05 20 2025 class

Cesai Li

May 2025

Outline

Course Information

Lectures will be audio-recorded for accommodation reasons. Questions are encouraged, but if you do not wish to be recorded, you may ask during the break time, after class, or in office hours.

Instructor: Cesai Li

Office: 665 Commonwealth Ave, CCDS 410

E-mail: cesai@bu.edu

Office hours: Monday and Tuesday 12-1pm, Wednesday 2-3pm

Course Meeting: Monday, Tuesday, Wednesday, Thursday, 3-4:50 pm, MCS B37

No Class Dates:

1. 05/26 MON no class, 05/30 FRI 3-5pm make-up class
2. 05/19 THU no class, no make-up class

Textbook: *STATISTICS: Informed Decisions Using Data*, Michael Sullivan, **7e** with MyLab Statistics. To register for MyLab Statistics of this course, see the Student Registration Instructions on the course Blackboard site or the course homepage

https://cesail.github.io/courses/25s1ma116/

ISBN

Physical copy: 9780138253332

Etext: 9780138253592

If you want to purchase a physical copy but are unsure about edition numbers, etc. Talk to me or send me an email.

Course Info

Evaluation:	**Attendance**	30%
	Quizzes	$5\% \times 4 = 20\%$
	Homework	$5\% \times 5 = 25\%$
	Final exam	25%

Attendance Policy

- There are **22 class meetings** in total:
  - **21 regular lectures** (May 20–June 25)
  - **1 final exam** (held on June 26 during the usual 2-hour lecture slot)
- **Full attendance credit** is awarded for attending **at least 18 of the 21 regular lectures**.
- If fewer than 18 lectures are attended, the attendance grade is calculated as:
  - **1.2% per lecture attended** (e.g., a student who attends 15 regular lectures gets 18% for their attendence credit. If this student gets full mark on quizzes, homework, and the final, then their course grade is 18%+20%+25%+25%=88%.)
- This policy is designed to accommodate occasional absences due to illness, emergencies, or other unavoidable circumstances. It is not intended to imply that students are encouraged to skip up to 3 lectures without cause.

Quizzes

- Four 1-hour (actually 50 minutes) quizzes are scheduled to be held on 05/30, 06/05, 06/12, 06/18, 3pm-4pm in class.

Calculators are allowed but not required. Any smart electronic devices (e.g., those capable of web searching, AI Q&A, storing images, or sending messages) are strictly prohibited. These devices must be turned off and stored in your backpack during quizzes.

Homework

- Homework will be assigned in MyLab Statistics. Due days are 05/27, 06/02, 06/09, 06/16, 06/23. Homework will be due at 23:59pm on its due day. You may discuss homework questions with peers.

Final Exam Requirement

- A minimum score of **30% on the final exam** is **required to pass the course**, independent of other grades. This policy will be strictly enforced.

Academic Conduct

Students' work and conduct in this course are governed by the **BU Academic Conduct Code**. It is each **student's responsibility** to be familiar with and adhere to the provisions of the Code. During in-class quizzes and the final exam, no collaboration is permitted. Boston University maintains a strict policy against academic dishonesty. Any form of cheating, plagiarism, collusion, unauthorized access to materials during quizzes and exams, or ghostwriting will not be tolerated. Penalties for violations of the Academic Conduct Code may include suspension or expulsion from the University.

You are allowed to discuss homework questions with peers, or consult help from technology and any resources. You have 10 attempts for each question (unless the question is multiple-choice style, for which you still have multiple attempts.)

Examples of violations of the BU Academic Conduct Code

Unauthorized downloading, uploading, sharing, and/or duplicating course materials including assignments, slides, quizzes without the instructor's express permission. This includes downloading/uploading/viewing/selling copyrighted material found on commercial notes-sharing websites such as Course Hero.

Plagairism

Copying the answers of another student on an examination; copying or restating the work or ideas of another person/persons or AI on an examination without appropriately citing the source; and collaborating with someone else in an academic endeavor without acknowledging their contribution.

Break during class

3:00-3:50pm Lectures/Quizzes
3:50-4:05pm 15 min Break
4:05-4:50pm Lectures

MyLab Statistics Registration

CAS MA 116

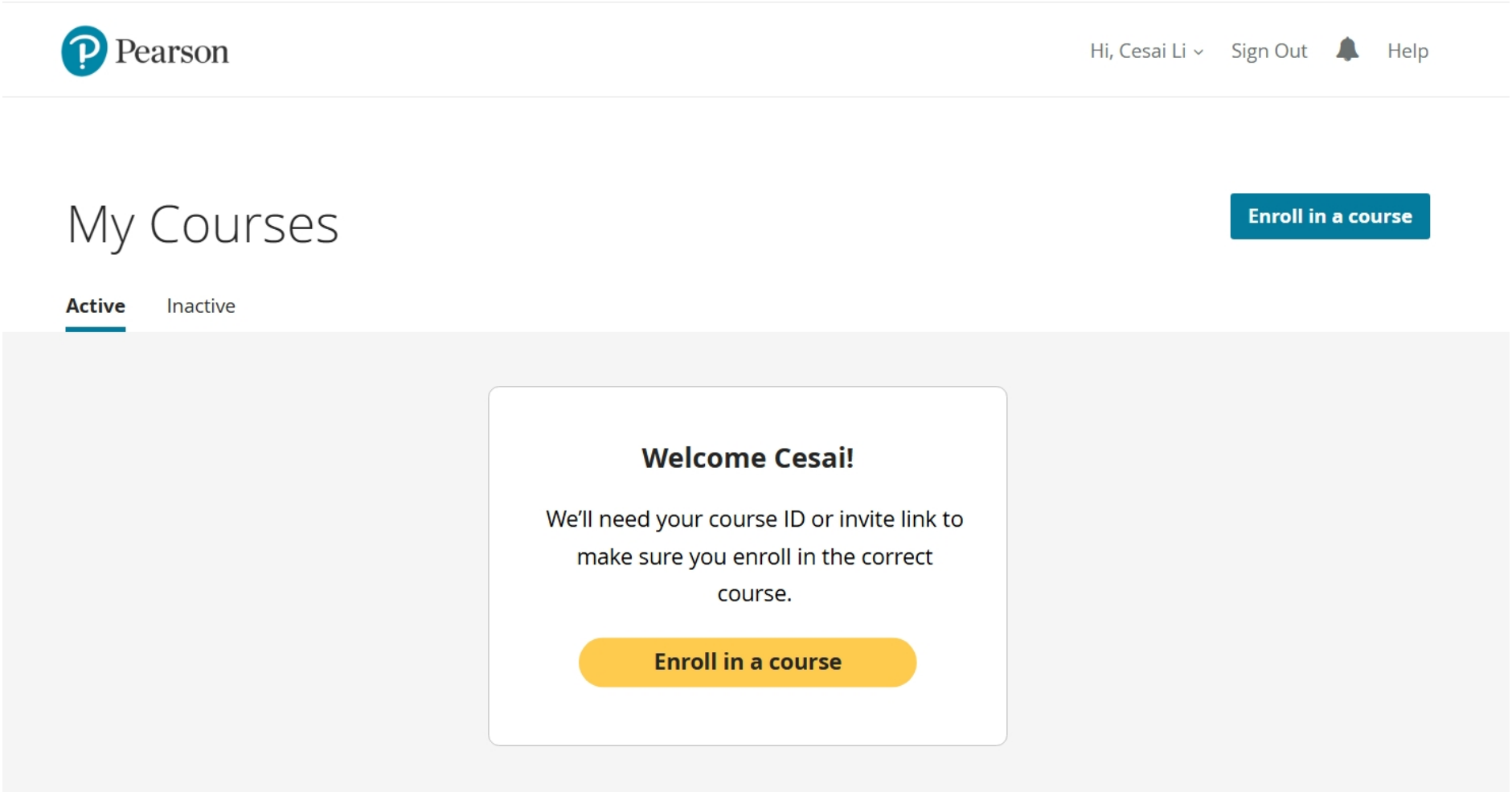
This is the syllabus .
MyLab Statistics registration guide. Students are encouraged **NOT** to purchase any course materials, including the MyLab Statistics access code or textbook, until after the first lecture on 05/20.

BU Academic Conduct Code

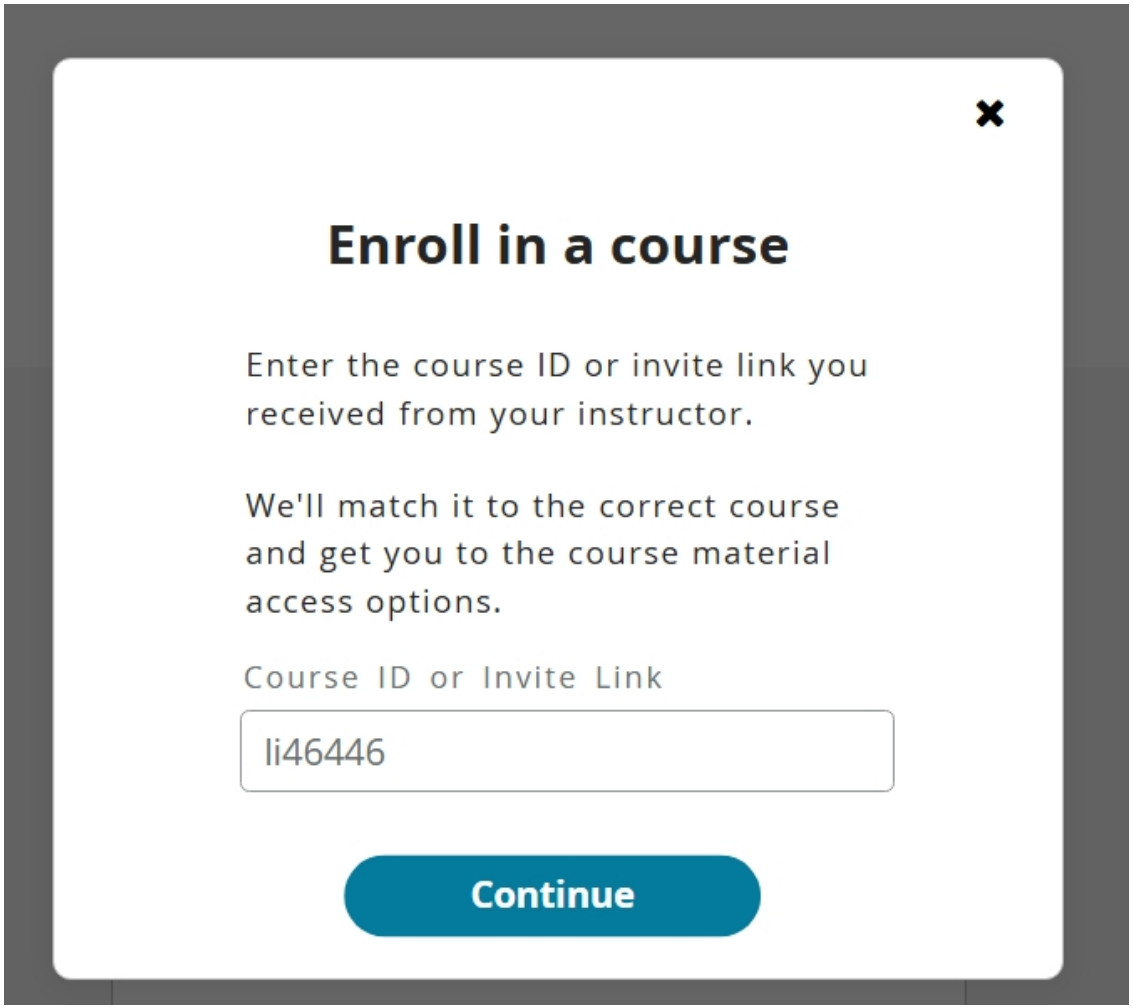
Course Log

05/20	Review of basic concepts in statistics and probability, normal distribution	First day of Summer 1
05/21	Normal distribution, sampling distributions	
05/22		
05/26	Holiday (Memorial Day), classes suspended	
05/27		Last day to drop without 'W' grade; HW1 due at 11:59pm

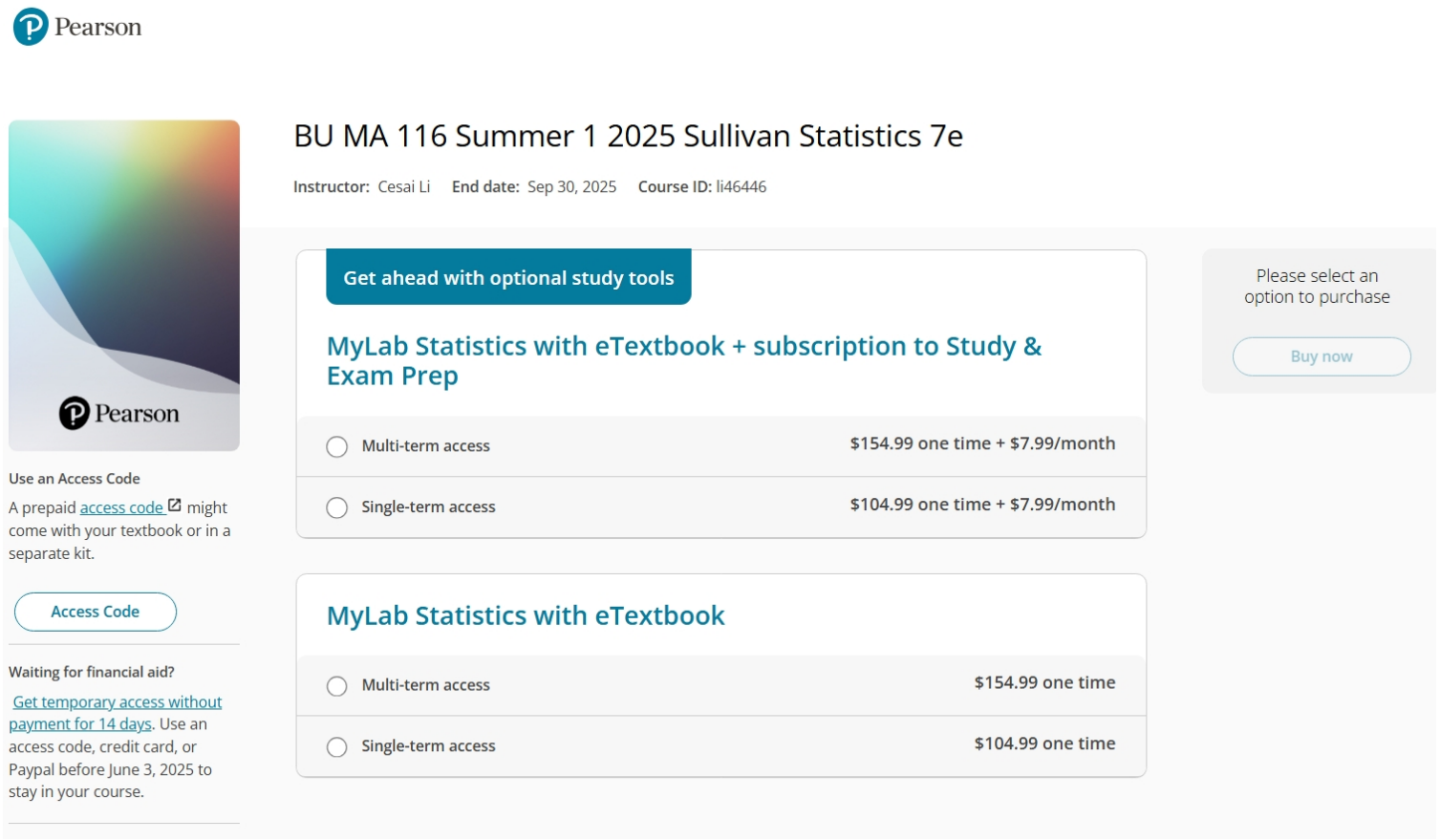
MyLab Statistics Registration



MyLab Statistics Registration

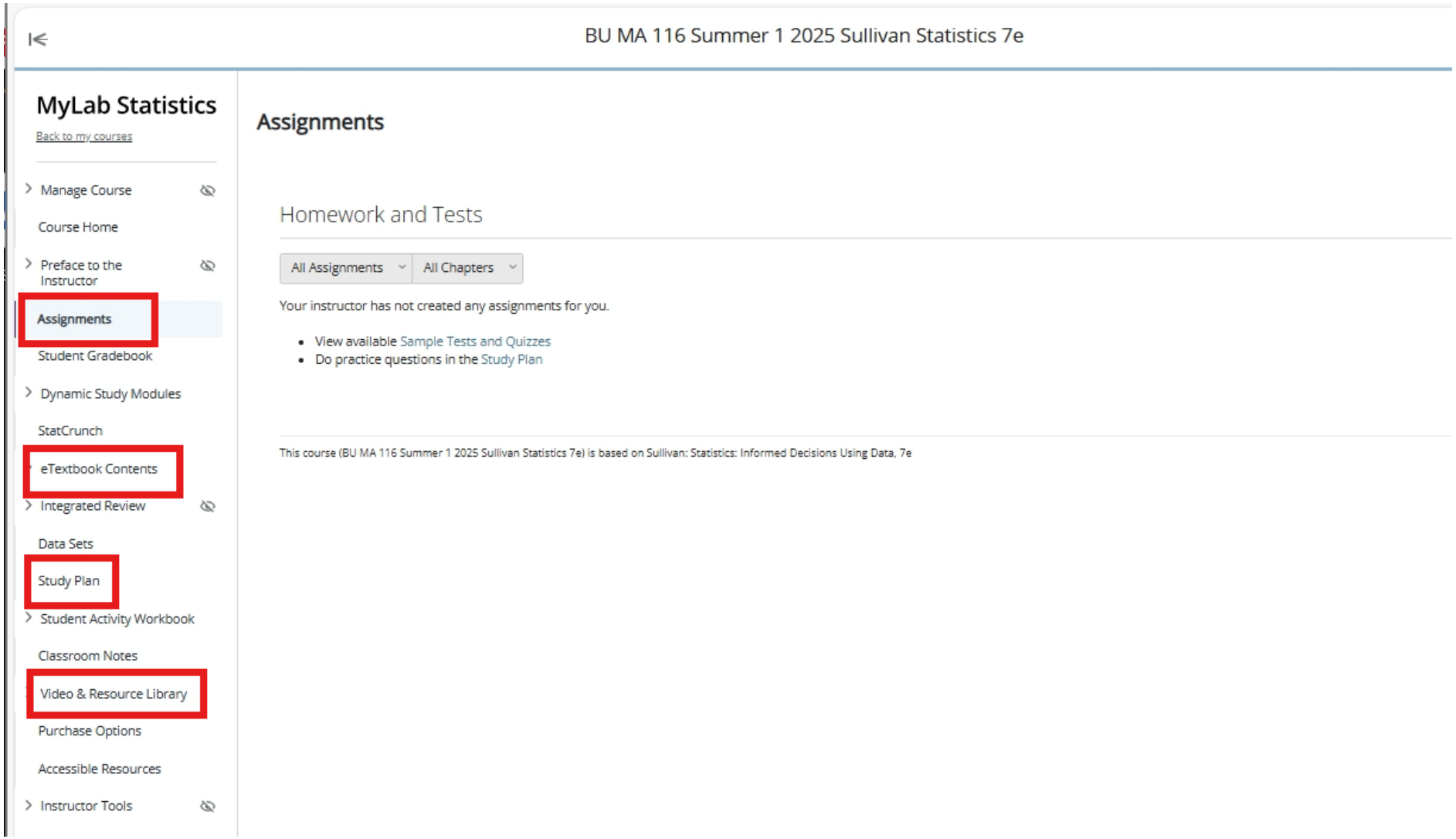


MyLab Statistics Registration



RMK. There is an option for **temporary access**. Use this option if you may drop this course. If you want to purchase a physical copy but are unsure about edition numbers, etc. Talk to me or send me an email. If you purchased a wrong edition textbook, let me know immediately.

MyLab Statistics resources



Plan of the course

Today's content

Review of basic statictical concepts and probability, up to continuous random variable and some normal distribution.

Part I	(2)	Reviewing Normal Distribution and Sampling Distributions
Part II	(6)	Confidence Interval, Tests of Hypothesis (CH. 9, 10, 11)
Part III	(5)	Regression (simple linear/multiple) (CH. 14)
Part IV	(6)	Selected topics from Chapter 12, 13, 15.

RMK. We are not going to study every concepts in a given chapter. In particular, in Part II we are going to skip some topics.

This course focuses on the idea of **statistical inference** in Statistics. We wish to use sample numerical descriptive measures to make inferences about the corresponding measures for a population. *Gain info about population from analyzing a sample.*

RMK. Quantitative data set (eg. age) vs. Qualitative (categorical) data set, a quantitative data set can be made into a qualitative data set by quartiles, for example.

Poll activity 1

Test poll

When poll is active respond at **PollEv.com/cesaili326**

How difficult do you expect MA 116 to be for you?

1. Very Easy
2. Easy
3. Manageable
4. Hard
5. Very Hard

Powered by Poll Everywhere

Poll activity 2

Which of the following data sets are generally considered qualitative/categorical data sets? Choose all that apply.

1. Amount of dollars in 100 person's bank account
2. A data set consisting of 200 randomly selected BU undergraduate students and which program they are in
3. A data set consisting of all BU undergraduate students and which program they are in
4. Size of all cells in a cell culture sample that contains $\sim 10^5$ cells
5. Size of 5 cells carefully measured and recorded from the same cell culture sample sample

RMK. It is not true to say a data set is either quantitative or qualitative–It often depends on how you interpret the data set.

Poll Activity 3

Suppose we want to gain information about the size of cells in a cell culture sample that contains $\sim 10^5$ cells. We measure 100 cells' sizes and produced a data set of 100 data points. What terminology best describe this data set? Choose all that apply.

1. Population
2. Variable
3. Sample
4. Sample space
5. A single outcome of a (statistical) experiment
6. A collection of 100 outcomes of a (statistical) experiment

RMK. 'Sample' and 'Experiment' are statistical terminologies and biological terminologies at the same time, with different meanings.

Poll Activity 4

Suppose the average/mean of the 100 measured cell sizes is 10.15 micrometers. Which of the following statements are true? Select all that apply.

1. If the 100 cells are randomly selected, we may make inference and claim that the cell size of any cell in this cell culture sample is about 10.15 micrometers.
2. If the 100 cells are randomly selected, we may make inference and claim that the average cell size of this cell culture sample is about 10.15 micrometers.
3. We can not make such claims. The collection of sizes of all cells is our population. In order to get information of a population, we have to measure the sizes of all cells in this culture sample.

numerical methods describing quantitative data set

1. **Central tendency** means the tendency of the data to cluster, or center, about certain numerical values. Examples:
   - Sample **mean** which is $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$.
   - Sample **median** M which is the middle number when the measurements are arranged in ascending or descending order (or the mean of the middle two numbers if n is even).
2. **Variability** measures the spread of the data.

RMK. n is the size of our sample. Let $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ denote the sample mean. It's important to clearly distinguish between the sample mean $\overline{x}$ and the **population mean** μ.

Measure of variability of a quantitative data set, examples

1. Sample **range** = largest measurement - smallest measurement
2. Sample **variance** $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$.
   Sample **standard deviation** $s = \sqrt{s^2}$.

Measure of relative standing: How one data pt compare to the whole sample

1. Quartiles: lower quartile Q_L, midddle quartile M (median), upper quartile Q_U.
2. Sample **z-score** $z = \dfrac{x - \overline{x}}{s}$ where $\overline{x}$ is the sample mean and s is the sample standard deviation.

An observation/measurement that is unusually large or small relative to other values in data set is called an **outlier**. Detect outlier: box plots*diagram*, hinges, IQR=$Q_U - Q_L$, inner/outer fences.

Summarize

Suppose our population is quantitative. In this setting, a quantitative data set might be our population or a sample from this population.

Measure of central tendency of population	Population mean μ, Population median η
Measure of variability of population	Population range, Population variance σ^2, Population standard deviation σ
Measure of relative standing of one data point x in population	Which quartile it lies in, Population z-score: $z = \dfrac{x - \mu}{\sigma}$

Summarize

Suppose we have a population but we do not know its mean μ, median η, range, variance σ^2, or standard deviation σ. We want to make inferences about such information from a sample. How to describe a sample as a quantitative data set?

Measure of central tendency of a sample	Sample mean $\overline{x}$, Sample median M
Measure of variability of a sample	Sample range, Sample variance s^2, Sample standard deviation s
Measure of relative standing of one data point x in this sample	Which quartile it lies in, Sample z-score: $z = \dfrac{x - \overline{x}}{s}$

Important concepts: Population Parameter vs. Sample Statistic.

	Population Parameter	**Sample Statistic**
Mean	μ	$\overline{x}$
Median	η	M
Variance	σ^2	s^2
Standard Deviation	σ	s

Poll Activity 5

Let our population be the ages of a group of 6 people: $\{24, 14, 8, 11, 23, 80\}$. From this population we choose a sample of size 5: $\{24, 14, 8, 11, 23\}$. Which of the following statements are correct? Choose all that apply.

1. $\overline{x} = 16$
2. $\mu = 16$
3. $\mu = 160/6 \cong 26.67$
4. $\overline{x} = 160/6 \cong 26.67$

Write the numbers down on a piece of paper! The next question uses the same data.

Poll Activity 6

Which of the following statements are correct? Choose all that apply.

1. $s^2 = \frac{1}{5}((24-16)^2+(14-16)^2+(8-16)^2+(11-16)^2+(23-16)^2) = 41.2$
2. $s^2 = \frac{1}{6}((24-16)^2+(14-16)^2+(8-16)^2+(11-16)^2+(23-16)^2) \cong 34.33$
3. $s^2 = \frac{1}{4}((24-16)^2+(14-16)^2+(8-16)^2+(11-16)^2+(23-16)^2) \cong 51.5$
4. $\sigma^2 = \frac{1}{5}((24-16)^2+(14-16)^2+(8-16)^2+(11-16)^2+(23-16)^2+(80-16)^2) = 4302/5 = 860.4$
5. $\sigma^2 = \frac{1}{6}((24-26.67)^2+(14-26.67)^2+(8-26.67)^2+(11-26.67)^2+(23-26.67)^2+(80-26.67)^2) \cong 3619.33/6 \cong 603.22$
6. $\sigma^2 = \frac{1}{5}((24-26.67)^2+(14-26.67)^2+(8-26.67)^2+(11-26.67)^2+(23-26.67)^2+(80-26.67)^2) \cong 3619.33/5 \cong 723.87$

Short review of probability

Terminology

1. A **sample point** of an experiment is a single outcome of the experiment.
2. The **sample space** of an experiment is the collection of all its sample points.
3. The **probability** of a sample point is a number between 0 and 1 that measures the likelyhood of getting this outcome when this experiment is performed, denoted p_i is the sample point is denoted by i.

In many cases we do not know p_i and we can't get to know the exact value of p_i by doing this experiment. However, we can repeat this experiment sufficiently many times to approximate p_i by $\frac{n_A}{n}$.

If our sample space is finite, the probabilities of sample points in this sample space sum up to 1.

Terminology

1. An **event**, denoted by A, is a specific collection of sample points.
2. The **probability of an event** $P(A)$ is calculated bysumming the probabilities of the sample points in A.
3. For A, B two events of a same experiment, the **union** of A and B, denoted by $A \cup B$, is the collection of sample points that lie in either A or B. We say event $A \cup B$ occurs at a single performance of this experiment, if either event A or event B occurs at this performance.
4. For A, B two events of a same experiment, the **intersection** of A and B, denoted by $A \cap B$, is the collection of sample points that lie in both A and B. We say event $A \cap B$ occurs at a single performance of this experiment, if both event A and event B occur at this performance.
5. The **complement** of an event A, denoted by A^c, is the event that A does not occur. Then $P(A^c) = 1 - P(A)$.

Calculate probability

Rule. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

We say two events A and B of a same experiment are **mutually exclusive** if $P(A \cap B) = 0$. I.e. If one occurs then the other does not occur. If two events A and B are mutually exclusive, then by the rule we have $P(A \cup B) = P(A) + P(B)$. If two events A and B are not mutually exclusive, then it is false to claim $P(A \cup B) = P(A) + P(B)$.

Conditional probability

Fix an experiment. Let A and B be two events of this experiment such that $P(B) \neq 0$. Let's now **assume** that event B would occur in a performance of this experiment. Then, what is the likelyhood that event A occurs in this performance?

The likelyhood that event A occurs in a performance, assuming event B would occur in this performance, is called the **conditional probability** of A with respect to B, and is denoted $P(A \mid B)$.

Formula. $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$. (Defined only if $P(B) \neq 0$.)

Theorem.

If $P(A)$, $P(B)$ are both nonzero, then
$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$.

This theorem, deducted from the previous formula, allows us to calculate $P(A \mid B)$ if we know $P(B)$, $P(B \mid A)$, $P(A)$.

End of probability review

We say events A and B of a same experiment of nonzero probability of occuring are **independent events** if $P(A \mid B) = P(A)$. Then from the formula $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$ we deduct that $P(B \mid A) = P(B)$.

The idea is that if two events are independent, then assuming one occurs does not affect the likelyhood of the other occurs.

Theorem.

If events A and B of a same experiment of nonzero probability of occuring are independent events, then $P(A \cap B) = P(A)P(B)$.

This theorem is deducted from the previous theorem and our assumption of independence. Probability review content won't be on quiz/exam

Discrete random variables: Terminology

1. Fix an experiment and let a variable x be associated to outcomes of this experiment. If performances of this experiment give numerical results, we say x is a **random variable**.
2. Two types of random variable: discrete and continuous. A random variable of an experiment is **discrete** if any sample point i has a well-defined probability p_i such that $p_i \geq 0$ and $\sum p_i = 1$ where the summation is taken over our sample space. (We can think $p_i > 0$ for all sample points i if we assume that our sample space consists of *possible* outcomes.) A random variable of an experiment is **continuous** if $p_i = 0$ for any sample point i. Moreover, there can be **mixed** random variable where the result of the experiment is mixed with discrete and continuous outcomes, but we do not discuss it here.

Discrete random variable: concepts and formulas

1. The **probability distribution** of a discrete random variable specifies the probability p_i of each sample point i in the sample space.
2. The **mean** (or **expected value**) of a discrete random variable x is $\sum i \times p_i$, where the summation is taken over our sample space. The mean/expected value is denoted by μ or $E(x)$.
3. The **variance** of a discrete random variable x is calculated by $E[(x-\mu)^2] = \sum (i-\mu)^2 p_i = \sum i^2 p_i - \mu^2$, and is denoted by σ^2.
4. The **standard deviation** of a discrete random variable is equal to the square root of the variance and is denoted by σ.

Formulas of discrete random variable won't be on quiz/exam

Binomial random variable

This topic won't be on quiz/exam **Binomial random variable** is an example of discrete random variables where the experiment consists of n identical trials such that each trial has two outcomes S (for success) and F (for failure). The possibility of getting S in a trial is p, a fixed number $0 \leq p \leq 1$. Then the possibility of getting F in a trial is $1 - p$, which we denote by q. We further require that the n trials are independent; that is, the result of a later trial does not depend on the result of a previous trial. We let an outcome of our experiment be the total number of S's in our n-trials. Let a variable x be associated to outcomes of this experiment. Then we say such x is a binomial random variable.

Such experiment is called a binomial experiment. Be careful that the sample space of a binomial experiment is $\{0, 1, \ldots, n\}$, and the sample space is not $\{S, F\}$.

Binomial random variable: results

Theorem. (Binomial probability distribution.)

Suppose x is a binomial random variable in a binomial experiment, which is to say x associates to the number of S's we get if we perform a trial n times. Then $p(x) = \binom{n}{x} p^x q^{n-x}$ for $x = 0, 1, ..., n$.

Lemma. Binomial expansion $(a+b)^c = \sum_{i=1}^{c} a^i b^{c-i}$.

Recall. $\binom{n}{n} = 1$, $\binom{n}{0} = 1$, $\binom{n}{m} = 0$ if $m < 0$ or $m > n$.

We may deduct the following results:

Suppose x is a binomial random variable, n is the number of trials in an experiment, and p is the possibility of getting S in a trial.

1. Mean/expected value $E(x) = np$
2. Variance $\sigma^2 = npq$
3. Standard deviation $\sigma = \sqrt{npq}$

Continuous random variable

Let x be a random variable associated to the outcomes of an experiment. We say x is a continuous random variable if the sample space is some intervals.

A characteristic of a continuous random variable is that for any sample point i in the sample space, the probability of getting i is exactly zero, but it doesn't mean it's impossible to get i.

Be careful. If we fix a measurement accuracy, then an experiment that would normally give rise to a continuous random variable often instead gives rise to a discrete one. Example: Consider an experiment: I throw a ball and measure how far away it lands. Suppose I can never throw the ball beyond 2 meters. If I fix the measurement accuracy by using a tool that measures to the nearest $1/10$ meter, then the variable x is discrete and the sample space is $\{0.0m, 0.1m, 0.2m, \ldots, 1.9m, 2.0m\}$. In this class we ignore this issue—for example, by assuming that the tool can measure any real distance r meters such that $0 \leq r \leq 2$. Under this assumption, x is treated as a continuous random variable.

Idea

1. Measurement accuracy does **discretize** a continuous variable in practice.
2. We think a continuous variable to be more associated to the underlying abstract quantity (distance, in this example), rather than the experiment outcomes.

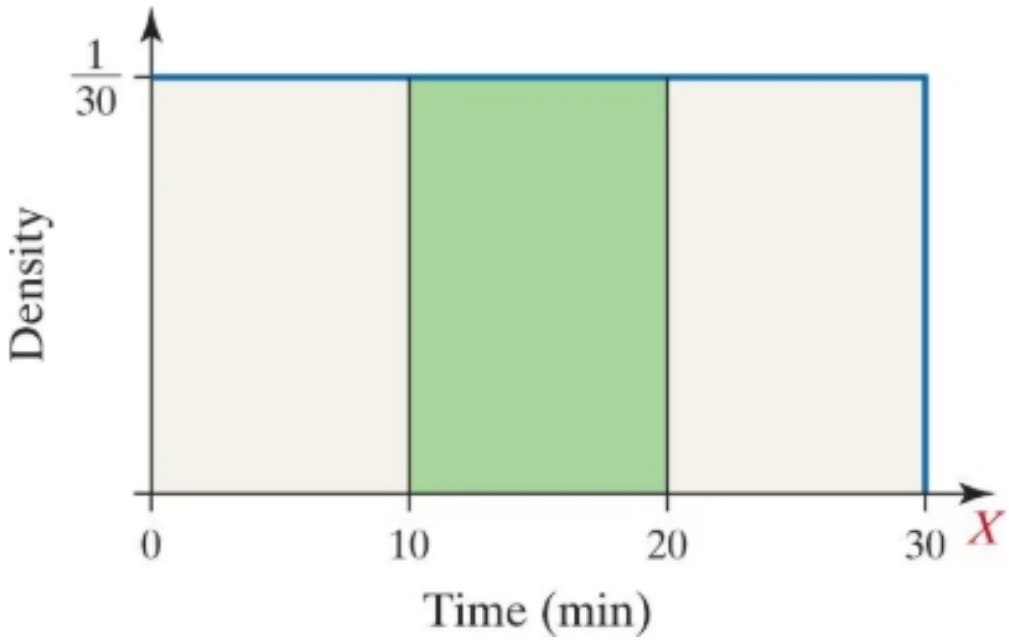
Since $p_i = 0$ for any sample point i, it only makes sense to talk about the probability that one experimental outcom r falls into some interval, say, $1 \leq r \leq 2$.

Probability density function

For a continuous random variable x there associates a **probability density function** (or called a **probability distribution**) $f(x)$ such that for any sample point i we have $0 \leq f(i) \leq 1$, and the total area under $f(x)$ is exactly 1. The probability of an experimental outcom r lands into $a \leq r \leq b$ is exactly the area under the graph of $f(x)$ from $x = a$ to $x = b$.

Examples of continuous random variables

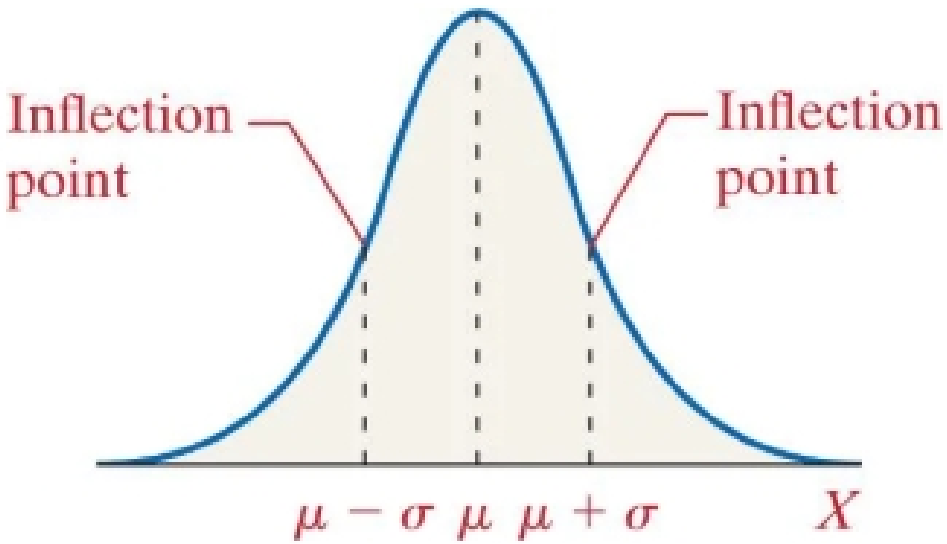
Example 1. Uniform distribution. This topic won't be on quiz/exam $f(x)$ is defined on $c \leq x \leq d$ and has a constant value $\frac{1}{d-c}$ whenever it's defined. **formulas: mean, standard deviation, probability p(a ≤ x ≤ b).**



Example 2. of continuous random variables Normal distribution / bell curve.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-1/2[(x-\mu)/\sigma]^2)$$

A (hypothetical) variable x is called a **normal variable** if its probability distribution is exactly this $f(x)$. If $\mu = 0$ and $\sigma = 1$ we say such a normal variable is a **standard normal variable** and $f(x)$ is a standard normal distribution. We often denote a standard normal variable by z.



Key questions to ask:

1. If we know a data set is (approximately) a normal distribution, how can we extract information we need from the normality?
2. Given a data set, how can we tell if it can be approximated by a normal distribution?

To answer question 1, be careful with **Population** vs. **Sample** vs. **Model**

Population vs. Sample vs. Model

Suppose we perform an experiment and get a quantitative data set (assume it's a sample). We plot the sample points and notice the graph look *roughly* like a straight line, a quadratic curve, or graphs of some other simple nice functions $F(x)$, we can probably make use of the nice properties of $F(x)$ to obtain useful information about our data set.

1. First, we need to make sure that it is reasonable to approximate our data set by $F(x)$. There are methods with different strength, from looking at graphs or comparing means/median/standard deviation to performing careful numerical analysis.
2. Second, we obtain the [means/median/standard deviation/other info] from our nice simple function $F(x)$, and claim that the [means/median/standard deviation/other info] of our data set is approximately that of $F(x)$.

If we claim that the probability distribution of our variable x **can be approximated** by a nice mathematical function $F(x)$, then we call $F(x)$ a **model**.

Poll Activity 7

The function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-1}{2}[(x-\mu)/\sigma]^2\right)$$

is called the Normal distribution. It is a sort of

1. Sample
2. Population
3. Experiment outcome
4. Variable
5. None of the above

This question will not be on quiz/exam.

Properties of a bell curve

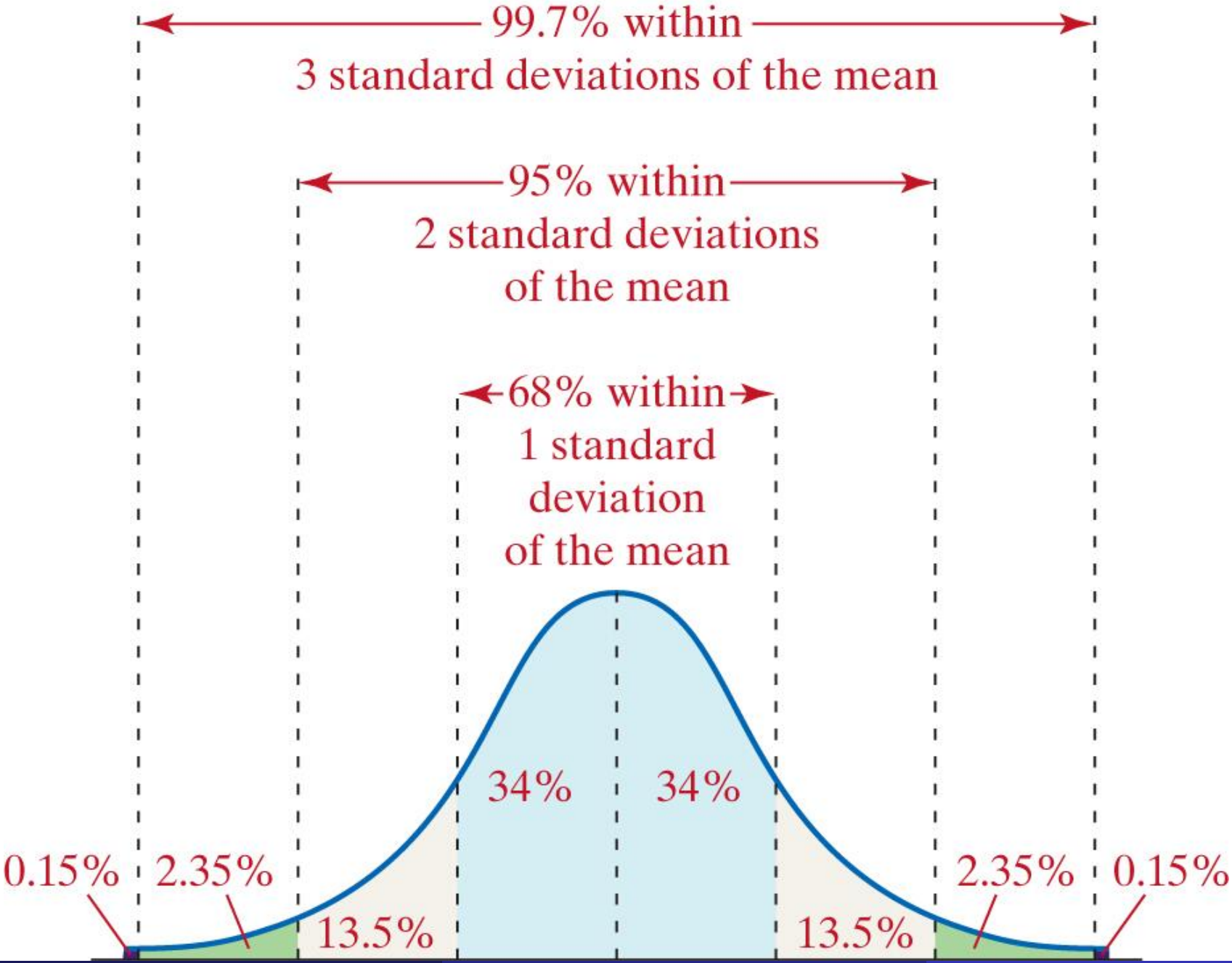
1. The bell curve is symmetric about its mean μ.
2. The area under the bell curve is 1.
3. The curve gets closer and closer to 0 as x increases/decreases away from μ, but never really gets to 0.

7. The Empirical Rule:

   - Approximately 68% of the area under the normal curve is between $x = \mu - \sigma$ and $x = \mu + \sigma$;
   - approximately 95% of the area is between $x = \mu - 2\sigma$ and $x = \mu + 2\sigma$;
   - approximately 99.7% of the area is between $x = \mu - 3\sigma$ and $x = \mu + 3\sigma$.

4.

Normal Distribution



Poll Activity 8

Suppose x is a normal random variable with mean μ and standard deviation σ. What is $P(x < \mu - \sigma)$?

1. 34%
2. 50%
3. 13.5%
4. 16%
5. None of the above
6. Depends on μ and σ