; \\ \frac{x_{i_m} + x_{i_{m+1}}}{2}, & \text{if } n \text{ is even of the form } n = 2m. \end{cases}$$

Percentile

The *K*-th percentile of a sample or population is a number p such that at least $K\%$ of all sample values are less than or equal to p, and no more than $K\%$ of all sample values are strictly less than p.

Method to compute the k-th percentile of a given sample: Suppose we have a sample $\{x_1, x_2, \ldots, x_n\}$, such that after ordering the values, we get $x'_1 \leq x'_2 \leq \cdots \leq x'_n$.

- Simple cases $k = 0$ and $k = 100$:
  - The 0th percentile is x'_1, the smallest value in the sample.
  - The 100th percentile is x'_n, the largest value in the sample.
- For $0 < K < 100$: First compute $\frac{K}{100}n$.
  - If $\frac{K}{100}n$ is NOT a whole number, round it up to get the whole number m. Then the K-th percentile is x'_m.
  - If $\frac{K}{100}n$ is a whole number, say we call it M, then the K-th percentile is $\frac{x'_M + x'_{M+1}}{2}$.

Percentile

Suppose we have the following sample of ten numbers:

$$\{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}.$$

What is the k-th percentile of this sample for $k = 15, 20, 25, 48$?

Solution:

- $k = 15$: We compute $\frac{15}{100} \times 10 = 1.5$, which rounds up to 2. So the 15th percentile is 3 (2nd smallest number in sample).
- $k = 20$: We compute $\frac{20}{100} \times 10 = 2$, which is a whole number. So the 20th percentile is 4 (average of 2nd and 3rd numbers).
- $k = 25$: We compute $\frac{25}{100} \times 10 = 2.5$, which rounds up to 3. So the 25th percentile is 5 (3rd smallest number in sample).
- $k = 48$: $\frac{48}{100} \times 10 = 4.8$, which rounds up to 5. So the 48th percentile is 9 (5th smallest number in sample).

The k-th percentile may not necessarily be a value in the sample. (For example, the 20th percentile is not in the sample!)

Example 4

Let S be the sample $\{10, 3, 5\}$.

- What is the range, mean and median of S?
- What is the 50th percentile of S? In general, does "50th percentile" mean the same as "median"?

Example 4

Solution: The range is $10 - 3 = 7$.

The mean is $\bar{x} = (10 + 3 + 5)/3 = 6$.

The median is $\tilde{x} = 5$.

To find the 50th percentile, we rearrange the sample as $3, 5, 10$. We compute $\frac{50}{100} \times 3 = 1.5$, which rounds up to 2, so the 50th percentile is the second number 5.

Note: The "median" is exactly the same as "50th percentile"!

Frequency and Relative Frequency

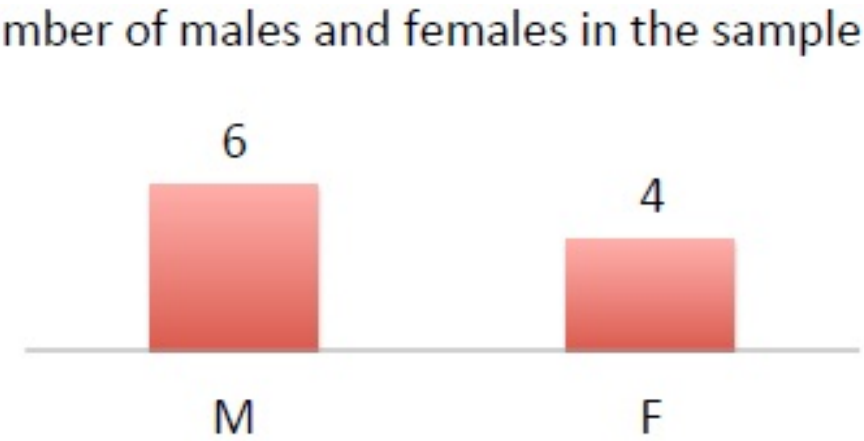
Consider a data set for some discrete variable x.

- The **frequency** of any particular value that x can take, is defined to be the number of occurrences of that value in the data set
- The **relative frequency** of a value is defined to be the fraction or proportion of occurrences of that value in the data set:

$$\text{relative frequency of a value } k = \frac{\text{number of times the value } k \text{ occurs}}{\text{total number of observations in the data set}}$$

Example 5: Frequency and relative frequency

Let x be the variable representing the gender of a student, which has two possible values, M or F. Suppose we have a sample of size $n = 10$ for this variable x, summarized as follows:

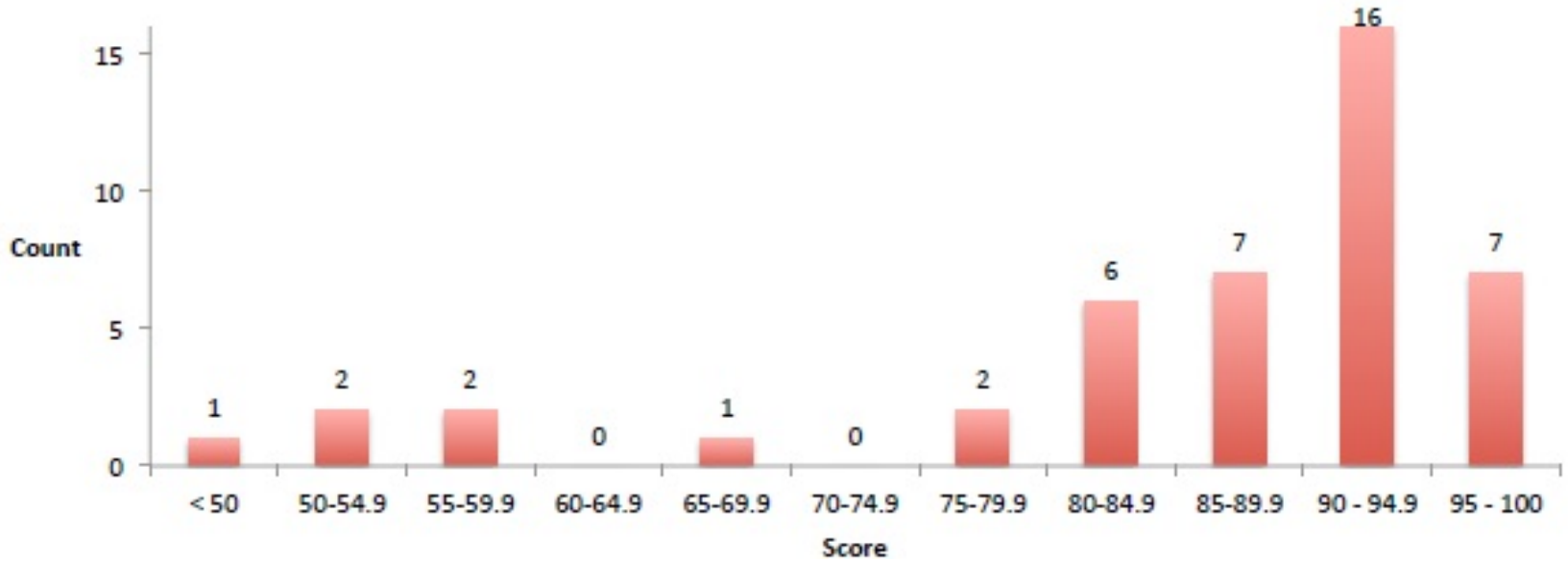


Then the frequencies of M and F are 6 and 4 respectively, and the relative frequencies of M and F are 0.6 and 0.4 respectively.

Sometimes, we use the word "class" interchangeably with "value". We could say that x is a variable with two possible classes M or F.

Example 6: Reading histograms

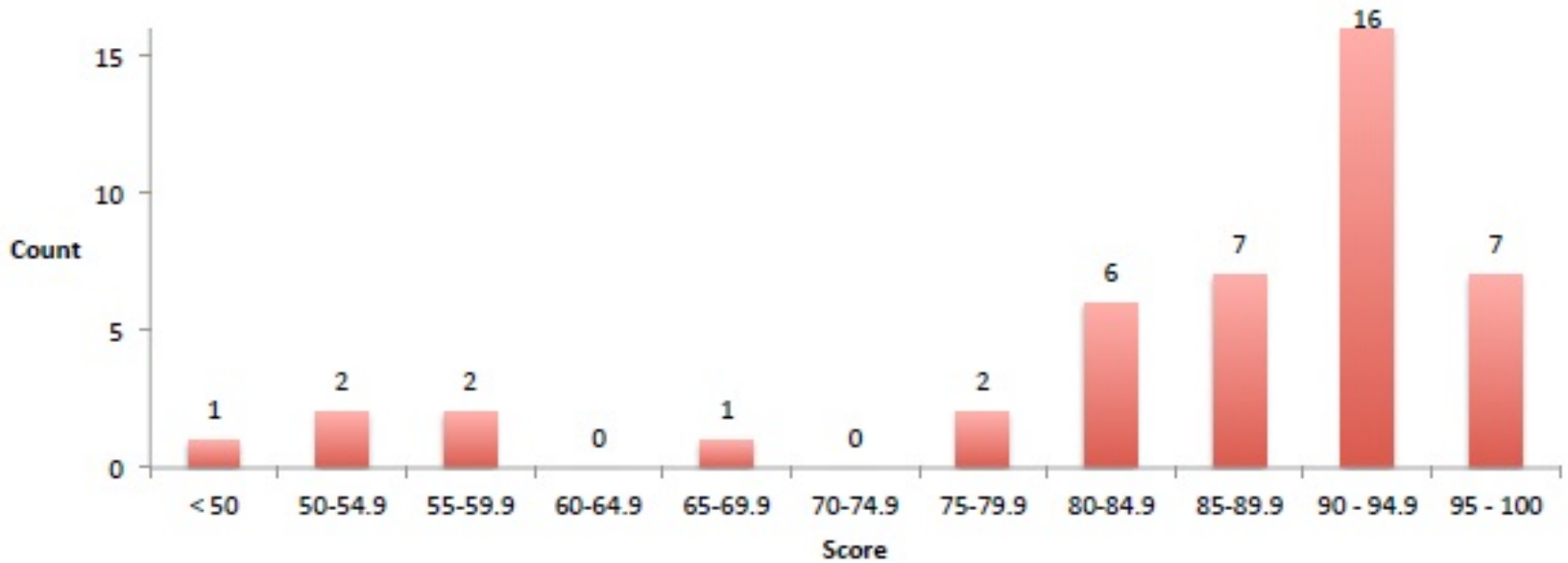
Consider the following histogram representing the exam scores of a subject (past year, $n = 44$ students):



Let x be a discrete variable that takes on one of the following 11 possible values/classes: $[0, 50)$, $[50, 55)$, ..., $[95, 100)$.
(Each value/class is an interval of exam scores.)

- What are the frequencies of all 11 possible values for x?
- What are the relative frequencies of all 11 possible values for x?
- What is the proportion of scores at least 90?

Example 6: Reading histograms (Solution)



Possible values for variable x: $[0, 50)$, $[50, 55)$, ..., $[95, 100)$.

- The frequencies of all values are 1, 2, 2, 0, 1, 0, 2, 6, 7, 16, and 7.
- The relative frequncies are 0.0227, 0.0455, 0.0455, 0, 0.0227, 0, 0.0455, 0.1364, 0.1591, 0.3636, and 0.1591, which sum up to be 1.
- The proportion of scores ≥ 90 is $0.3636 + 0.1591 = 0.5227$.

Constructing a histogram for discrete data

Steps to construct a histogram to represent frequency distribution.

- Choose the number of classes for your variable x. There is no fixed rule on how to choose this number. Typically, 5–20 classes will be satisfactory for most data sets. If you want to have more than 20 classes, then a reasonable rule of thumb is

$$\text{number of classes} \approx \sqrt{\text{number of observations}}$$

- Determine the frequency and relative frequency of each possible class for x.
- Mark all possible classes on a horizontal scale.
- Above each class, draw a rectangle whose height is the frequency or relative frequency of the corresponding class.

Example 7: Histograms

The following data represent the length of life in seconds, of 90 fruit flies subject to a new spray in a laboratory experiment:

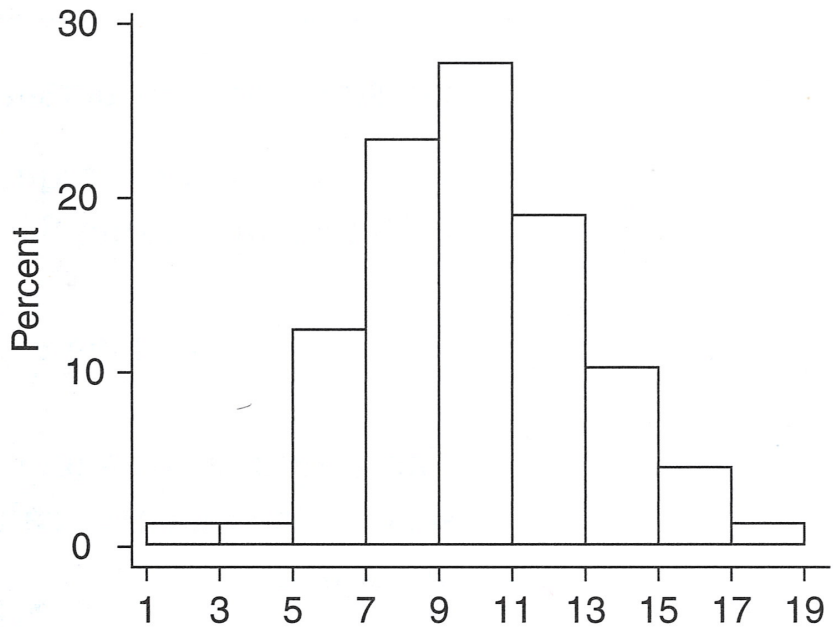
2.97	4.00	5.20	5.56	5.94	5.98	6.35	6.62	6.72	6.78
6.80	6.85	6.94	7.15	7.16	7.23	7.29	7.62	7.62	7.69
7.73	7.87	7.93	8.00	8.26	8.29	8.37	8.47	8.54	8.58
8.61	8.67	8.69	8.81	9.07	9.27	9.37	9.43	9.52	9.58
9.60	9.76	9.82	9.83	9.83	9.84	9.96	10.04	10.21	10.28
10.28	10.30	10.35	10.36	10.40	10.49	10.50	10.64	10.95	11.09
11.12	11.21	11.29	11.43	11.62	11.70	11.70	12.16	12.19	12.28
12.31	12.62	12.69	12.71	12.91	12.92	13.11	13.38	13.42	13.43
13.47	13.60	13.96	14.24	14.35	15.12	15.24	16.06	16.90	18.26

(a) Construct a relative frequency histogram.
(b) How many peaks are there in the histogram?

Example 7: Histograms

We can classify the 90 observations into $\sqrt{90} \approx 9$ classes.

Class	1−<3	3−<5	5−<7	7−<9	9−<11	11−<13	13−<15	15−<17	17−<19
Frequency	1	1	11	21	25	17	9	4	1
Relative frequency	.011	.011	.122	.233	.278	.189	.100	.044	.011



There is only one peak in the histogram.

Bivariate histograms

Histograms may also be used to show the distribution of two variables, such as the height and weight of students. A typical bivariate histogram looks like this:

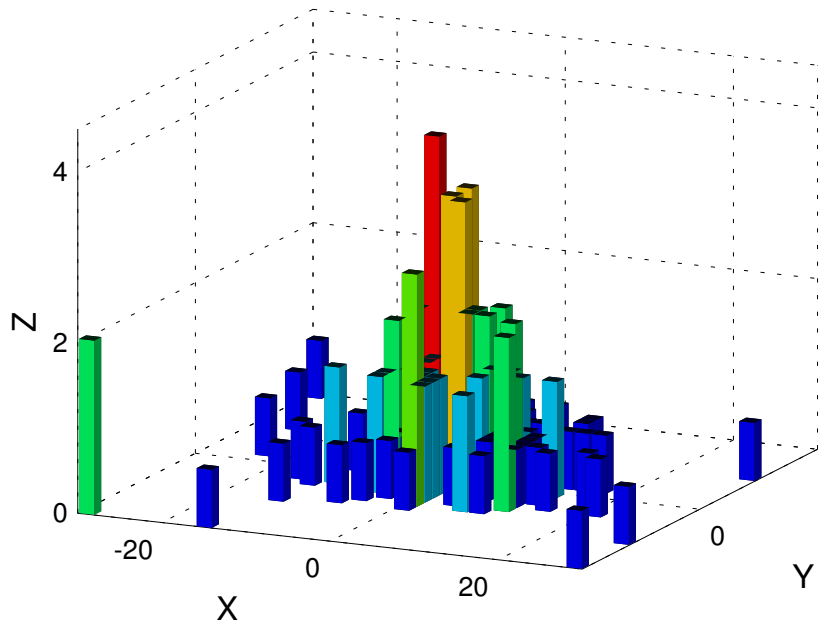


Image Source: Chen Lu Jie.

Summary

- Population, Sample, Variable
- Probability vs. Statistics
- Descriptive and Inferential statistics
- Frequency and Relative frequency
- Range, Mean, Median
- Percentile, Histogram

Announcement:
Due to the festive season next week, there will be a **make-up Lecture 4**, which will be held **next Friday**, 2:30pm-4pm, at LT5.