Statistics and Machine Learning, B.S.

B.S. in Statistics and Machine Learning

Peter Freeman, *Undergraduate Program Director*
Location: Baker Hall 229
pfreeman@andrew.cmu.edu

Zach Branson, *Assistant Director of the Undergraduate Program*
Location: Baker Hall 232
zbranson@andrew.cmu.edu

Amanda Mitchell, *Lead Senior Academic Advisor*
Sylvie Aubin, *Academic Program Manager*
Location: Baker Hall 129
statadvising@andrew.cmu.edu

The Bachelor of Science in Statistics and Machine Learning trains students in statistical theory, methods, applications, and judgment, such that they can make real-world impacts. Furthermore, students gain in-depth training in machine learning and computer science, so that students not only apply state-of-the-art computational methods, but also design and analyze workflows, algorithms, and data structures to efficiently-but-reliably analyze complex data. This degree is especially relevant for students interested in the underpinnings of computational tools that make advanced statistical methods possible in practice, as well as those who want to pursue graduate programs at the intersection of computation and data science.

Curriculum

1. Mathematical Foundations (Prerequisites) 51-64 units

Mathematics is the language in which statistical models are described and analyzed, so some experience with basic calculus and linear algebra is an important component for anyone pursuing a program of study in Statistics and Machine Learning.

Complete the following:

21-090	Precalculus	10
Complete one of the following options:		
21-111	Differential Calculus	10
21-112	Integral Calculus	10
OR		
21-120	Differential and Integral Calculus	10
And one of the following four courses:		
21-256	Multivariate Analysis	9
21-259	Calculus in Three Dimensions	10
21-266	Vector Calculus using Matrix Algebra	10
21-268	Multidimensional Calculus	11
And one of the following three courses:		
21-240	Matrix Algebra with Applications	10
21-241	Matrices and Linear Transformations	11
21-242	Matrix Theory	11

Mathematical Theory

21-127	Concepts of Mathematics *	12

Notes:

- Passing the Mathematical Sciences assessment tests available during First-Year Orientation is an acceptable alternative to completing 21-090 and/or 21-120.
- It is recommended that students complete the calculus requirement during their freshman year.
- 21-266 and 21-268 are intended only for students with a very strong mathematical background.
- The linear algebra requirement needs to be completed before taking 36-401 Modern Regression.
- 21-241 and 21-242 are intended only for students with a very strong mathematical background.

*Students with little to no previous experience in theoretical mathematics and/or theoretical proofs are encouraged to consider taking 21-108 Introduction to Mathematical Concepts either prior to or with 21-127 for additional support.

2. Data Analysis 36 units

Data analysis is the art and science of extracting insight from data. The art lies in knowing which displays or techniques will reveal the most interesting features of a complicated data set. The science lies in understanding the various techniques and the assumptions on which they rely. Both aspects require practice to master.

The Beginning Data Analysis courses give a hands-on introduction to the art and science of data analysis. The courses cover similar topics but differ slightly in the examples they emphasize. 36-200 draws examples from many fields and satisfies the Dietrich College Core Requirement in Statistical Reasoning. This course is therefore required for students in the college. (Note: a score of 5 on the Advanced Placement [AP] Exam in Statistics may be used to waive this requirement). 36-220 emphasizes examples in engineering and architecture.

The Intermediate Data Analysis courses build on the principles and methods covered in the introductory course, and more fully explore specific types of data analysis methods in more depth.

The Advanced Data Analysis courses draw on students' previous experience with data analysis and understanding of statistical theory to develop advanced, more sophisticated methods. These core courses involve extensive analysis of real data with emphasis on developing the oral and writing skills needed for communicating results.

Beginning Data Analysis

Choose *one* of the following courses:

36-200	Reasoning with Data *	9
36-220	Engineering Statistics and Quality Control	9

*A score of 5 on the Advanced Placement (AP) Exam in Statistics may be used to waive this requirement. 36-220 emphasizes examples in engineering and Architecture.

Note: Students who enter the program with credit for probability and statistical inference should discuss options with an advisor.

Sequence 1

Intermediate Data Analysis

Choose *one* of the following courses:

36-202	Methods for Statistics & Data Science *	9
36-290	Introduction to Statistical Research Methodology	9
36-303	Sampling, Survey and Society	9
36-309	Experimental Design for Behavioral & Social Sciences	9
36-315	Statistical Graphics and Visualization	9
*Must take prior to 36-401 or will need to take an additional Advanced Data Analysis Elective		

Advanced Data Analysis Electives

Choose *two* of the following courses:

36-311	Statistical Analysis of Networks	9
36-313	Statistics of Inequality and Discrimination	9
36-318	Introduction to Causal Inference	9
36-396	Tartan Athletics Analytics	9
36-460	Special Topics: Sports Analytics	9
36-461	Special Topics: Statistical Methods in Epidemiology	9
36-462	Special Topics: Statistical Machine Learning	9
36-463	Special Topics: Multilevel and Hierarchical Models	9
36-464	Special Topics: Psychometrics: A Statistical Modeling Approach	9
36-465	Special Topics: Conceptual Foundations of Statistical Learning	9
36-466	Special Topics: Statistical Methods in Finance	9
36-467	Special Topics: Data over Space & Time	9
36-468	Special Topics: Text Analysis	9

36-469	Special Topics: Statistical Genomics and High Dimensional Inference	9
36-470	Special Topics: Statistical Methods in Health Sciences	9
36-471	Special Topics: Time Series	9
36-472	Special Topics: Computational Statistical Methods in Life Sciences	9
36-473	Special Topics: Statistical Principles of Generative AI	9
36-490	Undergraduate Research	9
36-493	Sports Analytics Capstone	9
36-497	Corporate Capstone Project	9

All Special Topics (36-46x or 36-47x) courses are not offered every semester. They are on a rotation and new Special Topics are regularly added.

Sequence 2 (For students beginning later in their college career)

Advanced Data Analysis Electives

Choose *three* of the following courses:

36-311	Statistical Analysis of Networks	9
36-313	Statistics of Inequality and Discrimination	9
36-318	Introduction to Causal Inference	9
36-396	Tartan Athletics Analytics	9
36-460	Special Topics: Sports Analytics	9
36-461	Special Topics: Statistical Methods in Epidemiology	9
36-462	Special Topics: Statistical Machine Learning	9
36-463	Special Topics: Multilevel and Hierarchical Models	9
36-464	Special Topics: Psychometrics: A Statistical Modeling Approach	9
36-465	Special Topics: Conceptual Foundations of Statistical Learning	9
36-466	Special Topics: Statistical Methods in Finance	9
36-467	Special Topics: Data over Space & Time	9
36-468	Special Topics: Text Analysis	9
36-469	Special Topics: Statistical Genomics and High Dimensional Inference	9
36-470	Special Topics: Statistical Methods in Health Sciences	9
36-471	Special Topics: Time Series	9
36-472	Special Topics: Computational Statistical Methods in Life Sciences	9
36-473	Special Topics: Statistical Principles of Generative AI	9
36-490	Undergraduate Research	9
36-493	Sports Analytics Capstone	9
36-497	Corporate Capstone Project	9

All Special Topics (36-46x or 36-47x) courses are not offered every semester. They are on a rotation and new Special Topics are regularly added.

3. Probability Theory and Statistical Theory 18 units

The theory of probability gives a mathematical description of the randomness inherent in our observations. It is the language in which statistical models are stated, so an understanding of probability is essential for the study of statistical theory. Statistical theory provides a mathematical framework for making inferences about unknown quantities from data. The theory reduces statistical problems to their essential ingredients to help devise and evaluate inferential procedures. It provides a powerful and wide-ranging set of tools for dealing with uncertainty.

To satisfy the theory requirement, complete the following:

36-235	Probability and Statistical Inference I *	9
36-236	Probability and Statistical Inference II **	9

Note: Students who enter the program with credit for probability and statistical inference should discuss options with an advisor.

*It is possible to substitute 36-218, 36-219, 36-225, 15-259, or 21-325 for 36-235 . 36-235 is the standard (and recommended) introduction to probability, 36-219 is tailored for engineers and computer scientists, 36-218 and 15-259 are more mathematically rigorous classes for Computer Science students and more mathematically advanced (students need advisor approval to enroll), and 21-325 is a rigorous probability theory course offered by the Department of Mathematics.)

**It is possible to substitute 36-226 or 36-326(honors course) for 36-236. 36-236 is the standard (and recommended) introduction to statistical inference.

Please note that students who complete 36-235 are expected to take 36-236 to complete their theory requirements. Students who choose to take 36-225 instead will be required to take 36-226 afterward. They will not be eligible to take 36-236.

Comments:

(i) In order to meet the prerequisite requirements, a grade of at least a C is required in 36-235 (or equivalent) and 36-236 (or equivalent).

4. Statistical Computing 9 units

Fundamental to the practice of statistics and data science is the ability to effectively code data processing and analysis tasks. Within the domain of statistics, the use of the programming language R is ubiquitous, and thus we expose students to it throughout the curriculum (and in depth in Statistical Computing).

36-350	Statistical Computing	9

5. Modern Regression and Advanced Methodology 18 UNITS

Central to the practice of statistics is the implementation and interpretation of statistical models. The purpose of statistical models is to represent data-generating processes, such that predictions and inferential conclusions can be made about real-world phenomena. Proper modeling involves not just coding, but also thinking critically about data, research goals, and the validity of the models themselves, given their intrinsic assumptions. The courses 36-401 and 36-402 focus on the theory of statistical models (especially linear models and their extensions), how they are applied in real data analyses, and how to interpret and present these analyses in written reports.

To satisfy these requirements, complete the following:

36-401	Modern Regression	9
36-402	Advanced Methods for Data Analysis	9

Notes:

- 36-401 and 36-402 must be taken here at Carnegie Mellon University. Transfer credit will not be accepted.
- In order to meet the prerequisite requirements, a grade of at least a C is required in 36-401.

6. Machine Learning/Computer Science 57-60 units

Statistical modeling in practice nearly always requires computation in one way or another. Computational algorithms are sometimes treated as “black boxes," whose innards the statistician need not pay attention to. But this attitude is becoming less and less prevalent, and today there is much to be gained from a strong working knowledge of computational tools. Understanding the strengths and weaknesses of various methods allows the data analyst to select the right tool for the job; understanding how they can be adapted to work in new settings greatly extends the realm of problems that he/she can solve. While all majors in Statistics & Data Science are given solid grounding in computation, extensive computational training is really what sets the B.S. in Statistics and Machine Learning program apart. Note that we would advise those students who are considering receiving course credit for Fundamentals of Programming and Computer Science given their score on the AP Computer Science A exam to actually take the course at Carnegie Mellon instead, as within data science as a whole Python is far more widely used than Java.

15-112	Fundamentals of Programming and Computer Science	12
or 02-120	Programming for Scientists	
15-122	Principles of Imperative Computation	12
15-351	Algorithms and Advanced Data Structures	12
or 15-451	Algorithm Design and Analysis	
10-301	Introduction to Machine Learning	12
or 07-280	Artificial Intelligence and Machine Learning I	

Note: 07-280 is meant only for students pursuing a minor or additional major in Artificial Intelligence. Otherwise, students should plan on taking 10-301 to complete the requirement.

and take *one* of the following Machine Learning Advanced Electives:

05-434	Machine Learning in Practice	12
07-380	Artificial Intelligence and Machine Learning II ***	12
10-403	Deep Reinforcement Learning & Control	12
10-703	Deep Reinforcement Learning & Control *	12
10-405	Machine Learning with Large Datasets (Undergraduate)	12
10-605	Machine Learning with Large Datasets *	12
10-417	Intermediate Deep Learning	12
10-418	Machine Learning for Structured Data	12
10-707	Advanced Deep Learning *	12
11-344	Machine Learning in Practice	12
11-411	Natural Language Processing	12
11-441	Machine Learning with Graphs	9
11-485	Introduction to Deep Learning	12
11-661	Language and Statistics *	12
15-386	Neural Computation	9
15-387	Computational Perception	9
16-311	Introduction to Robotics	12
16-385	Computer Vision	12
16-720	Computer Vision *	12
*PhD level ML course as approved by Statistics advisor		
** Independent research with an ML faculty member as approved by Statistics Advisor		
*** 07-380 is a requirement for Artificial Intelligence minors and additional majors. Please double check with your advisor if there are any double counting policy concerns before using this course for this curriculum requirement.		
****This is not an exhaustive list. Please contact your Academic Advisor if there is a course you are considering taking that is not on this list.		

Total number of units for the major	**180–205 Units**
Total number of units for the degree	**360 Units**

Recommendations

Students in the Dietrich College of Humanities and Social Sciences who wish to declare a Statistics and Machine Learning major are advised to complete both the calculus requirement (one Mathematical Foundations calculus sequence) and the Beginning Data Analysis course 36-200 Reasoning with Data by the end of their Freshman year.

The linear algebra requirement is a prerequisite for the course 36-401 . It is therefore essential that students complete this requirement by their junior years at the latest.

Additional Major in Statistics and Machine Learning

Students who elect Statistics and Machine Learning as an additional major must fulfill *all* degree requirements.

With respect to double-counting courses, it is departmental policy that students must have at least six courses (three Computer Science/Machine Learning and three Statistics and Data Science) that do *not* count for their primary major. If students do not have at least six, they will need to take additional advanced data analysis or ML electives, depending on where the double counting issue is.

Students are advised to begin planning their curriculum (with appropriate advisors) as soon as possible. This is particularly true if the other major has a complex set of requirements and prerequisites or when many of the other major's requirements overlap with the requirements for the B.S. in Statistics and Machine Learning.

Sample Programs

The following sample programs illustrates two ways (of many) to satisfy the requirements for the B.S. in Statistics and Machine Learning. However, keep in mind that the program is flexible enough to support *many* other possible schedules and to emphasize a wide variety of interests.

The second schedule is an example of the case when a student enters the program through 36-235 and 36-236.

Schedule 1

First-Year		Second-Year	
Fall	**Spring**	**Fall**	**Spring**
21-111 Differential Calculus	Intermediate Data Analysis course	21-256 Multivariate Analysis	36-236 Probability and Statistical Inference II
36-200 Reasoning with Data	21-112 Integral Calculus	36-235 Probability and Statistical Inference I	21-241 Matrices and Linear Transformations
15-112 Fundamentals of Programming and Computer Science	21-127 Concepts of Mathematics	15-122 Principles of Imperative Computation	36-350 Statistical Computing
-----	-----	-----	-----

Third-Year		Fourth-Year	
Fall	**Spring**	**Fall**	**Spring**
10-301 Introduction to Machine Learning	36-402 Advanced Methods for Data Analysis	36-3xx or 36-4xx Advanced Data Analysis Elective	36-3xx or 36-4xx Advanced Data Analysis Elective
36-401 Modern Regression	15-351 Algorithms and Advanced Data Structures	Machine Learning Elective	-----
-----	-----	-----	-----
-----	-----	-----	-----

*In each semester, ----- represents other courses (not related to the major) which are needed in order to complete the 360 units that the degree requires.

Schedule 2

First-Year		Second-Year	
Fall	**Spring**	**Fall**	**Spring**
21-090 Precalculus	21-127 Concepts of Mathematics	21-256 Multivariate Analysis	36-236 Probability and Statistical Inference II
36-200 Reasoning with Data	21-120 Differential and Integral Calculus	36-235 Probability and Statistical Inference I	21-241 Matrices and Linear Transformations
15-112 Fundamentals of Programming and Computer Science	-----	15-122 Principles of Imperative Computation	36-3xx or 36-4xx Advanced Data Analysis Elective
-----	-----	-----	-----

Third-Year		Fourth-Year	
Fall	**Spring**	**Fall**	**Spring**
36-350 Statistical Computing	36-402 Advanced Methods for Data Analysis	15-351 Algorithms and Advanced Data Structures	Machine Learning Advanced Elective
36-401 Modern Regression	10-301 Introduction to Machine Learning	36-3xx or 36-4xx Advanced Data Analysis Elective	36-3xx or 36-4xx Advanced Data Analysis Elective
36-3xx or 36-4xx Advanced Data Analysis Elective	-----	-----	-----
-----	-----	-----	-----

*In each semester, "-----" represents other courses (not related to the major) which are needed in order to complete the 360 units that the degree requires.