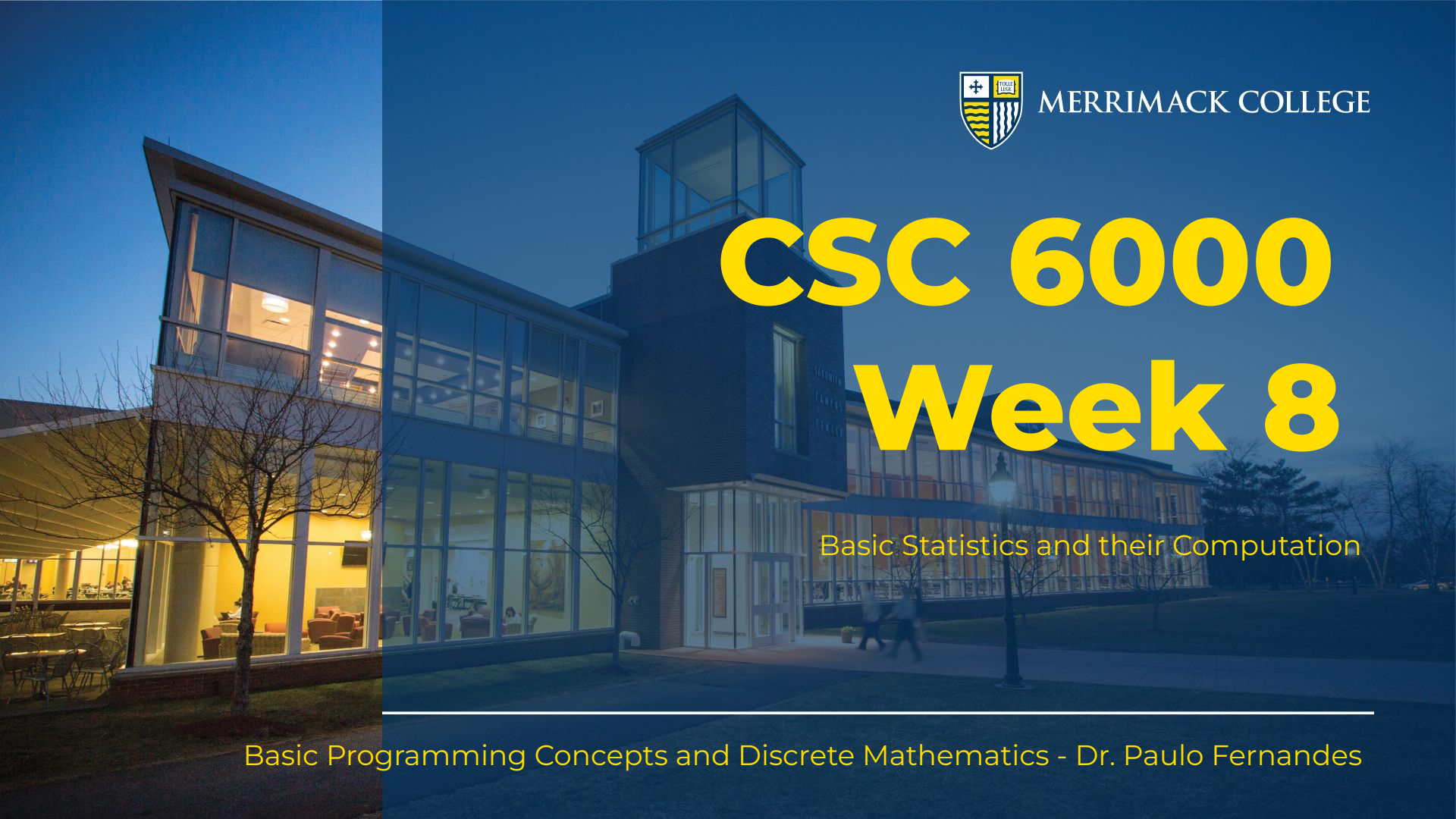


MERRIMACK COLLEGE

CSC 6000 Week 8

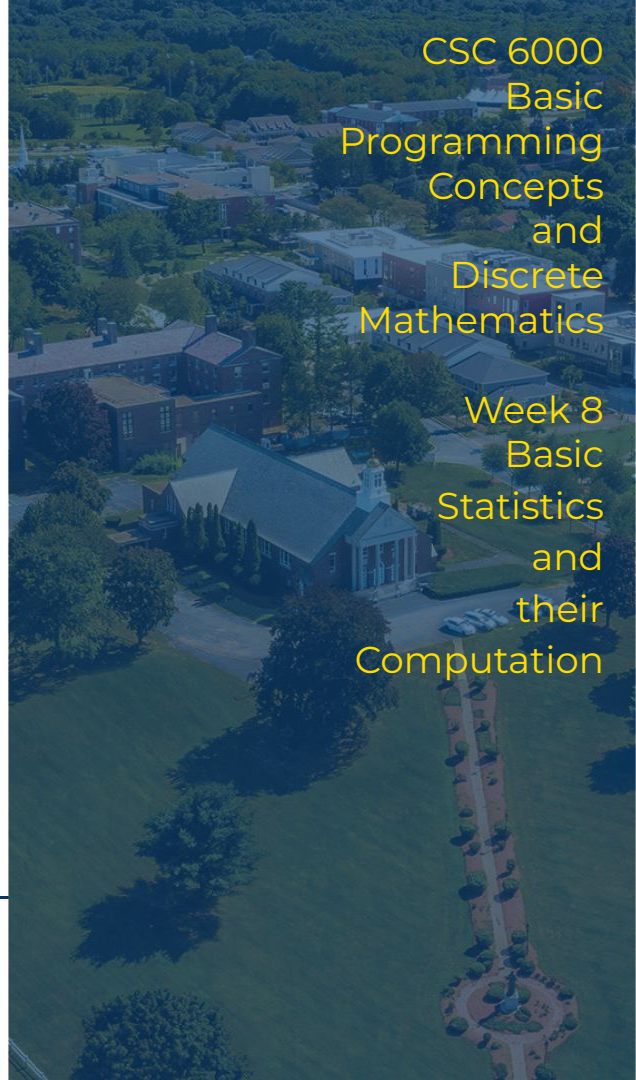
Basic Statistics and their Computation

Basic Programming Concepts and Discrete Mathematics - Dr. Paulo Fernandes

Presentation Agenda

Week 8

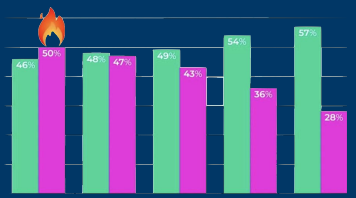
- Statistics
  a. Statistics Basics
  b. Sampling
  c. Basic Measures
- Wrapping the course up
  a. This Week's tasks
  b. The Final Exam
  c. Evaluation
  d. The Next Steps

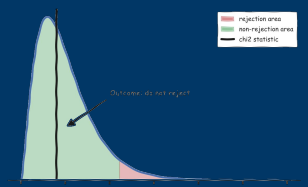


Statistics

Facts are stubborn things, statistics are more pliable.

Mark Twain





There's a large number of Probability & Statistics courses available, which helps to build the confusion that the two areas are the same.

Actually, probabilities and statistics are components of the same things, just like sowing and reaping.

We will try to set these concepts apart by discussing:

- **Statistics Basics**,
- Sampling,
- Basic Measures.

Statistics

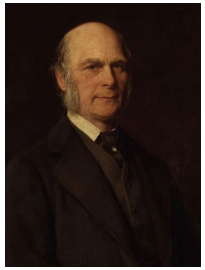
It started as an application of Probability theory, and yet the term "statistics" was written by Ghilini in 1647. Statistics became an independent area of Mathematics by the beginning of the XX Century, mostly as an effort of Galton and Pearson, establishing the basics of the area including basic measures like standard deviation. It is notable, however, that since the beginning of modern statistics, the concepts were almost always applied ones, not theoretical.

This, probably, is the reason why statistics got a bad reputation (such as Mark Twain's opinion of it).

Girolamo **Ghilini**
1589 - 1632 - 1668

Jacob **Bernoulli**
1655 - 1684 - 1705

Carl Friedrich **Gauss**
1777 - 1840 - 1855



Francis **Galton**
1822 - 1882 - 1911

Karl **Pearson**
1857 - 1912 - 1936

Wikipedia: Statistics.

Statistics and Probabilities

What happened?

- While:
  - in Probabilities you are trying to guess what will happen based on the likelihood of events to occur, which is fundamentally based on assumption and previous experiences,
  - in Statistics you are trying to figure out what happened in a specific phenomenon, crunching the data which you've gathered by observing some aspects of the phenomenon.
- We employ Probabilities while planning (before execution);
- We employ Statistics while analyzing (after execution).

Text: Probability vs Statistics.

Statistics Goals

- Understanding
  - Find the proper measurements to summarize large amounts of data:
    - Data Transformations, Visualization;
  - Look for correlations among phenomena:
    - Identify Causes and Consequences, Detect Dependency;
- Credibility
  - Find numerical reasons to trust the measurements:
    - Statistical Relevance, Accountability;
  - Assign numerical values to back up opinions:
    - Assessment, Evaluation.

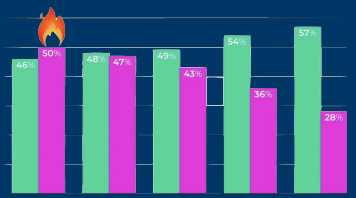
MERRIMACK COLLEGE

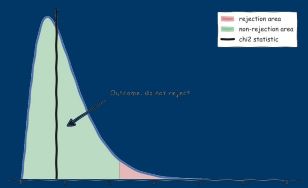
Video: What is Statistics?.

Statistics

Facts are stubborn things, statistics are more pliable.

Mark Twain





The number of Probability & Statistics courses around help to build the confusion that the two areas are the same.

Actually, probabilities and statistics are components of the same things, just like sowing and reaping.

We will try to set these concepts apart by discussing:

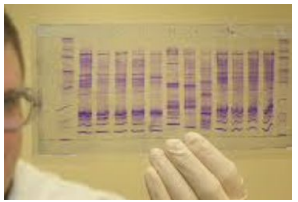
- Statistics Basics,
- **Sampling**,
- Basic Measures.

Sampling

A **population** is composed by **individuals**.
An **individual is anything**, including, but not only, people.

Collecting data

- Sometimes you have a vast amount of data.
  - Let's call it a **statistical population**, which is all data of a given phenomenon;
- If this statistical population is too big to be analyzed or even observed, we may need to set aside a portion of it to work with:
  - Let's call it a **statistical sample**, which is a (hopefully) representative subset of the statistical population.



This is traditionally what we call sampling - the techniques to pinpoint a subset of a population to be the sample to be analyzed.

Wikipedia: Sampling (statistics).

Sampling

A **sample** is composed by **individuals**.
An **individual** is represented solely by its data.

Collecting data

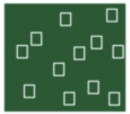
- Thanks to the Internet and modern technologies, the means of collecting and analyzing large amounts of data have increased significantly.
  - often all data of a given population can be collected.

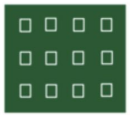
- Even in such situations it is also possible to refer to it as a sample, since even when collecting all data, this can be seen as a **sample of possible data**.
  - For example, the universe that exists was just one possible instance, thus a sample of a theoretical multiverse.

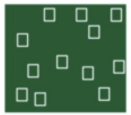
Wikipedia: Sampling (statistics).

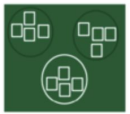
Sample Characterization

- Sampling Assumptions
  - Population Profile - the population and individuals' info
- Sampling Frame
  - Size - how many are there
  - Particular Characteristics - how are they distributed
  - Meta Information - information about the information
- Sampling Method
  - Random - just pick some
  - Convenience - just pick what is possible/easy
  - Systematic - follow a pattern
  - Cluster - pick some packs
  - Stratified - try to reproduce the population







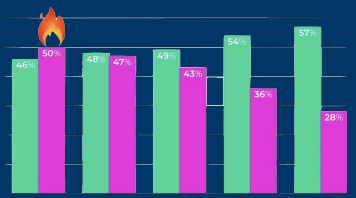


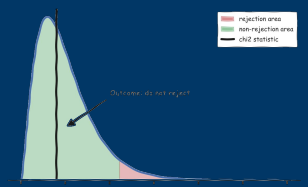
Video: Sampling.

Statistics

Facts are stubborn things, statistics are more pliable.

Mark Twain





The number of Probability & Statistics courses around help to build the confusion that the two areas are the same.

Actually, probabilities and statistics are components of the same things, just like sowing and reaping.

We will try to set these concepts apart by discussing:

- Statistics Basics,
- Sampling,
- **Basic Measures**.

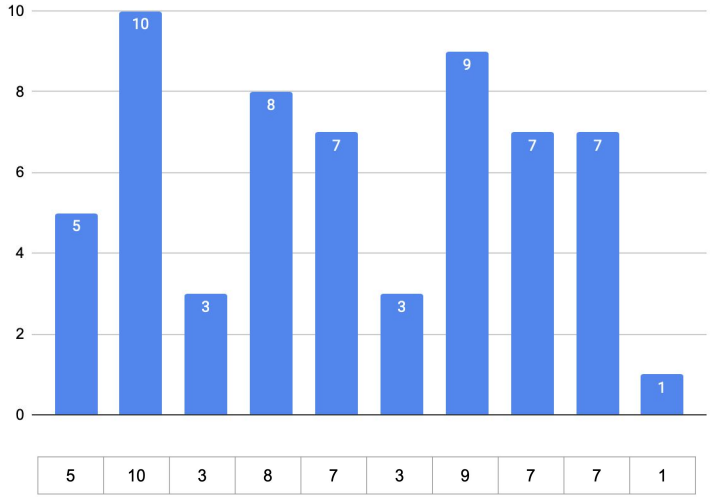
Basic Measurements

Given a sample composed of ***n*** values:

To characterize a given sample of ***n*** elements we can:

- write down all elements or ...
- compute some measurements that give the idea how the elements are:
  - the average (or mean);
  - which element is the most common one in the sample;
  - which element is the smallest one;
  - which element is the largest;
  - how spread out they are.

n = 10



Wikipedia: Statistical Dispersion.

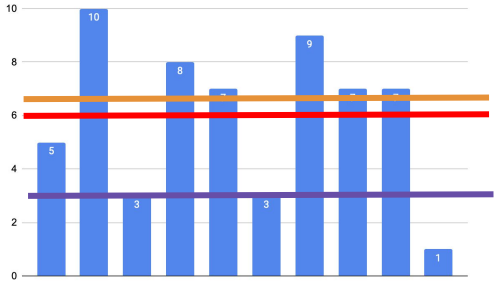
Basic Measurements

Given a sample composed by ***n*** values.

- Mean or Arithmetic Average
  - The sum of all elements of the sample divided by their number;
- Range
  - The minimum and maximum value within the sample's elements;
- Mode
  - The most frequent value within the sample's elements;
- Median
  - If sorted, the one (or average of the two) in the middle of the sample;
- Variance
  - The variability of the value with respect to the mean;
- Standard Deviation
  - The dispersion of values of the sample with respect to the mean.

Video: Finding mean, median, and mode.

Mean



- Mean or Arithmetic Average
  - The sum of all elements of the sample divided by their number.

A =	5	10	3	8	7	3	9	7	7	1
w =	2	3	1	3	2	1	2	3	1	3

- ***n = 10***
- average of $A = \frac{\sum_{i=0}^{9} a_i}{10}$
- $\bar{A} = \frac{\sum_{i=0}^{9} a_i}{10} = \frac{60}{10} = 6$

Arithmetic Average

$$\bar{A} = \frac{\sum_{i=0}^{n-1} a_i}{n}$$

- Harmonic Average ***H = 3.8003***

$$H = \frac{n}{\sum_{i=0}^{n-1} \frac{1}{a_i}}$$

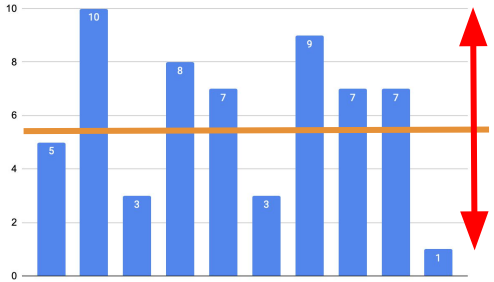
- Weighted Average

$$\bar{A}_w = \frac{\sum_{i=0}^{n} w_i a_i}{\sum_{i=0}^{n} w_i}$$

$\bar{A}_w$ = 6.3333

Wikipedia: Average.

Range



- Values covered by the sample
  - The minimum and maximum value within the sample's elements.

A =	5	10	3	8	7	3	9	7	7	1

- ***n = 10***
- ***min(A) = 1***
- ***max(A) = 10***
- ***range(A) = 1 - 10***

Range

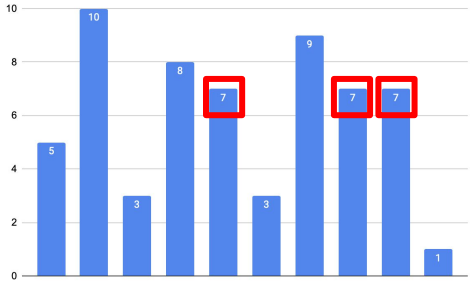
- ***min - max***

Mid-Range

- ***(max(A) + min(A)) / 2***
  - ***(1 + 10) / 2***
    - ***5.5***

Wikipedia: Mid-range.

Mode



- The highest probability in the sample
  - The most frequent value within the sample's elements.

A =	5	10	3	8	7	3	9	7	7	1

- ***n = 10***
- ***the element 7 appears three times***
- *the element 3 appears two times*
- ...

Mode

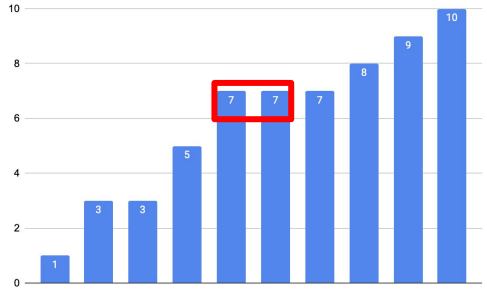
- ***mode(A) = 7***

Number of Modes

- ***Unimodal*** - 1 mode
- ***Bimodal*** - 2 modes
- ***Multimodal*** - more than 2 modes

Wikipedia: Mode (statistics).

Median



- A less stiff measure than the mean
  - If sorted, the element (or average of two) in the middle of the sample.

A =	1	3	3	5	7	7	7	8	9	10

- ***n = 10***
- there is a even number of elements
- if sorted, the two in the middle are ***7*** and ***7***
- ***(7 + 7) / 2 = 7***

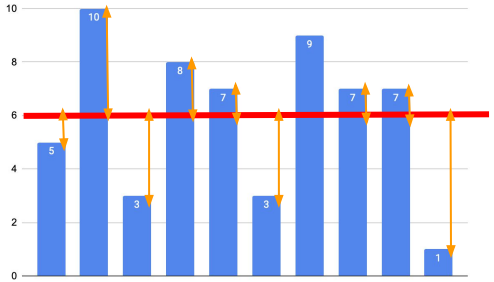
Median

- ***median(A) = 7***

There are efficient algorithms to compute the median without sorting the sample, for example, QuickSelect

Wikipedia: Median.

Variance (of a population)



- Basic dispersion of elements
  - The variability of the value with respect to the mean.

A =	5	10	3	8	7	3	9	7	7	1
	1	4	3	2	1	3	3	1	1	5

- ***n = 10***
- ***$\bar{A}$ = 6***
- $Var(A) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})^2}{n}$

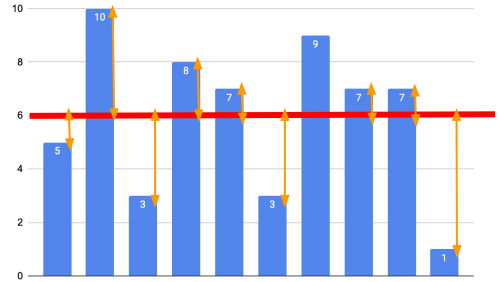
Variance

- ***var(A) = 7.6***

The variance is the average of the square of the difference between each element and the mean.

Wikipedia: Variance.

Standard Deviation (population)



- A proportional version of the variance
  - The dispersion of values of the sample with respect to the mean.

A =	5	10	3	8	7	3	9	7	7	1
	1	4	3	2	1	3	3	1	1	5

- ***n = 10***
- ***$\bar{A}$ = 6***
- $\sigma(A) = \sqrt[2]{\frac{\sum_{i=1}^{n}(a_i - \bar{A})^2}{n}}$

Standard Deviation

- ***σ(A) = 2.7568***

The square root of the variance.

MERRIMACK COLLEGE

Wikipedia: Standard Deviation.

Wrapping the course up

... and now the end is near.

Paul Anka (My Way lyrics)

Here we are, eight weeks after starting and (I hope) a lot more knowledgeable about Basic Programming and Discrete Mathematics.

You have learned a considerable amount of new concepts, or at least revisited concepts seen before.

Let me talk to you about what comes now and beyond:

- This Week's Tasks
- The Final Exam,
- Evaluation,
- The Next Steps.

This Week's tasks

- You have to take Quiz #8 as soon as possible (deadline is this Friday).
- Post in the discussion if the course has met your expectations.
- You have to make sure all assignments from the previous weeks are submitted, as after this Friday they will not be considered anymore.

- The final exam is available for you to take just after the last Live Session (it opens on Thursday, 8:00pm EST) and you have to take it by this Sunday at 11:59pm.

One more push and you are done!

Eighth Quiz - Week 8 - Q#8

- The eighth quiz in this course covers the topics of Week 8;
- The quiz is available already, and it is composed of 10 questions;
- The quiz should be taken on Canvas (Module 8), and it is not timed:
  - You can take as long as you want to answer it;
- The quiz is open book, open notes, and you can even use any language interpreter to answer it;
- Yet, the quiz is evaluated and you are allowed to submit it only once.

Your task:

- Go to Canvas, answer the quiz and submit it within the deadline:
  - The deadline for the quiz is this Friday (yes, not much time for this one).

This quiz counts towards the Quizzes grade.

Post Discussion - Week 8 - D#8

Post in the discussion if the course has met your expectations:

- You can mention topics you expected to see and if they were covered in the level of detail you expected.
- You can also mention topics that you didn't expect to see, but you were positively or negatively surprised by.
- You can share your satisfaction comparing the amount of time you spent and how much you have learned.

Your task:

- Post your discussion in the message board whenever you feel like, but no later than this Friday;
- Reply to the posts of your colleagues in the message board freely.

This task counts towards the Discussions grade.

The Final Exam

Before starting this course, most of you took the Mastery Skill Profiler.

- You had an hour and three quarters to answer 50 questions about the very same topics covered here in class.
- The final exam is your chance to prove to yourselves how well you did over the last eight weeks.

Now as the final task of this course, you have to take the final exam. It is very similar to the Profiler, and all the quizzes.

- The main difference is that it is timed and you have just 75 minutes to answer the 50 questions.
    - The questions might be a little, very, or not at all similar to those on the Profiler and your quizzes.

This task counts towards the Final Exam grade.

Evaluation

The evaluated tasks are weighted as such (103%):

Activity	Published grades	Percentage
Projects (7)	Evaluated from 0 to 100 each	42.0%
Quizzes (8)	Evaluated from 0 to 10 each	32.0%
Discussions (8)	Evaluated from 0 to 100 each	4.0%
Final Exam	Evaluated from 0 to 50	25.0%

All tasks are due by their stated deadline, the late penalties reduce 10% of the evaluation, plus another 2% for each full day of delay.

Final deadline on Friday of the course's last week

Evaluation

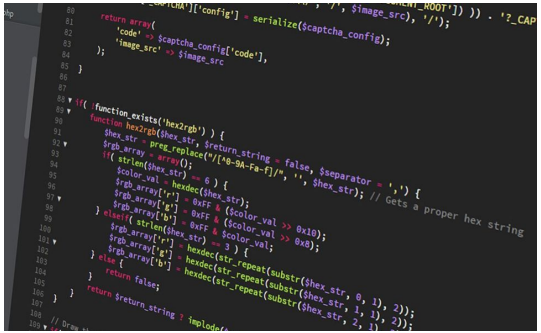
The final letter grade is computed according to this table:

A	A-	B+	B	B-	C+	C	C-	F
95 and up	90 to 94.9	87 to 89.9	83 to 86.9	80 to 82.9	77 to 79.9	73 to 76.9	70 to 72.9	69.9 and low

There are no grade D+, D, or D- in graduate courses. C- is not a passing grade for MSCS. An average grade B is required to remain in the Masters of Science in Computer Science, an average lower than that puts you in probation.

Final grades published first Monday after the course's end.

The Next Steps

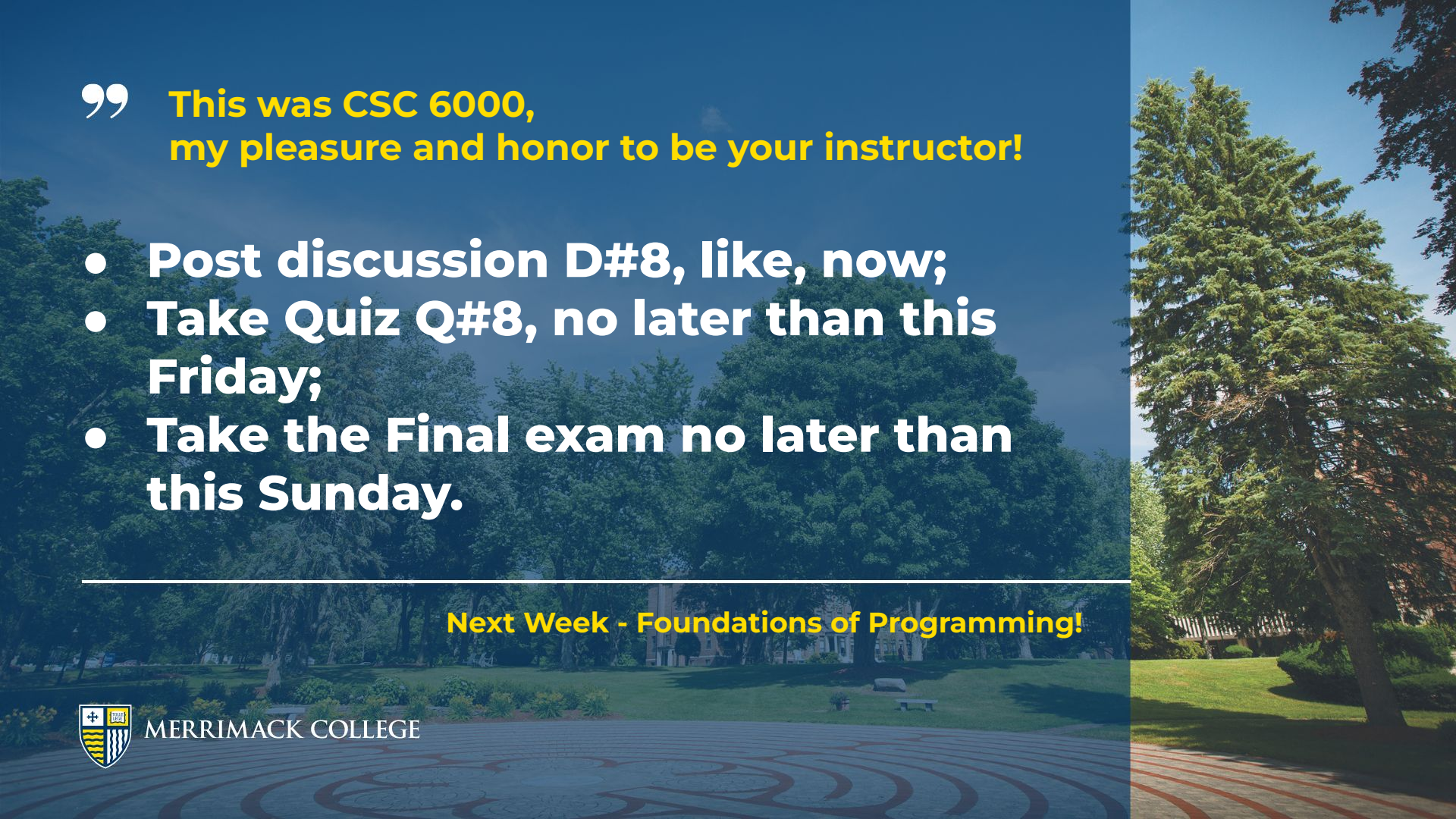


Now that you finished the course, successfully
I hope, this is your path in the MSCS Program:

- CSC 6003 Foundations of Programming
- CSC 6013 Algorithms & Discrete Structures
- CSC 6023 Advanced Algorithms
- CSC 6033 Automata, Languages, and Computability
  - DSE 6003 Data Governance and Privacy
- CSC 6301 Software Design & Documentation
- CSC 6302 Database Principles
- CSC 6303 Systems & Languages Survey
- CSC 6304 Advanced Programming Concepts
- CSC 6400 Capstone Project (optional)

MERRIMACK COLLEGE

Keep in touch with the success coach!



> This was CSC 6000,
> my pleasure and honor to be your instructor!

- Post discussion D#8, like, now;
- Take Quiz Q#8, no later than this Friday;
- Take the Final exam no later than this Sunday.

Next Week - Foundations of Programming!

MERRIMACK COLLEGE